Predictions and Policy Optimization in Online Decision Making

Thesis by
Yiheng Lin

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 13, 2025

ACKNOWLEDGEMENTS

I was extremely fortunate to meet my Ph.D. advisors Adam Wierman and Yisong Yue in 2019 and spend five years as their graduate student from 2020. Without their invaluable guidance and enormous support, this thesis would not have been possible. Since I started my first research project on smoothed online optimization as a junior undergraduate student, Adam has been a constant resource of wisdom from whom I would never hesitate to seek advice. As an advisor, Adam has done the best to prepare me at every stage of my graduate study: Problem solving, presentation, asking the right questions, and forming a complete thesis. His mentorship is not limited to research. During the time with great uncertainties, Adam's encouragements help me build up confidence to face many unprecedented obstacles. Yisong has set up a role model for me to conduct the most impactful research in an era of groundbreaking AI advances. Besides providing detailed suggestions about my projects, he always reminds me to not get bogged down in technical challenges and question the real significance/impact before diving into a specific problem. I was deeply impressed by his inspiring tutorial on how to find a "research vision" and build a "research taste," which help me gain momentum in my daily research work and communication. Yisong generously supported several group retreats, creating a relaxing atmosphere for group members to discuss everything from research to daily life like a big family.

I would like to thank my thesis committee members, Eric Mazumdar and Rayadurgam Srikant, as well as my candidacy committee member, Sicun Gao, who have devoted much time and effort. Eric always made himself available to provide advice on new project ideas, related literature, and presentations. I was fortunate to meet Rayadurgam Srikant multiple times throughout my Ph.D. to present my work and hear his insightful feedback. After years, I still got inspiration from his advice on asking the most fundamental questions, even in complicated settings. And I also appreciate the stimulating questions Sicun asked during my candidacy talk, which prompted me to start new projects on policy optimization.

I want to express my sincere gratitude to my excellent collaborators. In particular, I thank my undergraduate thesis advisor Longbo Huang for his great mentorship during my final year at Tsinghua University. Special thanks to Gautam Goel, Haoyuan Sun, and Guanya Shi for showing me what excellent teamwork looks like when I first started my research journey as a junior undergraduate student. I would like to thank Guannan Qu, who kept providing brilliant ideas and knowledgeable

suggestions in a sequence of projects on prediction power and networked MARL. I was lucky to collaborate with James A. Preiss, Yingying Li, Tongxin Li, and Zaiwei Chen in multiple projects on policy optimization. They shared with me not only their expertise on research but also heartfelt experiences in their academia careers. I would also like to thank Yash Kanoria for guiding me throughout our Networked OCO project and providing a detailed introduction to the writing style of the operation research community. I would like to thank Tianyu Chen for leading our collaboration on the adaptive bitrate streaming project, and I also appreciate the guidance provided by Ramesh K. Sitaraman, Mohammad Hajiesmaili, and Zahaib Akhtar on this project.

My graduate study would not have been this enjoyable without my dear colleagues and friends at Caltech. I want to thank Navid Azizan and Na (Lina) Li for hosting my brief visit to Boston in the summer of 2023. I also want to thank Steven Low for providing valuable advice on both research and career. Many thanks to my friends at Caltech CMS: Jeremy Bernstein, Geeling Chau, Nico Christianson, Lauren Conger, Wenqi Cui, Victor Dorobantu, Lu Gan, Linqi (Daniel) Guo, Tinashe Handina, Yujia Huang, Jiaqing Jiang, Eitan Levin, Zongyi Li, Chen Liang, Ziqi Ma, Laixi Shi, Yuanyuan Shi, Jialin Song, Peicong Song, Yu Su, Jennifer Sun, Sabera Talukder, Xuefei (Julie) Wang, Zihui (Ray) Wu, Fengze Xie, Yiheng Xie, Changhe Yang, Christopher Yeh, Jing Yu, Yizhou Zhang, Bingliang Zhang, Hongkai Zheng, Fengyu Zhou, and everyone at the Wierman and Yue group. I also want to thank the CMS administrative staff, especially Jolene Brink and Christine Ortega, for being extremely supportive on everything big or small. I also appreciate the support from all my other friends throughout my Ph.D. journey.

I would like to thank my parents and other family members for their unwavering love and support throughout this journey. Their encouragement, patience, and belief in me have been a constant source of strength, especially during challenging times. Reaching this milestone would not have been possible without their steadfast support and guidance.

ABSTRACT

Predictions are ubiquitous in modern systems, offering insights into how environments might evolve by encoding our prior knowledge and assumptions. Recent advances in artificial intelligence have significantly expanded the scope and accuracy of such models, creating vast new opportunities across domains. At the same time, online decision making remains a fundamental challenge in many real-world problems, concerned with challenges such as limited information, delayed feedback, and irrevocable actions. This dissertation focuses on the interplay between predictions and online decision making—how predictive information can be effectively leveraged to improve performance in dynamic, uncertain environments.

While incorporating predictions often enhances decision-making, the degree of improvement can vary substantially. This variability arises from two key factors. First, the potential benefit of using predictions is fundamentally determined by both the nature of the predictions (e.g., their targets, errors, and distributions) and the characteristics of the decision-making process (e.g., costs and dynamics). Second, standard predictive policies frequently fall short of realizing such potential, especially in changing environments or when critical system parameters are unknown.

This dissertation introduces a unified theoretical framework to quantify the benefit of leveraging predictions across a broad range of online decision-making problems. To close the gap between the maximum potential and achievable performance, we formulate a general policy optimization framework and design efficient algorithms capable of tracking optimal (predictive) policies in time-varying settings. Additionally, we address practical considerations such as scalability and computational efficiency, enabling the application of our methods in large-scale networks and on resource-constrained devices.

PUBLISHED CONTENT AND CONTRIBUTIONS

Lin, Yiheng, Zaiwei Chen, et al. (2025). “Maximizing the value of stochastic predictions in control: Accuracy is not enough.” In: *Under submission*.
Y.L. participated in the conception of the project, formulated the problem setting, proved the main theorems, and led the writing of the paper.

Preiss, James A. et al. (2025). “Fast non-episodic adaptive tuning of robot controllers with model-based online policy optimization.” In: *Under submission*.
Y.L. participated in the conception of the project, participated in algorithm design, and participated in the writing of the paper.

Chen, Tianyu et al. (2024). “SODA: An adaptive bitrate controller for consistent high-quality video streaming.” In: *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 613–644. DOI: `10.1145/3651890.3672260`.
Y.L. participated in the conception of the project and the development of algorithms, proved the theoretical guarantees, and participated in the writing of the paper.

Lin, Yiheng, James A. Preiss, Fengze Xie, et al. (2024). “Online policy optimization in unknown nonlinear systems.” In: *The Thirty Seventh Annual Conference on Learning Theory*. Proceedings of Machine Learning Research, pp. 3475–3522. URL: `https://proceedings.mlr.press/v247/lin24a.html`.
Y.L. participated in the conception of the project, proposed the meta-framework, proved the main theorems, and led the writing of the paper.

Lin, Yiheng, James A. Preiss, Emile Anand, et al. (2023). “Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations.” In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., pp. 53508–53521. URL: `https://proceedings.neurips.cc/paper_files/paper/2023/file/a7a7180fe7f82ff98eee0827c5e9c141-Paper-Conference.pdf`.

Lin, Yiheng, Judy Gan, et al. (2022). “Decentralized online convex optimization in networked systems.” In: *International Conference on Machine Learning*. Proceedings of Machine Learning Research, pp. 13356–13393. URL: `https://proceedings.mlr.press/v162/lin22c/lin22c.pdf`.
Y.L. participated in the conception of the project, formulated the networked models, proved the main theorems, and co-led the writing of the paper.

Lin, Yiheng, Yang Hu, Guannan Qu, et al. (2022). “Bounded-regret MPC via perturbation analysis: Prediction error, constraints, and nonlinearity.” In: *Advances in Neural Information Processing Systems* 35, pp. 36174–36187. URL: `https://proceedings.neurips.cc/paper_files/paper/2022/file/eadeef7c51ad86989cc3b311cb49ec89-Paper-Conference.pdf`.
Y.L. participated in the conception of the project, proposed the pipeline framework, proved the main theorems, and co-led in the writing of the paper.

Lin, Yiheng, Yang Hu, Guanya Shi, et al. (2021). "Perturbation-based regret analysis of predictive control in linear time varying systems." In: *Advances in Neural Information Processing Systems* 34, pp. 5174–5185. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/298f587406c914fad5373bb689300433-Paper.pdf.
Y.L. participated in the conception of the project, proved the main theorems, and co-led the writing of the paper.

Lin, Yiheng, Guannan Qu, et al. (2021). "Multi-agent reinforcement learning in stochastic networked systems." In: *Advances in Neural Information Processing Systems* 34, pp. 7825–7837. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/412604be30f701b1b1e3124c252065e6-Paper.pdf.
Y.L. participated in the conception of the project, formulated the problem, proved the main theorems, and led the writing of the paper.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Part I

Introduction

Chapter 1

INTRODUCTION

Online decision making studies the problem of making sequential decisions under uncertainty, where information and feedback are typically revealed over time, potentially with delays. Due to limited information, a decision made at any given moment may become suboptimal as new information emerges. As a result, predictions about future uncertainties are extremely valuable since they expand the available information for making each decision, enabling more sophisticated decision-making policies that may achieve better performances. Predictions are increasingly employed across a wide range of decision-making tasks. For instance, in autonomous driving, machine learning (ML) models can predict different types of obstacles and their potential trajectories, enabling the controller to plan maneuvers that avoid collisions and ensure a smooth ride. Similarly, in video streaming, forecasts of future network throughput allow the controller to select the bitrate of video segments that minimizes re-buffering delays and maximizes viewing quality (see Figure 1.1 for an illustrative example). When applied to electric vehicle charging, predictions of future electricity prices and departure times inform the controller's selection of charging rates, ultimately improving the consumers' satisfaction rate and the service provider's profit. Despite their broad applicability, different prediction tasks and methods exhibit varied performance characteristics. Sophisticated ML-based predictors often excel in environments similar to those represented in their training data or simulators, but their performance can deteriorate when faced with unexpected or changing conditions. In contrast, simpler predictors may produce relatively larger errors yet remain more robust under diverse real-world scenarios.

Incorporating predictions into control systems often enhances performance, yet the extent of improvement can vary greatly across different scenarios. In many applications, even limited predictions within a short lookahead window can yield substantial cost reductions. In contrast, there are empirical and theoretical counterexamples in which more sophisticated predictive methods and controllers do not perform well. This discrepancy raises a critical question: under what conditions do predictions reliably improve control performance? Addressing this question would help system designers determine whether to adopt specific predictive methods or other alternative approaches in a control task. While prior research has studied this

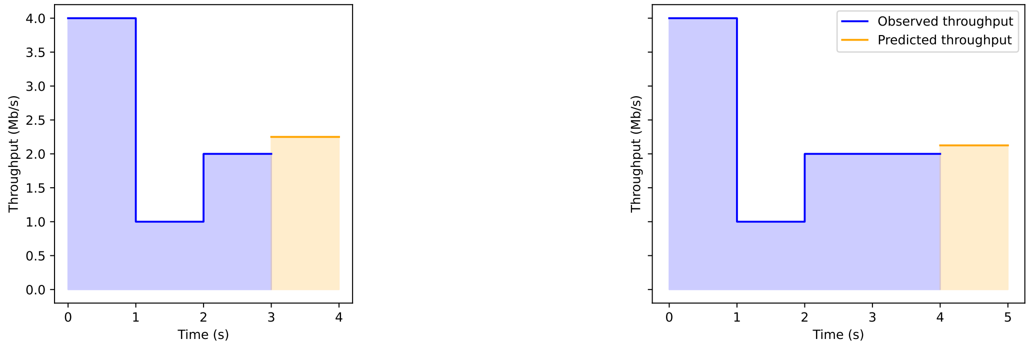

Figure 1.1: An illustrative example of making predictions in adaptive bitrate streaming: Given the observed network throughput in past t seconds (blue curve), predict the throughput for the next second (orange curve).

topic within specific combinations of prediction models, dynamical systems, and predictive controllers, a unified analytical framework that addresses this question more broadly across diverse settings remains a critical open challenge.

While leveraging predictions can improve the performance of online decision making significantly, one important aspect is understanding how to develop controllers to use predictions with time-varying reliability. For instance, ML-based predictors often suffer from distribution shifts, meaning their reliability may change over time. To address this challenge, we can introduce parameters to the predictive control policies that allow us to change how we use predictions. Taking the predictive control with confidence coefficients (Li, Yang, Qu, Shi, et al., 2022; Lin, Preiss, Anand, et al., 2023) as an example, the adaptivity is achieved by tuning the coefficients that control how much the controller trusts each prediction entry. More broadly, the problem of finding the optimal way to use predictions can be viewed as a special case of learning the parameters of any control policy to optimize its performance. There is a need for a general framework to optimize control policy adaptively in time-varying environments with provable guarantees. The framework is helpful in many real-world applications, such as quadcopters experiencing fluctuating wind conditions or variable payloads, where a fixed control policy may fail to provide near-optimal responses in hindsight.

For online decision making and policy optimization algorithms to be practical, achieving scalability and efficiency in large-scale or complex systems is essential. For instance, in a large networked system where the nodes collaborate to maximize their average total reward, each node must rely only on localized information, as global communication is often too costly. As another example, when deploying an online policy optimization algorithm on a quadcopter, minimizing memory and computational complexity is critical due to hardware constraints and the need for frequent updates. While many prior works focus primarily on theoretical guarantees, they often overlook the practical challenges of scalability and efficiency. Addressing these concerns is crucial for ensuring that proposed algorithms have significant impacts when applied to real-world applications.

This thesis aims to provide analytical frameworks to characterize the benefit of using predictions in control under general prediction/dynamical models and propose efficient/scalable policy optimization algorithms with provable guarantees.

1.1 Major Challenges and Prior Work

Two key factors fundamentally shape the performance of leveraging predictions in online decision making: the *prediction power* and the *policy optimization* process. Prediction power characterizes the intrinsic value of the predictions themselves—the extent to which they can improve the control performance in principle. It depends on how the predictions relate to unknown system parameters, underlying dynamics, and cost structures. Even highly accurate predictions may offer limited benefit if they do not align with the decision-relevant uncertainties. Complementing this, policy optimization focuses on how to realize the full potential of predictions in practice. It involves identifying or tracking the optimal decision policy under constraints such as limited feedback, computational resources, and time-varying environments. Together, these two dimensions—what predictions can offer and how well that offer can be harnessed—define the central challenges in this domain.

Prediction Power. We use the term *prediction power* to refer to the benefit of using predictions in online decision making compared to the no-prediction case. To attain such benefit, we rely on control policies that can leverage future predictions in deciding the control actions, termed *predictive control*. Among various design philosophies of predictive control, perhaps the most prominent approach is Model Predictive Control (MPC), also known as receding horizon control. Generally speaking, at each time step, an MPC-style controller leverages all available predictions to solve an optimal control problem in a short horizon and commits to the first action in the planned trajectory. MPC's flexibility allows it to accommodate challenges from time-varying/nonlinear dynamics and constraints on the state or control actions, and it is known to work well in practice. Theoretically, while many classical results are about MPC's asymptotic behaviors (e.g., stability and convergence), there is a growing interest in establishing non-asymptotic learning guarantees that quantify how the sub-optimality of MPC reduces as the predictions become more powerful. However, existing results are limited to linear time-invariant (LTI) dynamics with quadratic costs because the proof approaches require explicitly writing down optimal control actions and MPC's control actions. Even with more flexibility allowed in controller design rather than focusing on MPC-style controllers, handling time-varying dynamics and/or constraints is still challenging.

While many works on predictive control start with the ideal model where the predictions are exact, most predictions in real-world applications are imperfect. Characterizing the impact of such inaccurate predictions is essential in two ways: 1) Showing the robustness of a predictive control approach (e.g., MPC) against bounded (or even adversarial) prediction errors and 2) providing guidance on how to reduce the control cost by choosing or improving the predictions. Many prior works have taken a natural path to extend the results under an ideal exact-prediction model: They model the observed predictions as the actual targets plus bounded adversarial perturbations. These works derive regret bounds that depend on the *prediction errors*—the magnitude of such perturbations—and show predictive controllers maintain similar regret guarantees with the exact-prediction case if the prediction errors are sufficiently small. However, obtaining near-accurate predictions is too much to hope for in many applications, while weaker predictions with stochastic correlations with their targets still proved useful. As the prediction error grows, the worst-case regret bounds become overly conservative because they overlook the potential stochastic dependencies between predictions and future uncertainties. Further, optimizing the prediction accuracy does not always lead to better control costs.

Multiple practical challenges may arise when we apply predictive control to real-world problems. First, scalability issues are significant when implementing an MPC-based predictive control approach in large networked systems. Each node cannot afford the complexity of gathering information from the whole network and solving the predictive optimization problem globally. Second, critical assumptions that lead to theoretical guarantees may break under application-specific objectives and constraints. For example, in high-quality video streaming, the key assumption about the exponentially decaying perturbation property of the optimal control problem breaks due to the structure of the objectives and buffer constraints. Lastly, the optimal predictive controller may depend on the unknown joint distribution of predictions and disturbances or change as the environment changes. Therefore, the user cannot directly implement the optimal predictive controller without learning the key system/controller parameters.

Policy Optimization. Optimization is the process of choosing the best option in a set of feasible alternatives. Policy optimization can be viewed as a special class of optimization problems, where the objective is the state/control cost and each solution corresponds to a control policy. Compared with classic offline optimization settings where the solver has full knowledge of the objective function/constraints, the online nature of the decision-making process brings the major challenge for policy optimization: When updating the policy at an intermediate time step, the agent only receives incremental feedback without a “whole picture” that includes, for example, how the costs/dynamics will change in the future.

One challenge of online policy optimization arises from the dynamical system, where each action has impacts beyond the current time step. Prior works on online policy optimization build on techniques from online optimization, but their results often rely on the convexity of the objectives with respect to previous policy parameters – an assumption that typically confines applications to linear dynamics with specific policy classes. In practice, this convexity assumption can break down easily for more general policy classes (e.g., state-feedback controllers) or nonlinear dynamics.

Another challenge in policy optimization is that the underlying dynamical (or transition) model is often partially or completely unknown to the control agent. In many control applications, the user usually has some knowledge about a nominal system, so a common approach is to follow a *model-based* framework. Specifically, a standard technique is to apply random perturbations to collect data, enabling the agent to obtain a sufficiently accurate approximation of the dynamical model with high probability. With the learned model, one can solve the optimal policy or apply online policy optimization algorithms that use the models to compute the gradients. However, the approach is generally restricted to linear systems with specific policy classes, because extrapolating from local data to approximate global models works for time-invariant linear dynamics but not for nonlinear or time-varying systems.

Practical challenges in implementing policy optimization algorithms may arise from complexity and scalability issues—not only due to the structure of the control policy class, but also from the optimization process itself. Even when the policy class is simple or low-dimensional, identifying and tracking the optimal policy can remain computationally demanding, particularly in settings with multi-agent coordination requirements or limited resources. Consider a large-scale networked system where the agents work together to maximize the global average reward. Although each agent's policy that we wish to optimize might only depend its local state, the optimal policies depend on the global interaction structure in general. However, the size of the global state/action space grows exponentially with respect to the number of agents, so one cannot afford to run a classic policy optimization approach on the global scale. As another example, when applying online policy optimization on quadcopter, the limited resources onboard put constraints on complexity of the algorithm in terms of both computation and memory. In addition, the fast-changing environment may require the policy be updated at a high frequency, so the policy optimization algorithm must run efficiently.

Connections between Prediction Power and Policy Optimization. To understand the relationship between the two major parts of this thesis on a high level, we draw an intuitive analogue with a classic optimization problem. The performance of online decision making can be viewed as the "objective," while each predictive control policy corresponds to a "solution." Under this analog, the study of *prediction power* asks the question of *how good an optimal solution can be.* In contrast, policy optimization aims to *find/learn a (near-)optimal solution*.

Policy optimization can help achieve the prediction power. Specifically, predictive control methods such as MPC can be viewed as policy classes parameterized by how they incorporate predictions. For example, in an MPC framework with confidence coefficients, the controller fully trusts the predictions when the coefficient is one and ignores them entirely when the coefficient is zero. To fully exploit the benefits of available predictions, the agent must adopt the optimal predictive policy within a suitable policy class. However, deriving closed-form optimal policies is often intractable in time-varying or large-scale systems. Policy optimization approaches help overcome this difficulty by learning a (near-)optimal predictive policy using limited observations and/or computing.

On the other hand, prediction power is a fundamental quantity to characterize before using predictions and implementing policy optimization. For example, if an evaluation of prediction power suggests that the potential benefit is small, there is no need to obtain such predictions and/or deploy policy optimization algorithms. Otherwise, if the evaluation suggests the prediction power is large, meaning the potential benefit could be significant, we will focus on finding the optimal predictive policy with possible challenges from time-varying environments or unknown parameters. Policy optimization provides powerful tools for this subsequent step.

1.2 Motivating Applications

In this section, we introduce three motivating applications of our theoretical frameworks on prediction power and policy optimization.

Adaptive Bitrate Streaming. Adaptive bitrate streaming studies the problem of deciding the bitrates for video download sequentially, where the controller has access to (unreliable) predictions about network throughputs. As online video streaming becomes increasingly popular nowadays, the users watch videos from devices with different hardware capabilities and connect to the internet in a variety of ways, leading to a diverse range of network conditions. In addition, the network throughput can also be unstable due to congestion and other complicated network issues. The goal of adaptive bitrate streaming is to enhance the users' experience by dynamically adjusting the video bitrate under changing network conditions.

Adaptive bitrate streaming is challenging because the controller needs to balance multiple objectives including optimizing video quality, minimizing rebuffering frequency, and reducing bitrate switching. In addition, since the buffer length is usually short in live streaming, the controller must react quickly to fluctuating network conditions and make robust decisions with volatile throughput decisions. Our theoretical framework provides a promising approach to tackle these challenges by formulating the problem as online optimal control: The dynamics capture the buffer level change as the controller downloads new video segments into the buffer, and the user consumes video segments to watch. The objective balances different metrics that may affect the user's experience. Using the insights from our theoretical results, we design MPC-based policies that can leverage throughput predictions while being robust against prediction errors.

Adaptive Tuning of Quadcopter Controllers. We consider an application of general policy optimization for quadcopter control that does not involve predictions. Quadcopter is a type of unmanned aerial vehicle that uses four rotating propellers to generate lift and maneuver. It becomes increasingly popular in a variety of real-world applications ranging from agriculture to photography. Given the diversity of the deployment environments, the user cannot rely on a certain "expert" control policy that is set by default when the quadcopter is produced. For example, the user may put on additional attachments (e.g., carrying goods) or operate the quadcopter in an unexpected (e.g., high wind) environment. The default policy may become suboptimal in such scenarios.

Our works on policy optimization are promising for tackling the specific challenges arising in quadcopter control. First, our algorithms and results apply to general policy classes that satisfy the contraction properties, allowing sophisticated designs for safety while maintaining sufficient flexibility to adapt. Second, a primary goal of our online policy optimization framework is to adapt quickly to changes of the environment that may occur frequently during the flight (e.g., periodic wind disturbance). Third, we design policy optimization algorithms with high computation/memory efficiency, which make them applicable with limited computing hardware onboard.

Networked Systems. Networked systems are a class of multi-agent systems with special interaction/reward structures. Specifically, each node in the network corresponds to an individual agent, and it only interacts with its immediate neighbors. Each node has a local reward (or cost) function, and collectively, the nodes aim to maximize (or minimize) the sum of these local functions.

Various multi-agent systems can be captured under this general structure. One example is Wi-Fi networks, where each user (node) transmits packets to a nearby access point. A collision occurs if multiple users send packets to the same access point simultaneously. In this setting, a user's local state – namely, the packets waiting to be transmitted before a given deadline – depends on both its own actions and those of its neighbors. A second example is the Susceptible-Infected-Susceptible (SIS) epidemic network, in which an agent's state is either "susceptible" or "infected." The probability of transitioning from "susceptible" to "infected" depends on the number of infected neighbors as well as on whether the agent takes a preventive action at some cost. Lastly, consider a product network for multiple items sold on an online retail platform. Each product's demand is affected by its current and previous prices, as well as by the prices of complementary or supplementary products. To fit this setting into the networked system framework, we model each product as a node and use edge to encode complementary or supplementary relationships. The objective is to determine prices that maximize the platform's total revenue in a manner that is both efficient and interpretable.

By studying predictive control and policy optimization in networked systems, we demonstrate how to overcome the practical challenge of scalability by exploiting the interaction structure among nodes/agents.

1.3 Thesis Roadmap and Contributions

We introduce the basic settings of online optimization, online control, prediction models, and the networked systems in Chapter 2. Then, we present the main results of this thesis in two major parts.

Part II: Prediction Power

While the study of prediction power has received much attention recently, a novel perspective that sets our works apart is the *perturbation analysis*. Specifically, instead of analyzing the online process directly, we look at the underlying optimization problem (offline) and study the behaviors of the optimal solution as a function of problem parameters. Perturbation analysis plays a critical role in establishing prediction power bounds: On the one hand, it allows us to characterize the impact of having limited/inaccurate predictions at a single step; On the other hand, it helps bound the accumulative effect of per-step impacts along the whole horizon. Compared with prior works, our proof framework built upon perturbation analysis significantly generalizes the scope of settings where prediction powers can be characterized. It is particularly useful for MPC-style controllers, which represent the design principle behind a class of predictive policies that are flexible and empirically successful.

In this context, Part II of this thesis focuses on characterizing the benefit of using predictions in online decision making/control. We start by studying MPC under an adversarial prediction model that does not rely on any distributional assumptions. In Chapter 3, we provide a general analysis pipeline to establish finite-time optimality guarantees for model predictive control (Lin, Hu, Shi, et al., 2021). The pipeline reduces the study of MPC to the perturbation analysis, enabling the derivation of regret bounds of MPC under various settings.

- In Section 3.1, we first introduce the perturbation analysis for finite-time optimal control problem and define the exponential decaying properties that we want to establish. We discuss about its intuitions and derivations in classic settings.

- In Section 3.2, we present the main theorems of the pipeline. With the exponential decaying properties that we have derived in perturbation analysis, we use the pipeline to establish finite-time performance bounds for MPC. Our results show the insight that, although it becomes harder to get near-accurate predictions when we predict further into the future, they also have less impact on the performance than predictions that are closer to the current step.

- In Section 3.3, we extend our proof framework based on the perturbation analysis to online optimization in networked systems. The key observation is that we can establish the exponential decaying properties not only in “temporal” dimension—the time horizon that includes the future and the past—but also in “spatial” dimension, which implies the impact of a node on another node decays quickly as their graph distance in the network grows. Using the generalized exponential decaying properties, we propose a localized predictive control algorithm with provable performance guarantees. Our results show predictions into the future and information sharing among neighbors are both important, and the two “resources” should be balanced to achieve the best performance bound.

- In Section 3.4, we consider the application of adaptive bitrate streaming. We use the theoretical insight of the perturbation analysis to design the objective function of a novel MPC-style controller so that the exponential decay property holds. Our proposed approach achieves superior empirical performance in production experiments, but we will focus on discussing the theoretical insights in this thesis.

A limitation of the adversarial prediction model in Chapter 3 is that it overlooks the stochastic relationships between predictions and future uncertainties. As a result, the performance bounds may fail to characterize the benefit of using predictions, because the worst-case analysis is overly conservative. Therefore, in Chapter 4, we study the benefit of using predictions under a general stochastic model, under which the predictions and environment uncertainties are sampled from a joint distribution. In this context, we seek to characterize a general notion of the prediction power, which is the *maximum* cost improvement under the *optimal* predictive policy compared with the no-prediction scenario. In Section 4.2, we provide sufficient conditions for establishing the lower bound of the prediction power. The conditions put requirements on the landscape of the cost-to-go functions and the conditional covariance of the policy’s actions. In Section 4.3, we instantiate the general lower bound with specific online optimal control settings such as linear quadratic regulator (LQR). Our results highlight that even “weak” (in terms of stochastic dependencies) predictions have the potential to provide fundamental benefit in online control, and we provide examples to explain why strict improvements on prediction accuracy does not necessarily reduce the total state/control cost.

Part III: Policy Optimization

Our results for online policy optimization build upon a property called *contractive perturbation* that makes the goal of tracking the optimal policy in a dynamic environment tractable. Intuitively, contractive perturbation ensures each policy has its "preferred trajectory" and converges towards it if deployed for several steps in a row. It generalizes a key property of many standard control policies and helps policy optimization by containing the impact of past explorations. In Chapter 5, we consider online policy optimization on a single trajectory under contractive perturbation. We propose efficient online policy optimization algorithms that can adapt quickly with provable finite-time guarantees. This chapter is organized as following:

- In Section 5.2, we define the contractive perturbation property formally and provide examples to demonstrate its generality. Although our original definition requires the property to hold for a slowly-changing policy parameter sequence, we show it suffices to verify it under a fixed policy parameter.
- In Section 5.3, we present an efficient online policy optimization algorithm—Memoryless Gradient-based Adaptive Policy Selection (M-GAPS)—that leverages the first-order derivatives of the costs/dynamics to perform gradient-based updates on the policy parameters. Under the contractiveness assumption, we show that M-GAPS approximates the behavior of an ideal online gradient descent algorithm on the policy parameters. When convexity holds, M-GAPS achieves the optimal policy regret. When convexity does not hold, we establish a local regret bound for M-GAPS.
- In Section 5.4, we address the challenge of implementing online policy optimization when the dynamics are only partially known. Specifically, we assume there is an unknown component in the dynamics that can be time-varying and state-dependent. We develop a meta-framework that combines a module for learning the unknown component in dynamics with an online policy optimization algorithm like M-GAPS. Our theoretical results suggest that, to mimic the behavior of an online policy optimization algorithm when the true dynamics are known, it is unnecessary to identify the unknown component globally.
- In Section 5.5, we apply M-GAPS to select the parameters in a class of predictive control policies, so the controller can use the predictions adaptively based on their qualities and/or their stochastic relationship with future uncertainties. Intuitively, when the quality of predictions changes during the control process, M-GAPS adjusts the confidence coefficients that determine how much the MPC policy trusts the provided predictions. As a result, the controller "trusts" the predictions more when they are accurate and reduces the "trust" when they are bad. Our experiments show M-GAPS can learn the optimal

predictive control policy and adapt quickly to changes, building connections between online policy optimization with prediction power (Part III).

- In Section 5.6, we present a practical implementation of M-GAPS on quadcopter control. Focusing on the task of trajectory tracking, we use M-GAPS to tune the feedback gains of a quadcopter controller. In hardware experiments, M-GAPS can improve from a suboptimal initialization to near an expert controller that takes several days of efforts to tune manually. Further, we test M-GAPS with heavy payload or time-varying wind conditions. In these scenarios, M-GAPS can rapidly adapt to the disturbances and substantially reduce the cost compared with the expert controller. The experiment results demonstrate the hardware practicality of M-GAPS.

In Chapter 6, we use multi-agent reinforcement learning (MARL) to optimize the local policies for agents in a large-scale networked system. Classic centralized policy optimization algorithms are intractable because their complexities grow with the size of global state space, which is exponentially large in the number of agents. To address this challenge, we propose a scalable actor critic algorithm. Compared with its centralized counterpart, the most significant change is on the critic part, where we adopt localized truncation on the state/action space to reduce the computation/space complexity. Thus, we focus on the theoretical foundation of this key change in this thesis. In Section 6.2, we utilize the interaction structure among agents to derive the exponential decay property of local Q functions. With this property, we can confirm that localized information in the truncated state/action space is sufficient for approximating the local Q function, which depends on the global state/action in general. In Section 6.3, we introduce the connection between our truncated critic design and TD learning with state aggregation. Thus, we can use results on state aggregation to show finite-sample error bounds on Q function evaluation. Lastly, in Section 6.4, we apply our scalable actor critic algorithm to the settings of wireless and spreading networks. The results show that the proposed algorithm can improve the local policies effectively.

Chapter 2

BACKGROUND

In this chapter, we summarize the underlying problem settings in this thesis with the goal to build intuitions for our main results in Parts II and III. We start with the most basic setting of online optimization (Section 2.1) and discuss how it is generalized to study control with dynamical systems (Section 2.2). Then, in Section 2.3, we introduce different methods to model predictions and their relationship with future uncertainties. Lastly, in Section 2.4, we present two settings of networked systems that consider continuous or discrete decisions, respectively.

2.1 Online Optimization

The basic form of online optimization is a two-player game between the online agent and the (potentially adversarial) environment.

Classic online optimization. At each time step, the agent makes a decision $x_t \in \mathcal{X}$. A cost function $f_t : \mathcal{X} \to \mathbb{R}$ is revealed, and the agent incurs the stage cost $f_t(x_t)$. The agent's goal is to minimize the total cost $\sum_{t=1}^{T} f_t(x_t)$. The performance is usually evaluated by comparing against the optimal trajectory in hindsight (e.g., the static regret $\sum_{t=1}^{T} f_t(x_t) - \min_{x^* \in \mathcal{X}} \sum_{t=1}^{T} f_t(x^*)$).

In classic online optimization, each step's decision is independent with other steps. However, it is common for the current step's cost to also depends on the previous action. For example, in data center/power system operation, additional costs are incurred when the servers/generators are turned on or off. Thus, it is sometime preferable to have "smoother" decision trajectories where the decisions (e.g., the number of running servers) does not change dramatically between consecutive time steps. These considerations motivate the extension of the classic online optimization framework to include switching costs.

Smoothed online optimization. At each time step, a hitting cost function $f_t : \mathcal{X} \to \mathbb{R}$ and a switching cost function $c_t : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are revealed. The agent decides an action $x_t \in \mathcal{X}$ and incurs a stage cost $f_t(x_t) + c_t(x_t, x_{t-1})$. The agent's goal is to minimize the total cost $\sum_{t=1}^{T} f_t(x_t) + \sum_{t=2}^{T} c_t(x_t, x_{t-1})$. To evaluate the performance, we compare the algorithm's trajectory with the hindsight optimal trajectory $x_{1:T}^*$. The difference with classic online optimization is that $x_{1:T}^*$ can change over time.

Following the intuition to penalize changes between consecutive steps, researchers have considered different forms of switching costs such as the norm distance $c_t(x_t, x_{t-1}) = \|x_t - x_{t-1}\|$ or the squared ℓ_2-norm $c_t(x_t, x_{t-1}) = \frac{1}{2}\|x_t - x_{t-1}\|_2^2$. Indeed, the switching costs can be more general as long as they satisfy some assumptions jointly with the hitting costs. Other generalizations consider longer coupling with the previous time steps. For example, the switching cost $c_t(x_{t-k:t})$ depends on the decisions at $k > 1$ past time steps.

2.2 Online Control

In the setting of online optimization, we can choose any x_t from the set $\mathcal{X}$ (at some cost). However, in most control applications, we cannot steer the system to an arbitrary state in one step. Instead, we must pick control actions that can affect the system's state through the dynamics.

Dynamics. We consider a discrete-time dynamical system that is given by

$$x_{t+1} = g_t(x_t, u_t; w_t^*), \text{ where } x_t \in \mathcal{X}, u_t \in \mathcal{U}, \text{ and } w_t^* \in \mathcal{W}. \tag{2.1}$$

Here, x_t denotes the system's current state; u_t denotes the control action; and w_t^* denotes the disturbance or exogenous input. We use the star superscript to indicate that w_t^* is the true value (i.e., the actual disturbance experienced by the controller) and distinguish with any predictions/estimations. The dynamical function $g_t : \mathcal{X} \times \mathcal{U} \times \mathcal{W} \to \mathcal{X}$ decides how the current state and control action decide the next state x_{t+1} together. Function g_t is determined by the specific application, and its subscript t means it may change over time. A simple example is the linear time-invariant (LTI) dynamics, where function g_t has the form

$$g_t(x_t, u_t; w_t^*) = Ax_t + Bu_t + w_t^* \text{ for } x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^m, \text{ and } w_t^* \in \mathbb{R}^n. \tag{2.2}$$

Although many previous works on online control focus on the LTI dynamics in (2.2), some real-world applications demand different generalizations. For example,

- *Linear Time-Varying (LTV) Dynamics.* Dynamical matrices (A, B) are replaced by $\{(A_t, B_t)\}_{t=0,1,\ldots}$ that change over time.
- *Nonlinear Dynamics.* g_t can be a nonlinear function of (x_t, u_t).
- *State/Action Constraints.* $\mathcal{X}$ and $\mathcal{U}$ can be bounded sets.

Policy. The way we decide the control action u_t based on the state x_t is called the *control policy*, which is a function π_t that maps any state in $\mathcal{X}$ to an action in

$\mathcal{U}$. For example, a classic policy class for controlling the LTI dynamics in (2.2) decides the control action by taking a linear feedback on the state $u_t = -Kx_t$. To control general dynamical systems, one may consider adopting more sophisticated optimization-based or ML-based policies that uses additional information (e.g., future predictions or expert advice). The problem of choosing control actions can be converted to finding the optimal control policy. To formulate this problem rigorously, we consider a parameterized policy class

$$u_t = \pi_t(x_t; \theta_t), \tag{2.3}$$

where $\theta_t \in \Theta$ is the *policy parameter* to learn and update throughout the control process. The policy class $\pi_t(\cdot, \cdot)$ together with policy parameter set Θ encode all feasible control policies that we want to consider for the purpose of cost optimization.

Costs. At each time step, after taking the control action u_t, the agent incurs a stage cost that depends on the current state/action pair (x_t, u_t):

$$c_t = f_t(x_t, u_t). \tag{2.4}$$

Note that the subscript t allows f_t to change between time steps. The objective of online control is to minimize the total cost incurred through a finite horizon T, i.e., $\sum_{t=0}^{T-1} c_t$. Often, the cost functions f_t are defined by the user, so it is common to assume that they are known. However, if we need to consider uncertainties in costs, one way to model them is extending the cost function to be

$$c_t = f_t(x_t, u_t; w_t^*). \tag{2.5}$$

Here, w_t^* is an *uncertainty parameter* that combines the uncertainties at time step t, so it is shared with the dynamical function $x_{t+1} = g_t(x_t, u_t; w_t^*)$.

As a remark, stability is a fundamental issue in control theory, which means the system state must be kept in a region that is “safe.” For example, the linear dynamics approximation only holds if the state is not too far away from the equilibrium point. However, stability guarantees are not the primary focus of online control. Specifically, the user should design the policy class in online control carefully, so changing the policy parameter θ_t will not destabilize the system. As a result, online policy optimization algorithms can focus on minimizing the total cost $\sum_{t=0}^{T-1} c_t$.

Connection with Online Optimization. Smoothed online optimization is a special case of online control. One common approach to do the reduction is to consider a simple dynamics $x_{t+1} = x_t + u_t$, where the state x_t corresponds to the decision in online optimization. The stage cost $f_t(x_t, x_t - x_{t-1})$ is also general enough to capture different combinations of hitting/switching costs. Although online control is a more general setting, smoothed online optimization is still important for both theory and practice. For example, establishing the exponentially decaying perturbation property is easier for smoothed online optimization. Meanwhile, for some dynamical systems, one can reduce online control to smoothed online optimization with the help of control canonical form or uniform controllability assumptions.

2.3 Prediction Models

Prediction is a general term that includes any useful information about what may happen in the future. To model the prediction theoretically in the context of online optimization/control, we denote the prediction that the agent receives at time step t as $v_t \in \mathcal{V}$. Intuitively, the prediction v_t is provided by the environment in addition to the original observations (e.g., state x_t) that the agent can make at time t. Although one can combine v_t and x_t to form a "large" state $\bar{x}_t \coloneqq (x_t, v_t)$ when designing the control policy, we treat v_t and x_t separately when studying prediction power because it is easier to characterize the incremental benefit of observing v_t. Another reason for separating v_t is that the predictions are *oblivious* in many applications, meaning that they come from an exogenous source. As a result, the realization of v_t will not be affected by the agent's past actions $u_{0:t-1}$. In some settings, the assumption about oblivious predictions is critical for deriving the theoretical guarantees.

The goal of formulating the *prediction models* is to explain how the prediction v_t provided at time step t is related to the unknown future uncertainties $w^*_{t:T-1}$. In this thesis, we consider the following two types of prediction models.

Adversarial prediction model. At time step t, prediction v_t is a vector that combines the predictions of future uncertainties within a finite lookahead window k:

$$v_t = \left(w_{t|t}, w_{t+1|t}, \ldots, w_{t+k-1|t}\right), \tag{2.6}$$

where $w_{\tau|t}$ denotes the prediction of the true uncertainty parameter w^*_τ at time step t ($\tau \geq t$). Recall that we use w^*_τ to denote the true disturbance in dynamics (2.1) or true parameter in the stage cost (2.5) in online control. Under this model, the predictions and the ground-truth are assumed to be chosen by an adversary subject to certain constraints. For example, an ideal special case considers the setting that all predictions are exact (i.e., $w_{\tau|t} = w^*_\tau$ for $\tau = t, \ldots, t + k - 1$). To study the robustness of predictive controllers against prediction errors, we can relax the constraints to allow the adversary to provide predictions subject to some prediction error bounds (i.e., the distance $\left\|w_{\tau|t} - w_\tau^*\right\|$ is upper bounded). The adversarial is conservative because it consider the worst-case scenarios and overlooks any stochastic relationships.

Stochastic prediction model. In stochastic prediction model, we assume the predictions $v_{0:T-1}$ and the unknown parameters $w_{0:T-1}^*$ are sampled together from some joint distribution. Unlike (2.6), we do not need further assumptions on what v_t predicts about or its specific structure. Intuitively, the stochastic predictions are useful because they contain information about future uncertainties.

Compared with the adversarial prediction model, the stochastic prediction model is particularly useful for characterizing the benefit of "weak" predictions that are far from accurate. For example, some predictions may be useful because they have stochastic dependencies with the true uncertainty parameters. However, if we only look at their prediction errors, it may be difficult to distinguish them with "useless" predictions who contain no mutual information with future uncertainty parameters. As a result, we cannot conclude they are useful under the adversarial prediction model. On the other hand, one can see a limitation of the stochastic prediction model arises from the joint distribution assumption. Even if the assumption holds, another challenge is that the optimal predictive policy depends on the joint distribution, which we often do not know in practice. In contrast, the adversarial prediction model does not rely on any distributional assumptions.

2.4 Networked Systems

Networked systems is a special class of multi-agent systems that features localized interaction/dependency structures. Specifically, for a network (undirect graph) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, each node $i \in \mathcal{V}$ represents an agent, and an edge $e \in \mathcal{E}$ between two nodes indicate that they can directly affect each other. Each node has its own local state/action space. We consider different forms of interaction/dependency between nodes:

- **Transition probabilities in MDP.** The local state of node i at time step $t+1$ depends on the current local states/actions of node i's direct neighbors:

  $$P(s_i(t+1) \mid s_{N_i}(t), a_{N_i}(t)), \text{ where } N_i = \{j \mid \mathrm{dist}_{\mathcal{G}}(i, j) \leq 1\}.$$

  Here, $s_{N_i}(t), a_{N_i}(t)$ denotes the local states/actions of the nodes in N_i at time t, and $s_i(t+1)$ denotes the local state of node i at time $t+1$.

- **Cost functions in online optimization.** For each edge $e = (i, j) \in \mathcal{E}$, nodes i and j share a cost $s_t^e(x_t^i, x_t^j)$ that depends on their decisions jointly, where x_t^i and x_t^j denote the local decisions of nodes i and j, respectively.

Therefore, the decision of a node i can affect any other node j because the impact can travel through a multi-edge path from i to j even if they are not direct neighbors.

In this thesis, we focus on collaborative settings, where all nodes work together to optimize the global cost/reward. The global cost/reward can be decomposed to be the sum/average of local costs/rewards. Due to observation/complexity constraints, each agent adopts a localized policy that uses information within a finite neighborhood. Policy optimization is challenging in this setting due to the global dependence, and a centralized approach can be intractable for large-scale networks.

Part II

Prediction Power

Chapter 3

ADVERSARIAL PREDICTIONS

Predictions play a important role in enhancing online control performance, particularly when dealing with uncertain or adversarial environments. In contrast to purely stochastic models, the adversarial prediction model enables robust performance guarantees without relying on specific probabilistic assumptions regarding the predictions and uncertainties. Under this adversarial model, we show that the classic MPC-style policies, requiring only a short prediction horizon, can demonstrate near-optimal performance in a wide range of dynamical systems and cost functions. Moreover, with exact predictions, the performances of MPC-style policies improve at an exponential rate as the prediction horizon increases. With inexact predictions, the impact of an error for predicting further into the future decays exponentially with respect to the temporal distance. In turn, the theoretical analysis offers valuable insights into designing better MPC-style policies in applications such as adaptive bitrate streaming.

This chapter is mainly based on the following papers:

[Lin, Hu, Shi, et al., 2021] Lin, Yiheng, Yang Hu, Guanya Shi, Haoyuan Sun, Guannan Qu, and Adam Wierman. "Perturbation-based regret analysis of predictive control in linear time varying systems." Advances in Neural Information Processing Systems 34 (2021): 5174-5185.

[Lin, Gan, et al., 2022] Lin, Yiheng, Judy Gan, Guannan Qu, Yash Kanoria, and Adam Wierman. "Decentralized online convex optimization in networked systems." In International Conference on Machine Learning, pp. 13356-13393. Proceedings of Machine Learning Research, 2022.

[Lin, Hu, Qu, et al., 2022] Lin, Yiheng, Yang Hu, Guannan Qu, Tongxin Li, and Adam Wierman. "Bounded-regret MPC via perturbation analysis: Prediction error, constraints, and nonlinearity." Advances in Neural Information Processing Systems 35 (2022): 36174-36187.

[Chen, Lin, et al., 2024] Chen, Tianyu, Yiheng Lin, Nicolas Christianson, Zahaib Akhtar, Sharath Dharmaji, Mohammad Hajiesmaili, Adam Wierman, and Ramesh K. Sitaraman. "SODA: An adaptive bitrate controller for consistent high-quality video streaming." In Proceedings of the ACM SIGCOMM 2024 Conference, pp. 613-644. 2024.

3.1 Perturbation Analysis

Perturbation analysis examines how changes in a system's parameters—such as initial states, dynamic functions, or cost functions—affect the solutions of a finite-time optimal control problem. By quantifying how these solutions deviate when parameters are perturbed, perturbation analysis provides guarantees about how errors that come from inexact predictions or finite prediction horizons propagate through the online process of running an MPC-style policy.

A core component of both the design of MPC-style control policies and our analysis is the following *finite-time optimal control problem* (FTOCP). Given a time interval $[t_1, t_2]$, the FTOCP solves the optimal sub-trajectory subjected to the given initial state z, terminal cost F, and a sequence of (potentially noisy) parameters $\xi_{t_1:t_2-1}, \zeta_{t_2}$, as formalized in the following definition.

Definition 3.1.1 (FTOCP)**.** *The finite-time optimal control problem (FTOCP) over the time horizon $[t_1, t_2]$, with parameters $\xi_{[t_1,t_2]} \coloneqq (z_{t_1}, w_{t_1:t_2-1}, \zeta_{t_2})$ and terminal cost function $F(\cdot;\cdot)$, is defined as*

$$\begin{aligned}
\iota_{t_1}^{t_2}(\xi_{[t_1,t_2]}; F) \coloneqq \min_{x_{t_1:t_2}, u_{t_1:t_2-1}} & \sum_{t=t_1}^{t_2-1} f_t(x_t, u_t; w_t) + F(x_{t_2}; \zeta_{t_2}) \\
\textit{s.t. } & x_{t+1} = g_t(x_t, u_t; w_t), && \forall t_1 \le t < t_2, \\
& s_t(x_t, u_t; w_t) \le 0, && \forall t_1 \le t < t_2, \\
& x_{t_1} = z_{t_1}, && (3.1)
\end{aligned}$$

and a corresponding optimal solution as $\psi_{t_1}^{t_2}(\xi_{[t_1,t_2]}; F)$.

The components of $\xi_{[t_1,t_2]}$ correspond to different elements in FTOCP. z_{t_1} corresponds to the initial state. $w_{t_1:t_2-1}$ correspond to the uncertainty parameters for intermediate time steps. It is shared between the dynamics function g_t, cost function f_t, and the constraint function s_t. ζ_{t_2} is the parameter of the terminal cost. Sometimes, when we need to write down the components of $\xi_{[t_1,t_2]}$ and the context is clear, we use the shorthand $(z_{t_1}, w_{t_1:t_2})$ to denote $(z_{t_1}, w_{t_1:t_2-1}, w_{t_2})$.

As a remark, Definition 3.1.1 formulates FTOCP in a general form. In many settings, there might be less uncertainties that we want to consider (e.g., unconstrained, or

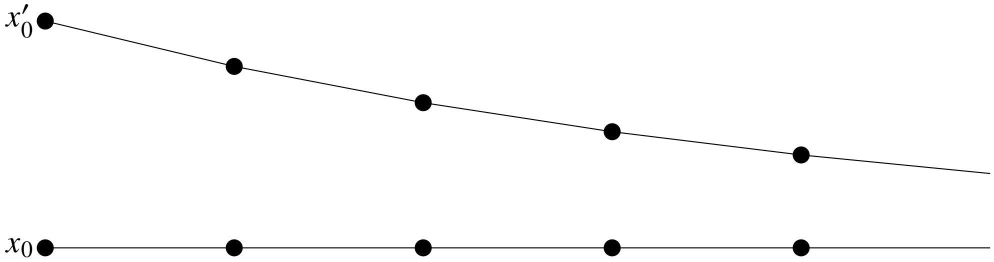

Figure 3.1: Illustration of the exponentially decaying perturbation.

deterministic cost functions). The FTOCP in (3.1) does not include a terminal constraint set. To compensate for this, we allow the terminal cost $F(\cdot; \zeta_{t_2})$ to take value $+\infty$ in some subset of $\mathbb{R}^n$. For example, a terminal cost function that we frequently use later is the indicator function of the terminal parameter ζ_{t_2}, where $\zeta_{t_2} \in \mathbb{R}^n$. We use $\mathbb{I}$ to denote such indicator terminal cost (i.e., $\mathbb{I}(y_{t_2}; \zeta_{t_2}) = 0$ if $y_{t_2} = \zeta_{t_2}$ and $\mathbb{I}(y_{t_2}; \zeta_{t_2}) = +\infty$ otherwise).

Theorem 3.1.1 (Meta Perturbation Bound)**.** *Consider the FTOCP defined in* (3.1)*. Given any parameters* $\xi_{[t_1,t_2]} := (z_{t_1}, w_{t_1:t_2-1}, \zeta_{t_2})$ *and* $\xi'_{[t_1,t_2]} := (z'_{t_1}, w'_{t_1:t_2-1}, \zeta'_{t_2})$*,*

$$\begin{aligned}&\left\| \psi_{t_1}^{t_2}(\xi_{[t_1,t_2]}; F)_{x_h} - \psi_{t_1}^{t_2}(\xi'_{[t_1,t_2]}; F)_{x_h} \right\| \\ &\leq C \left(\lambda^h \left\| z_{t_1} - z'_{t_1} \right\| + \sum_{\tau=0}^{t_2-t_1-1} \lambda^{|h-\tau|} \left\| w_\tau - w'_\tau \right\| + \lambda^{t_2-t_1} \left\| \zeta_{t_2} - \zeta'_{t_2} \right\| \right)\end{aligned}$$

hold for all time intervals $[t_1, t_2]$.

Instantiation with SOCO

Perhaps the most straightforward instantiation of Theorem 3.1.1 is in SOCO. Recall that the classic setting of SOCO is an online game played by an agent against an adversary: at each time step t, the adversary reveals a hitting cost function $\hat{f}_t$, a switching cost function $\hat{c}_t$, and a disturbance (or exogenous input) $\hat{w}_t$. The agent picks a decision point $\hat{x}_t \in \mathbb{R}^n$, and incurs a stage cost of $\hat{f}_t(\hat{x}_t) + \hat{c}_t(\hat{x}_t, \hat{x}_{t-1}; \hat{w}_{t-1})$. The agent seeks to minimize the total cost it incurs throughout the game. The offline optimal cost is defined as the minimum cost if the agent has full knowledge of the costs and disturbances at the start of the game. Instead of analyzing the performance of an online algorithm directly, our focus is on studying how the perturbations of the system parameters (initial state, terminal state, and disturbances) impact the offline optimal solution. We consider the case when the terminal state is fixed to $\hat{x}_p$, which is given as one of the problem parameters.

To begin, observe that when the initial state $\hat{x}_0$, terminal state $\hat{x}_p$, and the disturbances $\hat{w}$ are given, the optimal p-step trajectory of SOCO can be obtained from the unconstrained optimization problem

$$\hat{\psi}(\hat{x}_0, \hat{w}, \hat{x}_p) := \operatorname*{arg\,min}_{\hat{x}_{1:p-1}} \sum_{\tau=1}^{p-1} \hat{f}_\tau(\hat{x}_\tau) + \sum_{\tau=1}^{p} \hat{c}_\tau(\hat{x}_\tau, \hat{x}_{\tau-1}; \hat{w}_{\tau-1}), \tag{3.2}$$

where the objective is a convex function of the decision variables $\hat{x}_{1:p-1}$. Since (3.2) is an unconstrained optimization problem, the gradient of its objective equals zero at $\hat{\psi}(\hat{x}_0, \hat{w}, \hat{x}_p)$. Using this, we can further show that the directional derivative of $\hat{\psi}(\hat{x}_0, \hat{w}, \hat{x}_p)$ along some direction e, denoted by χ, satisfies the linear equation $M\chi = \delta$, where symmetric matrix M is the Hessian of the objective and vector δ is determined by the direction e. A special structure of the objective of (3.2) is that the correlations only occur in two consecutive time steps. This implies that its Hessian M is block tri-diagonal. Such tri-diagonal structure of M has been noted by previous work, e.g., Amos et al., 2018, and have been leveraged to solve the linear equation $M\chi = \delta$ quickly. In contrast, we focus on the exponential decay phenomena M^{-1} exhibits, i.e., the magnitudes of entries decay exponentially with respect to their distances to the main diagonal (Demko, Moss, and Smith, 1984). Bounding each entry of $\chi = M^{-1}\delta$ separately gives us the following perturbation bound.

Theorem 3.1.2. *Given a tuple $(\hat{x}_0, \hat{w}, \hat{x}_p)$ that contains the initial state, the disturbances, and the terminal state in this order, we consider the optimal solution of the SOCO problem*

$$\hat{\psi}(\hat{x}_0, \hat{w}, \hat{x}_p) \coloneqq \operatorname*{arg\,min}_{\hat{x}_{1:p-1}} \sum_{\tau=1}^{p-1} \hat{f}_\tau(\hat{x}_\tau) + \sum_{\tau=1}^{p} \hat{c}_\tau(\hat{x}_\tau, \hat{x}_{\tau-1}; \hat{w}_{\tau-1})$$

indexed by $1, \ldots, p-1$. Assume $\hat{f}_\tau : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex, $\hat{c}_\tau : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}$ is convex and ℓ-strongly smooth, and both are twice continuously differentiable for $\tau = 1, \ldots, p$, then

$$\begin{aligned} &\left\|\hat{\psi}(\hat{x}_0, \hat{w}, \hat{x}_p)_h - \hat{\psi}(\hat{x}_0', \hat{w}', \hat{x}_p')_h\right\| \\ &\leq C_0 \left(\lambda_0^{h-1}\left\|\hat{x}_0 - \hat{x}_0'\right\| + \sum_{\tau=0}^{p-1} \lambda_0^{|h-\tau|-1}\left\|\hat{w}_\tau - \hat{w}_\tau'\right\| + \lambda_0^{p-h-1}\left\|\hat{x}_p - \hat{x}_p'\right\|\right) \end{aligned}$$

for all $1 \leq h \leq p-1$, where $C_0 = (2\ell)/\mu$ and $\lambda_0 = 1 - 2 \cdot \left(\sqrt{1 + (2\ell/\mu)} + 1\right)^{-1}$.

We defer the formal proof of Theorem 3.1.2 to Appendix 3.A. As a remark, we do not require the hitting cost $\hat{f}_\tau$ to be strongly smooth, or the switching cost $\hat{c}_\tau$ to be strongly convex in Theorem 3.1.2.

Instantiations in Unconstrained LTV Systems

We study the FTOCP with LTV dynamics:

$$\begin{aligned} \psi_{t_1}^{t_2}(\xi_{[t_1,t_2]}; F) \coloneqq \operatorname*{arg\,min}_{x_{t_1:t_2}, u_{t_1:t_2-1}} & \sum_{t=t_1}^{t_2-1} f_t(x_t) + \sum_{t=t_1}^{t_2-1} c_t\,(u_t) + F(x_{t_2}; \zeta_{t_2}) \\ \text{s.t. } & x_{t+1} = A_t x_t + B_t u_t + w_t,\ t_1 \le t < t_2, \\ & x_{t_1} = z_{t_1}, \end{aligned} \tag{3.3}$$

where $\xi_{[t_1,t_2]} \coloneqq (z_{t_1}, w_{t_1:t_2-1}, \zeta_{t_2})$. Because we often set F to be the indicator function $\mathbb{I}$ that is defined as $\mathbb{I}(x;\zeta) = 0$ if $x = \zeta$ and $\mathbb{I}(x;\zeta) = +\infty$ otherwise, we introduce the shorthand $\psi_{t_1}^{t_2}(\xi_{[t_1,t_2]}) \coloneqq \psi_{t_1}^{t_2}(\xi_{[t_1,t_2]}; \mathbb{I})$ to simplify the notations.

As is standard in studies of regret and competitive ratio in linear control problems, we assume the cost functions are well-conditioned.

Assumption 3.1.1. *The cost functions satisfy the following conditions:*

1. *$f_t(\cdot)$ is both m_f-strongly convex and ℓ_f-strongly smooth for all t.*
2. *$c_t(\cdot)$ is both m_c-strongly convex and ℓ_c-strongly smooth for all t.*
3. *$f_t(\cdot)$ and $c_t(\cdot)$ are twice continuously differentiable for all t.*
4. *$f_t(\cdot)$ and $c_t(\cdot)$ are non-negative, and $f_t(0) = c_t(0) = 0$ for all t.*
5. *The terminal cost function F is the indicator function $\mathbb{I}$ or satisfies that (1) $F(\cdot;\zeta)$ is twice continuously differentiable and m_f-strongly convex for all ζ; (2) For a positive constant ℓ_F, we have*

$$\|\nabla_x F(x;\zeta) - \nabla_x F(x;\zeta')\| \le \ell_F \|\zeta - \zeta'\|, \textit{ for all } x, \zeta, \zeta'.$$

Note that assumptions (1) through (3) are quite common (Li, Chen, and Li, 2019; Li, Qu, and Li, 2021; Goel and Wierman, 2019; Goel, Lin, et al., 2019; Shi et al., 2020). Assumption (4) is less common, but can be satisfied via re-parameterization without loss of generality. Specifically, when the minimizers of state cost f_t and control cost c_t are nonzero, we perform the transformation

$$x_t' \leftarrow x_t - \operatorname*{arg\,min}_x f_t(x),\ u_t' \leftarrow u_t - \operatorname*{arg\,min}_u c_t(u),$$

$$w_t' \leftarrow w_t + A_t \arg\min_x f_t(x) + B_t \arg\min_u c_t(u).$$

The intuition of this transformation is that, when the minimizer of the cost function for the next step is known, we can always perform a translation in the state and control space to align the minimizer with the origin.

Additionally, we need to assume the dynamics are *controllable*. It is crucial that the dynamical system can be steered from an arbitrary initial state to an arbitrary final state via a finite sequence of admissible control actions. For linear time-invariant (LTI) systems, the full-rankness of the *controllability matrix* completely characterizes the reachability of the state space, which is generally used as a standard assumption for analysis (Zhang, Li, and Li, 2021; Mania, Tu, and Recht, 2019; Astrom and Murray, 2008). This can be generalized to parallel assumptions for LTV systems as follows. We begin with a definition.

Definition 3.1.2. *For a dynamical system with linear time-varying dynamics $x_{t+1} = A_t x_t + B_t u_t + w_t, t = 0, \dots, T-1$, the transition matrix $\Phi(t_2, t_1) \in \mathbb{R}^{n \times n}$ (from time step t_1 to t_2) is defined as*

$$\Phi(t_2, t_1) := \begin{cases} A_{t_2-1} A_{t_2-2} \cdots A_{t_1} & \text{if } t_2 > t_1 \\ I & \text{if } t_2 \leq t_1 \end{cases},$$

and the controllability matrix $M(t,p) \in \mathbb{R}^{n \times (mp)}$ is defined as

$$M(t,p) := \left[\Phi(t+p, t+1)B_t, \Phi(t+p, t+2)B_{t+1}, \dots, \Phi(t+p, t+p)B_{t+p}\right].$$

The dynamical system is called controllable if there exists a constant $d \in \mathbb{Z}_+$, such that the controllability matrix $M(t,d)$ is of full row rank for any $t = 0, \dots, T-d$. The smallest constant d with such property is called the controllability index of the system.

Given the above definition, we can state the key assumption necessary for the analysis of LTV systems. We use a slightly stronger assumption than being merely controllable, which we refer to as (d, σ)-uniform controllability. It is a natural generalization of its counterpart for LTI systems (see Assumption 2 in Mania, Tu, and Recht, 2019, where (d, σ) is instead named as (ℓ, ν)).

Assumption 3.1.2. *There exists positive constants a, b, and b', such that*

$$\|A_t\| \leq a, \ \|B_t\| \leq b, \text{ and } \|B_t^\dagger\| \leq b'$$

hold for all time steps $t = 0, \dots, T-1$, where $B_t^\dagger$ denotes the Moore–Penrose inverse of matrix B_t. Furthermore, there exists a positive constant σ such that

$$\sigma_{\min}\left(M(t, d)\right) \geq \sigma$$

holds for all time steps $t = 0, \dots, T-d$, where d denotes the controllability index.

Note that Assumption 3.1.2 implies $\sigma_{\min}(M(t,p)) \geq \sigma$ for all $p \geq d$ because appending more columns to a matrix with full row rank will not reduce its minimum singular value.

We now build upon the SOCO perturbation result to derive a perturbation result for LTV systems. In particular, we show an exponentially-decaying perturbation bound for our LTV system by reducing it to SOCO and apply Theorem 3.1.2. As we have discussed, LTV systems are more difficult than SOCO because the dynamics prevent the online agent from picking the next state x_{t+1} freely at a given state x_t. We overcome this obstacle by redefining the decision points as illustrated in Figure 3.2. Specifically, given state x_t at time step t as the last decision point, we then ask the online agent to decide state x_{t+d} at time step $(t+d)$ rather than x_{t+1} at time step $(t+1)$.

Since d is the controllability index, x_{t+d} can be picked freely from the whole space $\mathbb{R}^n$ regardless of x_t. We also utilize the *principle of optimality*, e.g., if $x_{t:t+k}, u_{t:t+k-1}$ is the optimal solution to $\psi_t^{t+k}((z_t, w_{t:t+k-1}, z_{t+k}))$, then $x_{i:j}, u_{i:j-1}$ is the optimal solution to $\psi_i^j((x_i, w_{i:j-1}, x_j))$ for any $t \leq i < j \leq t+k$. Therefore, the trajectory between time t and $(t+d)$ can be recovered by solving $\psi_t^{t+d}((x_t, w_{t:t+d-1}, x_{t+d}))$. So we are able to formulate a valid SOCO problem on the sequence of time steps $t, t+d, t+2d, \dots$.

Naturally, the hitting cost at time step $(t+d)$ remains the same, while the switching cost becomes $\Upsilon_t^{t+d}((x_t, w_{t:t+d-1}, x_{t+d}))$, where the function Υ_t^{t+p} is defined as

$$\Upsilon_t^{t+p}((z_t, w_{t:t+p-1}, z_{t+p})) \coloneqq \iota_t^{t+p}(z_t, w_{t:t+p-1}, z_{t+p}) - f_{t+p}(z_{t+p}). \tag{3.4}$$

An illustration of the reduction can be found in Figure 3.2. We would like to point out that our reduction from optimal control to SOCO is novel in that it leverages the principle of optimality to apply to more general LTV settings, as opposed to the reduction via control canonical forms in Li, Chen, and Li, 2019 that is specific to LTI systems. Unlike the switching costs in Goel, Lin, et al., 2019; Goel and Wierman, 2019; Chen, Goel, and Wierman, 2018; Argue, Gupta, and Guruganesh, 2020 which are explicitly defined as the ℓ_2-distance or squared ℓ_2-distance, the switching cost Υ_t^{t+p} here is defined implicitly as the optimal value of an optimization problem. Lemma 3.1.3 shows that the switching cost defined in (3.4) satisfies the requirements of Theorem 3.1.2, which allows us to obtain the desired perturbation bound.

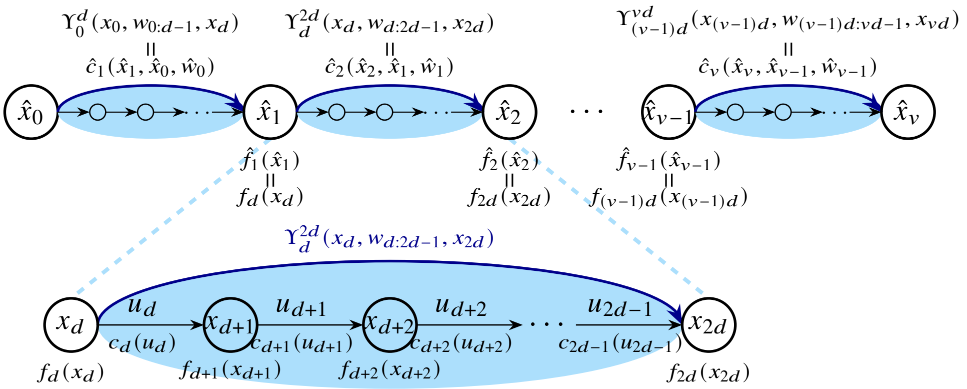
$\Upsilon_0^d(x_0, w_{0:d-1}, x_d)$
$\hat{c}_1(\hat{x}_1, \hat{x}_0, \hat{w}_0)$
$\Upsilon_d^{2d}(x_d, w_{d:2d-1}, x_{2d})$
$\hat{c}_2(\hat{x}_2, \hat{x}_1, \hat{w}_1)$
$\Upsilon_{(v-1)d}^{vd}(x_{(v-1)d}, w_{(v-1)d:vd-1}, x_{vd})$
$\hat{c}_v(\hat{x}_v, \hat{x}_{v-1}, \hat{w}_{v-1})$
$\hat{x}_0$
$\hat{x}_1$
$\hat{x}_2$
$\hat{x}_{v-1}$
$\hat{x}_v$
$\hat{f}_1(\hat{x}_1)$
$f_d(x_d)$
$\hat{f}_2(\hat{x}_2)$
$f_{2d}(x_{2d})$
$\hat{f}_{v-1}(\hat{x}_{v-1})$
$f_{(v-1)d}(x_{(v-1)d})$
$\Upsilon_d^{2d}(x_d, w_{d:2d-1}, x_{2d})$
x_d
u_d
$c_d(u_d)$
x_{d+1}
u_{d+1}
$c_{d+1}(u_{d+1})$
x_{d+2}
u_{d+2}
$c_{d+2}(u_{d+2})$
u_{2d-1}
$c_{2d-1}(u_{2d-1})$
x_{2d}
$f_d(x_d)$
$f_{d+1}(x_{d+1})$
$f_{d+2}(x_{d+2})$
$f_{2d}(x_{2d})$

Figure 3.2: Illustration of the reduction from LTV to SOCO. Here we consider a simple example where $t = 0$ and $p = vd$. At time step 0, the agent cannot steer the system to an arbitrary target state at the next time step due to dynamical constraints. However, given (d, σ)-uniform controllability, the controller is able to enforce an arbitrary target state after d time steps, which prompts the transformation to a SOCO problem with a decision point in every d time steps.

Lemma 3.1.3. *Under Assumption 3.1.1 and 3.1.2, for integer $p \geq d$, we have*

1. $\psi_t^{t+p}(\xi_{[t:t+p]})$ *is $L_1(p)$-Lipschitz in* $\xi_{[t:t+p]}$;
2. $\xi_t^p(\xi_{[t:t+p]})$ *is convex and $L_2(p)$-strongly smooth in* $\xi_{[t:t+p]}$.

Here $L_1(p) = C(p)\left(1 + \ell \cdot C(p)/m_c\right), L_2(p) = \ell \cdot C(p)^2 + \ell^2 \cdot C(p)^4/m_c$, *where* $\ell = \max(\ell_f, \ell_c)$,

$$C(p) = \begin{cases} O(a^{3p}) & \text{if } a > 1; \\ O(p^2) & \text{if } a = 1; \\ O(1) & \text{if } a < 1. \end{cases}$$

In Lemma 3.1.3, we use $O(\cdot)$ to hide quantities a, b, and $1/\sigma$; the precise expression of $C(p)$ and the proof of Lemma 3.1.3 can be found in Appendix 3.A. Using the reduction from LTV to SOCO, we obtain a perturbation bound for the LTV systems in Theorem 3.1.4, the proof of which is deferred to Appendix 3.A.

Theorem 3.1.4. *Consider the FTOCP defined in* (3.3) *and with a horizon length $p \geq d$. Under Assumptions 3.1.1 and 3.1.2, given any parameter sets*

$$\xi_{[t,t+p]} \coloneqq (x_t, w_{t:t+p-1}, \zeta_{t+p}) \text{ and } \xi'_{[t,t+p]} \coloneqq (x'_t, w'_{t:t+p-1}, \zeta'_{t+p}),$$

we have

$$\begin{aligned} &\left\|\psi_t^{t+p}(\xi_{[t,t+p]}; F)_{x_{t+h}} - \psi_t^{t+p}(\xi'_{[t,t+p]}; F)_{x_{t+h}}\right\| \\ &\leq C\left(\lambda^h \|x_t - x'_t\| + \sum_{\tau=0}^{p-1} \lambda^{|h-\tau|} \|w_{t+\tau} - w'_{t+\tau}\| + \lambda^{p-h} \|\zeta_{t+p} - \zeta_{t+p}\|\right). \end{aligned}$$

Here we define $L_0 = \max_{d \leq p \leq 2d-1} L_2(p)$, and the constants are given by

$$\begin{aligned} \lambda &= \left(1 - 2\left(\sqrt{1 + (2L_0/m_c)} + 1\right)^{-1}\right)^{\frac{1}{2d-1}}, \\ C &= \max\left\{1, \frac{\ell_F}{m_F}\right\} \cdot \frac{2L_0}{m_c} \cdot \left(1 - 2\left(\sqrt{1 + (2L_0/m_c)} + 1\right)^{-1}\right)^{-1}. \end{aligned}$$

Theorem 3.1.4 allows us to bound the distance between any two trajectories so long as they can be expressed as the optimal solutions of the FTOCP (3.3). For example, to bound the norm of each state in the predictive trajectory $\tilde{\psi}_t^p(x, \zeta; F)$, we only need to set $x' = 0, \zeta' = 0$ in the first inequality because an all zero trajectory can be expressed as $\tilde{\psi}_t^p(0, 0; F)$.

More examples of FTOCP exponentially decaying perturbation bounds can be found in Lin, Hu, Qu, et al., 2022 and Chen, Lin, et al., 2024. Note that this property does not hold for all dynamics/costs, and counterexamples are provided in Lin, Hu, Qu, et al., 2022.

3.2 From Perturbation Bounds to Control Performance Guarantees

We first introduce the general predictive online control problem including the settings, the objective, available information, and the predictive controller class. Then, we introduce the MPC algorithm, which is a widely-used predictive controller that we focus on. Specifically, we consider a general, finite-horizon, discrete-time optimal control problem with *time-varying costs, dynamics and constraints*, namely

$$\begin{aligned} \min_{x_{0:T}, u_{0:T-1}} \; & \sum_{t=0}^{T-1} f_t(x_t, u_t; w_t^*) + F_T(x_T; w_T^*) \\ \text{s.t. } & x_{t+1} = g_t(x_t, u_t; w_t^*), && \forall 0 \leq t < T, \\ & s_t(x_t, u_t; w_t^*) \leq 0, && \forall 0 \leq t < T, \qquad (3.5) \\ & x_0 = x(0). \end{aligned}$$

Here, $x_t \in \mathbb{R}^n$ is the *state*, $u_t \in \mathbb{R}^m$ is the *control input* or *action*; f_t is a time-varying *stage cost* function, g_t is a time-varying *dynamical* function, and s_t is a time-varying *constraint* function, all parameterized by a ground-truth parameter w_t^* (unknown to an online controller); and F_T is a terminal cost function parameterized by w_T^* that regularizes the terminal state.

The offline optimal trajectory OPT is obtained by solving (3.5) with the full knowledge of the true parameters $w_{0:T}^*$. In contrast, an online controller can only observe noisy estimations of the parameters in a fixed prediction horizon to decide its current action u_t at each time step t. For example, MPC picks u_t by calculating the optimal sub-trajectory confined to the prediction horizon. The objective is to design an online controller that can compete against the offline optimal trajectory OPT. We use *dynamic regret* as the performance metric, which is widely used to evaluate the performance of online controllers/algorithms in the literature of online control (Lin, Hu, Shi, et al., 2021; Yu et al., 2022; Zhang, Li, and Li, 2021) and online optimization (Li, Qu, and Li, 2021; Goel, Lin, et al., 2019; Lin, Goel, and Wierman, 2020). Specifically, for a concrete problem instance $(x(0), w_{0:T}^*)$, let $\text{cost}(\mathsf{OPT})$ denote the total cost incurred by OPT, and $\text{cost}(\mathsf{ALG})$ denote the total cost incurred by an online controller ALG. The *dynamic regret* is defined as the worst-case additional cost incurred by ALG against OPT, i.e., $\sup_{x(0), w_{0:T}^*} (\text{cost}(\mathsf{ALG}) - \text{cost}(\mathsf{OPT}))$.

Model Predictive Control

We focus on *Model Predictive Control (MPC)*, a popular predictive controller. In this subsection, we first define the available information (predictions) as well as its quality (prediction power), and how general predictive online controllers make decisions. Then, we introduce MPC as a predictive online controller.

We represent the uncertainties in cost functions, dynamics, constraints, and terminal costs as function families parameterized by w_t: $\mathcal{F}_t := \{f_t(x_t, u_t; w_t) \mid w_t \in \mathcal{W}_t\}$, $\mathcal{G}_t := \{g_t(x_t, u_t; w_t) \mid w_t \in \mathcal{W}_t\}$, $\mathcal{S}_t := \{s_t(x_t, u_t; w_t) \mid w_t \in \mathcal{W}_t\}$, and $\mathcal{F}_T := \{F_T(x_T; w_T) \mid w_T \in \mathcal{W}_T\}$. The online controller knows the function families $\mathcal{F}_{0:T}$, $\mathcal{G}_{0:T-1}$, and $\mathcal{S}_{0:T-1}$ as prior knowledge, but it does not know the true parameters $w_{0:T}^* \in \prod_{\tau=0}^{T} \mathcal{W}_\tau$. Instead, at time step t, the online controller has access to noisy predictions of these parameters for the future k time steps (where k is called the *prediction horizon*), represented by $w_{t:t+k|t} \in \prod_{\tau=t}^{t+k} \mathcal{W}_\tau$. The parameter space $\mathcal{W}_t$ at each time step t may have different dimensions.

We formally define the quality of predictions by introducing the following notion of prediction error.

Definition 3.2.1. *The prediction error is defined as* $\rho_{t,\tau} \coloneqq \left\| w_{t+\tau|t} - w^*_{t+\tau} \right\|$ *for an integer* $\tau \geq 0$. *The power of* τ*-step-away predictions (for parameter* w*) is defined as* $P(\tau) \coloneqq \sum_{t=0}^{T-\tau} \rho_{t,\tau}^2$.

Under this noisy prediction model, a general predictive online controller ALG decides the control action based on the current state and the latest available predictions of future parameters. We formally define the class of predictive online controllers considered in this paper in Definition 3.2.2, which includes MPC as a special case.

Definition 3.2.2. *A predictive online controller* ALG *is a function that takes the current state* x_t *and the available predictions* $w_{t:t+k|t}$ *as inputs at time* t *and outputs the current control action* u_t*, i.e.,* $u_t = \mathsf{ALG}(x_t, w_{t:t+k|t})$. *We use* $x_0 \xrightarrow{u_0} x_1 \xrightarrow{u_1} \cdots \xrightarrow{u_{T-1}} u_T$ *to denote the trajectory achieved by* ALG*, and use* $x_0 \xrightarrow{u_0^*} x_1^* \xrightarrow{u_1^*} \cdots \xrightarrow{u_{T-1}^*} u_T^*$ *to denote the offline optimal trajectory* OPT.

We formally introduce MPC using the definition of the FTOCP (Definition 3.1.1). The pseudocode of this online controller is given in Algorithm 1. Basically, at time step t, MPC_k solves a k-step predictive FTOCP using the latest available parameter predictions, and commits the first control action in the solution. When there are only fewer than k steps left, MPC_k directly solves a $(T-t)$-step FTOCP at time t until the end of the horizon, using the predicted real terminal cost $F_T(\cdot; w_{T|t})$. This MPC controller (and its variants) has a wide range of real-world applications.

Algorithm 1: Model Predictive Control (MPC_k)

```
Require: Specify the terminal costs $F_t$ for $k \leq t < T$.
for $t = 0, 1, \ldots, T-1$ do
    $t' \leftarrow \min\{t + k, T\}$
    Observe current state $x_t$ and obtain predictions $w_{t:t'|t}$.
    Solve and commit control action $u_t \coloneqq \psi_t^{t'}((x_t, w_{t:t'|t}); F_{t'})_{v_t}$.
end
```

Next, we give an overview of a novel analysis pipeline that converts a perturbation bound into a bound on the dynamic regret. We begin by highlighting the form of perturbation bounds required in the pipeline, and then describe the 3-step process of applying the pipeline. We apply this pipeline to obtain regret bounds for MPC in different settings.

Per-Step Error and Perturbation Bounds

A key challenge when comparing the performance of an online controller against the offline optimal trajectory is that the online controller's state x_t is different from the offline optimal state x_t^* at time step t. Due to such discrepancy in states, we cannot simply evaluate the online controller's action u_t via comparison against the offline optimal action u_t^*. To address this challenge, our pipeline uses the notion of per-step error (Definition 3.2.3) inspired by the performance difference lemma and its proofs in reinforcement learning (RL) (Lin, Hu, Shi, et al., 2021). Specifically, we compare u_t to the clairvoyant optimal action one may adopt at the same state x_t if all true future parameters $w_{t:T}^*$ are known, which leads to the definition of *per-step error* as follows.

Definition 3.2.3. *The per-step error e_t incurred by a predictive online controller* ALG *at time step t is defined as the distance between its actual action u_t and the clairvoyant optimal action, i.e.,*

$$e_t := \left\| u_t - \psi_t^T((x_t, w_{t:T}^*); F_T)_{u_t} \right\|, \text{ where } u_t = \mathsf{ALG}(x_t, w_{t:t+k|t}).$$

The clairvoyant optimal trajectory starting from x_t is defined as

$$x_{t:T|t}^* := \psi_t^T((x_t, w_{t:T}^*); F_T)_{x_{t:T}}.$$

Note that the clairvoyant optimal trajectory can be viewed as being generated by an MPC controller with long enough prediction horizon and exact predictions. This notion highlights the reason why MPC can compete against the clairvoyant optimal trajectory, since the per-step error in a system controlled by MPC_k becomes $e_t = \left\| \psi_t^{t+k}((x_t, w_{t:t+k|t}); F_{t+k})_{u_t} - \psi_t^T((x_t, w_{t:T}^*); F_T)_{u_t} \right\|$. Intuitively, the per-step error converges to zero as the prediction horizon k increases and the quality of predictions improves (i.e., $\left\| w_{t:t+k|t} - w_{t:t+k}^* \right\| \to 0$).

This intuition highlights the important role of perturbation bounds in comparing online controllers against (offline) clairvoyant optimal trajectories. As we have discussed in Section 3.1, the problem of establishing decaying perturbation bounds for different instances of the FTOCP (3.1) can be studied separately with the online implementation of predictive control. Perturbation bounds may take different forms, but for the application of our pipeline we require two types of perturbation bounds that are both common in the literature:

(a) *Perturbations of the parameters $w_{t_1:t_2}$ given a fixed initial state z_{t_1}:*

$$\left\| \psi_{t_1}^{t_2}\left((z_{t_1}, w_{t_1:t_2-1}, \zeta_{t_2}); F\right)_{u_{t_1}} - \psi_{t_1}^{t_2}\left((z_{t_1}, w'_{t_1:t_2-1}, \zeta'_{t_2}); F\right)_{u_{t_1}} \right\|$$

$$\leq \left(\sum_{t=t_1}^{t_2-1} q_1(t-t_1)\cdot \left\|w_t - w_t'\right\|\right)\left\|z_{t_1}\right\| + \sum_{t=t_1}^{t_2-1} q_2(t-t_1)\cdot\left\|w_t - w_t'\right\|$$
$$+ q_1(t_2 - t_1)\cdot\left\|\zeta_{t_2} - \zeta_{t_2}'\right\|\cdot\left\|z_{t_1}\right\| + q_2(t_2-t_1)\cdot\left\|\zeta_{t_2} - \zeta_{t_2}'\right\|, \tag{3.6}$$

where scalar functions q_1 and q_2 satisfy $\lim_{t\to\infty} q_i(t) = 0$, $\sum_{t=0}^{\infty} q_i(t) \leq C_i$ for constants $C_i \geq 1, i = 1, 2$. This perturbation bound is useful in bounding the per-step error e_t, as we will discuss in Lemma 3.2.1.

(b) *Perturbation of the initial state z_{t_1} given fixed parameters $w_{t_1:t_2}$:*

$$\left\|\psi_{t_1}^{t_2}\left((z_{t_1}, w_{t_1:t_2-1}, \zeta_{t_2}); F\right)_{(x_t,u_t)} - \psi_{t_1}^{t_2}\left((z_{t_1}', w_{t_1:t_2-1}, \zeta_{t_2}); F\right)_{(x_t,u_t)}\right\|$$
$$\leq q_3(t-t_1)\cdot\left\|z_{t_1} - z_{t_1}'\right\|, \text{ for } t \in [t_1, t_2], \tag{3.7}$$

where the scalar function q_3 satisfies $\sum_{t=0}^{\infty} q_3(t) \leq C_3$ for some constant $C_3 \geq 1$. This bound is useful in preventing the accumulation of per-step errors e_t throughout the horizon (see Lemma 3.2.2). Compared with (3.6), the right hand side of (3.7) has a simpler form.

Existing perturbation bounds usually combine the above two types ((3.6) and (3.7)) into a single equation that characterizes perturbations on z and $\xi_{t_1:t_2}$ simultaneously, e.g., Lin, Hu, Shi, et al., 2021; Shin and Zavala, 2021. Here, we decompose them into two separate types because they are used in different parts of our pipeline.

A 3-Step Pipeline from Perturbation Bounds to Regret

An overview of the pipeline is given in Figure 3.3, which illustrates the high-level ideas of the pipeline that starts by obtaining perturbation bounds, proceeds to bound the per-step error using perturbation bounds, and finally combines the per-step error and perturbation bounds to bound the dynamic regret. In the following we describe each step in detail.

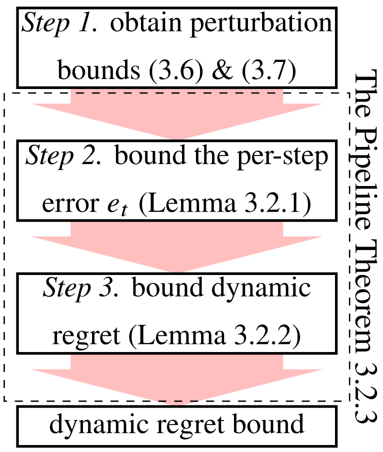
Step 1. obtain perturbation bounds (3.6) & (3.7)
Step 2. bound the per-step error e_t (Lemma 3.2.1)
Step 3. bound dynamic regret (Lemma 3.2.2)
dynamic regret bound
The Pipeline Theorem 3.2.3

Figure 3.3: Illustrative diagram of the 3-step pipeline from perturbation analysis to bounded regret.

Step 1: Obtain the perturbation bounds given in (3.6) and (3.7). The form of the perturbation bounds depends heavily on the specific form of the FTOCP, and thus the derivation requires case-by-case study (e.g., see Section 3.1). However, off-the-shelf bounds are available in most cases, as there has been a rich literature on perturbation analysis of control systems (e.g., Xu and Anitescu, 2019; Na and Anitescu, 2022; Shin, Zavala, and Anitescu, 2020; Shin and Zavala, 2021; Lin, Hu, Shi, et al., 2021 and the references therein). The following property summarizes precisely what is expected to be derived for bounds (3.6) and (3.7) in Steps 2 and 3.

Property 3.2.1. *Suppose there exists a positive constant R such that the perturbation bound* (3.6) *holds for the following specifications: with $t_1 = t$ and $t_2 = t + k$ for $t < T - k$,* (3.6) *holds for $F : \mathbb{R}^n \to \mathbb{R}^n$ be the identity function $\mathbb{I}$, and*

$$z_t \in \mathcal{B}(x_t^*, R);\ w_{t:t+k-1} \in \mathcal{W}_{t:t+k-1}, w'_{t:t+k-1} = w^*_{t:t+k-1};$$
$$\zeta_{t+k}, \zeta'_{t+k} \in \mathcal{B}(x^*_{t+k}, R) \subseteq \mathbb{R}^n;$$

with $t_1 = t$ and $t_2 = T$ for $t \geq T - k$, (3.6) *holds for $z_t \in \mathcal{B}(x_t^*, R)$; $w_{t:T} \in \mathcal{W}_{t:T}, w_{t:T} = w^*_{t:T}$; $F = F_T$. Further, perturbation bound* (3.7) *holds for any $z_{t_1}, z'_{t_1} \in \mathcal{B}(x^*_{t_1}, R)$ and $w_{t_1:t_2} = w^*_{t_1:t_2}$.*

Intuitively, Property 3.2.1 states that perturbation bounds (3.6) and (3.7) hold in a small neighborhood (specifically, a ball with radius R) around the offline optimal trajectory OPT, which is much weaker than the global exponentially decaying perturbation bounds required by previous work (e.g., Lin, Hu, Shi, et al., 2021) in the following sense: (i) in the general settings where the dynamical function g_t is non-linear, or where there are constraints on states and actions, one cannot hope the perturbation bound to hold globally for all possible parameters (Shin, Anitescu, and Zavala, 2022; Shin and Zavala, 2021; Na and Anitescu, 2022); (ii) the decay functions $\{q_i\}_{i=1,2,3}$ are only required to converge to zero and satisfy $\sum_{\tau=0}^{\infty} q_i(\tau) \leq C_i$, which means the exponential decay rate as in Lin, Hu, Shi, et al., 2021 is not necessary — in fact, polynomial decay rates can also satisfy these properties, which greatly broadens the applicability of our pipeline.

Step 2: Bound the per-step error e_t. The core of the analysis is to apply the perturbation bounds to bound the per-step error. For MPC_k, under Property 3.2.1, this step can be done in a universal way, as summarized in Lemma 3.2.1 below. A complete proof of Lemma 3.2.1 can be found in Section 3.B.

Lemma 3.2.1. *Let Property 3.2.1 hold. Suppose the current state x_t satisfies $x_t \in \mathcal{B}(x_t^*, R/C_3)$ and the terminal cost F_{t+k} of MPC_k is set to be the indicator function of some state $\bar{x}(w_{t+k|t})$ that satisfies $\bar{x}(w_{t+k|t}) \in \mathcal{B}(x^*_{t+k}, R)$ for $t < T - k$.*

Then, the per-step error of MPC_k *is bounded by*

$$e_t \le \sum_{\tau=0}^{k} \left(\left(\frac{R}{C_3} + D_{x^*} \right) \cdot q_1(\tau) + q_2(\tau) \right) \rho_{t,\tau} + 2R \left(\left(\frac{R}{C_3} + D_{x^*} \right) \cdot q_1(k) + q_2(k) \right). \tag{3.8}$$

Lemma 3.2.1 is a straight-forward implication of perturbation bound (3.6) specified in Property 3.2.1. To see this, for $t < T - k$, note that the per-step error e_t can be bounded by

$$e_t = \left\| \psi_t^{t+k}(x_t, w_{t:t+k-1|t}, \bar{x}(w_{t+k|t}); \mathbb{I})_{u_t} - \psi_t^T(x_t, w_{t:T}^*; F_T)_{u_t} \right\| \tag{3.9a}$$

$$= \left\| \psi_t^{t+k}(x_t, w_{t:t+k-1|t}, \bar{x}(w_{t+k|t}); \mathbb{I})_{u_t} - \psi_t^{t+k}(x_t, w_{t:t+k-1}^*, x_{t+k|t}^*; \mathbb{I})_{u_t} \right\| \tag{3.9b}$$

$$\begin{aligned} &\le \sum_{\tau=0}^{k-1} \left(\|x_t\| \cdot q_1(\tau) + q_2(\tau) \right) \rho_{t,\tau} \\ &\quad + \left(\|x_t\| \cdot q_1(k) + q_2(k) \right) \left\| \bar{x}(w_{t+k|t}) - x_{t+k|t}^* \right\|. \end{aligned} \tag{3.9c}$$

Here, we apply the principle of optimality to conclude that the optimal trajectory from x_t to $x_{t+k|t}^*$ (i.e., $\psi_t^{t+k}(x_t, w_{t:t+k-1}^*, x_{t+k|t}^*; \mathbb{I})$ in (3.9b)) is a sub-trajectory of the clairvoyant optimal trajectory from x_t (i.e., $\psi_t^T(x_t, w_{t:T}^*; F_T)$ in (3.9a)), and (3.9c) is obtained by directly applying perturbation bound (3.6). Note that $\|x_t\| \le \frac{R}{C_3} + D_{x^*}$, and that both $\bar{x}(w_{t+k|t})$ and $x_{t+k|t}^*$ are in $\mathcal{B}(x_{t+k}^*; R)$ by assumption and by perturbation bound (3.7) specified in Property 3.2.1, we conclude that (3.8) hold for $t < T-k$. The case $t \ge T - k$ can be shown similarly. We defer the detailed proof to Section 3.B.

Step 3: Bound the dynamic regret by $\sum_{t=0}^{T-1} e_t^2$. This final step builds upon perturbation bound (3.7), and aims at deriving dynamic regret bounds in a universal way, as stated in Lemma 3.2.2 below. Specifically, under the assumption that a local decaying perturbation bound in the form of (3.7) holds around the offline optimal trajectory OPT, and the property that per-step errors e_t are sufficiently small, we can show that the online controller will not leave the "safe region" near the offline optimal trajectory as specified in Property 3.2.1, and thus the dynamic regret of ALG is bounded as in (3.10) (note that ALG is not confined to MPC, but is allowed to be any algorithm with bounded per-step errors). We provide an intuitive illustration in Figure 3.4. A complete proof of Lemma 3.2.2 can be found in Section 3.B.

Lemma 3.2.2. *Let Property 3.2.1 hold. If the per-step errors of* ALG *satisfy* $e_\tau \le R/(C_3^2 L_g)$ *for all time steps* $\tau < t$*, the trajectory of* ALG *will remain close to* OPT *at*

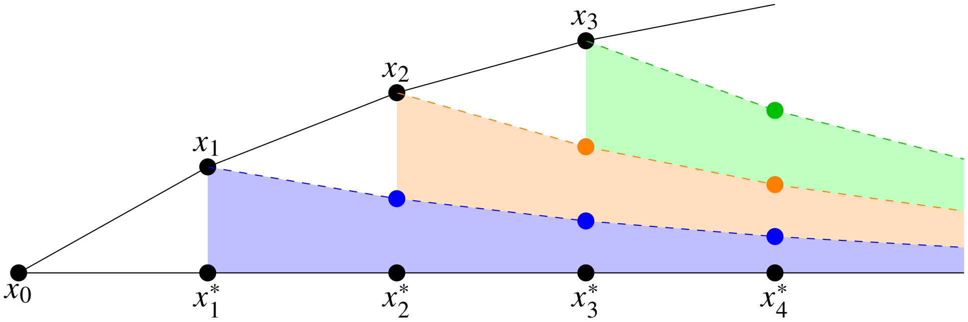

Figure 3.4: Illustration of the per-step error accumulation over time.

time t, i.e., $x_t \in \mathcal{B}(x_t^*, R/C_3)$. Further, if $e_t \le R/(C_3^2 L_g)$ for all $t < T$, the dynamic regret of ALG *is upper bounded by*

$$\mathrm{cost}(\mathsf{ALG}) - \mathrm{cost}(\mathsf{OPT}) = O\left(\sqrt{\mathrm{cost}(\mathsf{OPT}) \cdot \sum_{t=0}^{T-1} e_t^2} + \sum_{t=0}^{T-1} e_t^2\right). \tag{3.10}$$

Summary. Combining Steps 2 and 3 of the pipeline yields the following *Pipeline Theorem* for MPC_k (see Theorem 3.2.3). Basically it states that, when the prediction horizon k is sufficiently large and the prediction errors $\rho_{t,\tau}$ are sufficiently small, Lemma 3.2.1 and Lemma 3.2.2 can work together to make sure that MPC_k never leaves a (R/C_3)-ball around the offline optimal trajectory OPT; thus we obtain a dynamic regret bound.

Theorem 3.2.3 (The Pipeline Theorem)**.** *Let Property 3.2.1 hold. Suppose the terminal cost F_{t+k} of MPC_k is set to be the indicator function of some state $\bar{x}(w_{t+k|t})$ that satisfies $\bar{x}(w_{t+k|t}) \in \mathcal{B}(x_{t+k}^*, R)$ for all time steps $t < T - k$. Further, suppose the prediction errors $\rho_{t,\tau}$ are sufficiently small and the prediction horizon k is sufficiently large, such that*

$$\sum_{\tau=0}^{k}\left(\left(\frac{R}{C_3} + D_{x^*}\right) \cdot q_1(\tau) + q_2(\tau)\right)\rho_{t,\tau} + 2R\left(\left(\frac{R}{C_3} + D_{x^*}\right) \cdot q_1(k) + q_2(k)\right) \le \frac{R}{C_3^2 L_g}.$$

Then, the trajectory of MPC_k will remain close to OPT*, i.e., $x_t \in \mathcal{B}(x_t^*, R/C_3)$ for all time steps t, and the dynamic regret of MPC_k is upper bounded by*

$$\mathrm{cost}(\mathsf{MPC}_k) - \mathrm{cost}(\mathsf{OPT}) = O\left(\sqrt{\mathrm{cost}(\mathsf{OPT}) \cdot E} + E\right), \tag{3.11}$$

where $E \coloneqq \sum_{\tau=0}^{k-1} (q_1(\tau) + q_2(\tau))\, P(\tau) + \left(q_1(k)^2 + q_2(k)^2\right) T$.

The proof of Theorem 3.2.3 can be found in Section 3.B. To interpret the dynamic regret bound in (3.11), note that we have $\mathrm{cost}(\mathsf{OPT}) = O(T)$ as a result of our model assumptions. Thus, the dynamic regret of ALG is in the order of $\sqrt{TE} + E$. When there is no prediction error, the regret bound $O((q_1(k) + q_2(k)) \cdot T)$ reproduces the result in Lin, Hu, Shi, et al., 2021, and the bound will degrade as the prediction error increases. It is also worth noticing that, when the prediction power improves over time as the online controller learns the system better and $k = \Omega(\ln T)$, the dynamic regret can be $o(T)$.

Instantiation: Unconstrained LTV Systems

In this section, we consider the following special case of problem (3.5), where the dynamics is LTV and the prediction error can only occur on the disturbances w_t:

$$\begin{aligned} \min_{x_{0:T}, u_{0:T-1}} & \sum_{t=0}^{T-1} \left(f_t^x(x_t) + f_t^u(u_t) \right) + F_T(x_T) \\ \text{s.t. } & x_{t+1} = A_t x_t + B_t u_t + w_t^*, \qquad \forall 0 \le t < T, \qquad (3.12) \\ & x_0 = x(0). \end{aligned}$$

We summarize all necessary assumptions below in Assumption 3.2.1.

Assumption 3.2.1. *Assume the following holds for the online control problem instance* (3.12)*:*

- *Cost functions:* $\{f_t^x\}_{t=0}^{T-1}, \{f_t^u\}_{t=0}^{T-1}, F_T$ *are nonnegative* μ*-strongly convex and* ℓ*-smooth. And we assume* $f_t^x(0) = f_t^u(0) = F_T(0) = 0$ *without the loss of generality.*
- *Dynamical systems: The LTV system* $\{A_t, B_t\}$ *is* σ*-uniform controllable with controllability index* d [1]*, and* $\|A_t\| \le a$, $\|B_t\| \le b$, *and* $\|B_t^\dagger\| \le b'$ *hold for all* t*, where* $B_t^\dagger$ *denotes the Moore–Penrose inverse of matrix* B_t*.*
- *Predicted quantities:* $\|w_t\| \le D_w$ *holds for all* $w_t \in \mathcal{W}_t$ *and all* t*.*

Under Assumption 3.2.1, we can again apply the perturbation bounds shown in Lin, Hu, Shi, et al., 2021 to show Property 3.2.1. In particular, we already know that for some constants $H_1 \ge 1$ and $\lambda_1 \in (0, 1)$, perturbation bounds (3.6) and (3.7) hold globally for $q_1(t) = 0$, $q_2(t) = H_1 \lambda_1^t$, and $q_3(t) = H_1 \lambda_1^t$. Since both of these perturbation bounds hold globally, radius R in Property 3.2.1 can be set arbitrarily, and we shall take $R \coloneqq \max\left\{D_{x^*}, \frac{2L_g H_1^3}{(1-\lambda_1)^3}\right\}$ so that Theorem 3.2.3 can be applied to MPC_k with terminal cost $F_{t+k}(\cdot; w_{t+k|t}) \equiv \mathbb{I}(\cdot; 0)$. This leads to the following dynamic regret bound:

[1] Uniform controllability is defined in Assumption 3.1.2.

Theorem 3.2.4. *In the unconstrained LTV setting* (3.12)*, under Assumption 3.2.1, when the prediction horizon k is sufficiently large such that* $k \geq \ln\left(\frac{4H_1^3 L_g}{(1-\lambda_1)^2}\right)/\ln(1/\lambda_1)$*, the dynamic regret of* MPC_k *(Algorithm 1) with terminal cost* $F_{t+k}(\cdot; w_{t+k|t}) \equiv \mathbb{I}(\cdot; 0)$ *is bounded by*

$$\mathrm{cost}(\mathsf{MPC}_k) - \mathrm{cost}(\mathsf{OPT}) \leq O\left(\sqrt{T \cdot \sum_{\tau=0}^{k-1} \lambda_1^\tau P(\tau) + \lambda_1^{2k} T^2} + \sum_{\tau=0}^{k-1} \lambda_1^\tau P(\tau)\right).$$

A complete proof of Theorem 3.2.4 can be found in Appendix 3.B. When there are no prediction errors, the bound in Theorem 3.2.4 reduces to $O(\lambda_1^k T)$, which reproduces the result of Lin, Hu, Shi, et al., 2021. Further, it is also worth noticing that due to the form of discounted sum $\sum_{\tau=0}^{k-1} \lambda_1^\tau P(\tau)$, prediction errors for the near future matter more than those for the far future.

3.3 Application: Networked Online Convex Optimization

We consider online optimization in a networked system where each nodes individually decides on an action at each period and the objective is to minimize a global cost over a finite time horizon k. Specifically, we use a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the network where $\mathcal{V}$ denotes the set of nodes. Two nodes v and u interact with each other if and only if they are connected by an undirected edge $(v, u) \in \mathcal{E}$. At each period $t = 1, 2, \ldots, H$, each node v picks an n-dimensional local action $x_t^v \in D_t^v$, where n is a positive integer and $D_t^v \subset \mathbb{R}^n$ is a convex set of feasible actions. The global action at period t is the vector of all local actions $x_t = \{x_t^v\}_{v \in \mathcal{V}}$, and incurs a global cost, which is the sum of three types of local cost functions:

- **Node costs**: Each node v incurs a time-varying node cost $f_t^v(x_t^v)$, which characterizes the local preference for its local action x_t^v.
- **Temporal interaction costs**: Each node v incurs a time-varying temporal interaction cost $c_t^v(x_t^v, x_{t-1}^v)$, that characterizes how its previous local action interacts with the current one.
- **Spatial interaction costs**: Each pair of nodes (v, u) over an edge $e \in \mathcal{E}$ incurs a time-varying spatial interaction cost[2] $s_t^e(x_t^v, x_t^u)$. This characterizes how their local actions affect each other.

In our model, the node cost is the part of the cost that only depends on a node's current local action. If the other two types of costs are zero, each node will trivially pick the

[2] Since e is an undirected edge, the order in which we write the two inputs (the action of v and the action of u) does not matter. Note that s_t^e can be asymmetric for nodes v and u, e.g., $s_t^e(x_t^v, x_t^u) = s_t^e(x_t^u, x_t^v) = \left\|x_t^v + 2x_t^u\right\|^2$.

minimizer of its node cost. Temporal interaction costs encourage the local actions at each period to be "compatible" with the previous local actions. For example, a temporal interaction could be a switching cost that penalizes large deviations from the previous action in order to make the trajectory of local actions "smooth." Such switching costs can be found in work on single-node online optimization, e.g., Chen, Goel, and Wierman (2018), Goel, Lin, et al. (2019), and Lin, Goel, and Wierman (2020). In multi-product pricing, a general switching cost can be used to capture the impact of the previous price on the current demand. Spatial interaction costs, on the other hand, can be used to enforce some collective behavior among the nodes. For example, spatial interaction can model the probability that one node's actions affect its neighbor's actions in diffusion processes on social networks (Kempe, Kleinberg, and Tardos, 2015); or model interactions between complement/substitute products in multiproduct pricing (Candogan, Bimpikis, and Ozdaglar, 2012).

To this point, we summarize all necessary elements that define a specific instance of the Networked OCO problem in Definition 3.3.1. This is useful later for defining the class of Networked OCO problem we study and defining the performance metrics rigorously.

Definition 3.3.1. *An instance of the Networked OCO problem is characterized by a tuple with 4 entries, $(\mathcal{G}, H, x_0, \{f_t^v, c_t^v, s_t^e, D_t^v\}_{t\in[H], v\in\mathcal{V}, e\in\mathcal{E}})$, that contains the graph, the number of periods, the initial actions, and the set of all local cost functions/constraint sets.*

To study any instance of the Networked OCO problem, it is useful to separate the global cost at each period into two parts based on whether the cost term depends only on the current global action or whether it also depends on the previous action. Specifically, the part that depends only on the current global action x_t is the sum of all node costs and spatial interaction costs. We refer to this component as the *(global) hitting cost* and denote it as

$$f_t(x_t) \coloneqq \sum_{v\in\mathcal{V}} f_t^v(x_t^v) + \sum_{(v,u)\in\mathcal{E}} s_t^{(v,u)}(x_t^v, x_t^u).$$

The rest of the global cost involves the current global action x_t and the previous global action x_{t-1}. We refer to it as the *(global) switching cost* and denote it as

$$c_t(x_t, x_{t-1}) \coloneqq \sum_{v\in\mathcal{V}} c_t^v(x_t^v, x_{t-1}^v).$$

Given an instance p of the Networked OCO problem, the objective of the decentralized online algorithm is to minimize the total global stage costs in a finite horizon H starting from a given initial global action x_0 at period 0: $\text{cost}_p(\texttt{ALG}) \coloneqq \sum_{t=1}^{H} \left(f_t(x_t) + c_t(x_t, x_{t-1})\right)$, where ALG denotes any decentralized online algorithm used to solve the Networked OCO problem. The offline optimal cost is the clairvoyant minimum cost one can incur on the same sequence of cost functions and the initial global action x_0 at time step 0, i.e., $\text{cost}_p(\texttt{OPT}) \coloneqq \min_{x_{1:H}} \sum_{t=1}^{H} \left(f_t(x_t) + c_t(x_t, x_{t-1})\right)$.

We measure the performance of any online algorithm ALG by the competitive ratio (CR), which is a widely-used metric in the literature of online optimization, e.g., Chen, Goel, and Wierman (2018), Goel, Lin, et al. (2019), and Argue, Gupta, and Guruganesh (2020). The competitive ratio is defined as the worst-case ratio between $\text{cost}_p(\texttt{ALG})$ and $\text{cost}_p(\texttt{OPT})$ over a class of problem instances.

Definition 3.3.2. *The competitive ratio of a given online algorithm* ALG *for a class of problem instances $\mathcal{P}$ is the supremum of* $\text{cost}(\mathsf{ALG})/\text{cost}(\mathsf{OPT})$ *over problem instances in class $\mathcal{P}$, i.e.,*

$$CR_{\mathcal{P}}(\mathsf{ALG}) \coloneqq \sup_{p \in \mathcal{P}} \text{cost}_p(\mathsf{ALG})/\text{cost}_p(\mathsf{OPT}).$$

Finally, we define the partial hitting and switching costs over subsets of the nodes. In particular, for a subset of nodes $S \subseteq \mathcal{V}$, we denote the joint action over S as $x_t^S \coloneqq \{x_t^v \mid v \in S\}$ and define the partial hitting cost and partial switching cost over S as

$$\begin{aligned} f_t^S(x_t^{S_+}) &\coloneqq \sum_{v \in S} f_t^v(x_t^v) + \sum_{(v,u) \in \mathcal{E}(S_+)} s_t^{(v,u)}(x_t^v, x_t^u), \\ c_t^S(x_t^S, x_{t-1}^S) &\coloneqq \sum_{v \in S} c_t^v(x_t^v, x_{t-1}^v). \end{aligned} \tag{3.13}$$

This notation is useful for presenting decentralized online algorithms where the optimizations are performed over the r-hop neighborhood of each node.

Predictions and Locality

We assume that each node has access to local cost functions up to a *prediction horizon* k into the future, for themselves and their neighborhood up to an *observation radius* r. In more detail, recall that N_v^r denotes the r-hop neighborhood of a node v, i.e., $N_v^r \coloneqq \{u \in \mathcal{V} \mid d_{\mathcal{G}}(u, v) \leq r\}$. To pick a local action x_t^v at period t, node v can use k steps of future node costs, temporal interaction costs, and spatial interaction costs

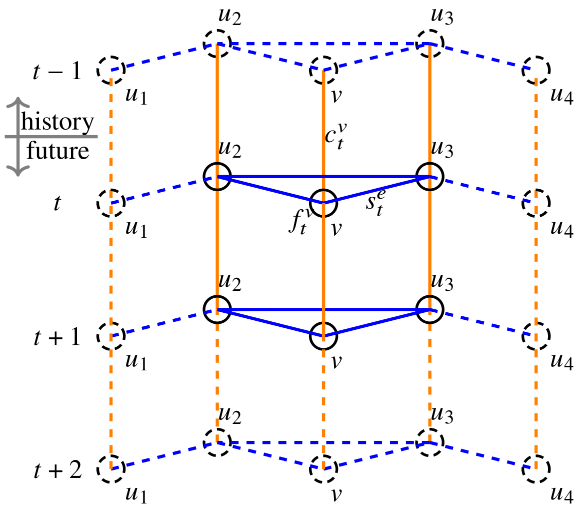

Figure 3.5: Illustration of available information for agent v at period t with $k = 2$ and $r = 1$, for the network with $\mathcal{V} = \{u_1, u_2, v, u_3, u_4\}$ and $\mathcal{E} = \{(u_1, u_2), (u_2, v), (v, u_3), (u_2, u_3), (u_3, u_4)\}$).

within its r-hop neighborhood, $\{\{(f_{\tau|t}^u, c_{\tau|t}^u) \mid u \in N_v^r\}, \{s_{\tau|t}^e \mid e \in \mathcal{E}(N_v^r)\}\}_{t \le \tau < t+k}$, and the previous local actions in N_v^r: $\{x_{t-1}^u \mid u \in N_v^r\}$. Here, $f_{\tau|t}^u$, $c_{\tau|t}^u$, and $s_{\tau|t}^e$ denote the best predictions for the future true local cost functions $f_\tau^u, c_\tau^u, s_\tau^e$ $(\tau > t)$ that we can make at the current period t.

We provide an illustration of the local cost functions known to node v at period t in Figure 3.5. In the figure, the black circles, blue lines, and orange lines denote the node costs, temporal interaction costs, and spatial interaction costs, respectively. The known functions are marked by solid lines. Note that, in addition to the local cost functions, node v also knows the local actions in N_v^r at period $t-1$, which are not illustrated in the figure.

To ease the presentation, we first focus on the case when the k-step predictions of cost functions are *exact*. Specifically, exact predictions mean $f_{\tau|t}^u = f_\tau^u, c_{\tau|t}^u = c_\tau^u, s_{\tau|t}^e = s_\tau^e$ for all $t \le \tau < t+k$ and $u \in \mathcal{V}, e \in \mathcal{E}$. Then, we discuss how to model the prediction errors when the predictions are inexact and how they affect the performance of the proposed algorithm.

Localized Predictive Control (LPC)

The design of LPC is inspired by the classical model predictive control (MPC) framework (García, Prett, and Morari, 1989), which leverages all available information at the current period to decide the current local action "greedily." In our context, when a node v decides its action x_t^v at time t, the available information includes previous local actions in the r-hop neighborhood and k-period predictions of all local node costs and temporal/spatial interaction costs. The boundaries of all available information, which are formed by $\{t-1\} \times N_v^r$ and $\partial N_{(t,v)}^{(k,r)}$, are illustrated in Figure 3.6.

The pseudocode for LPC is presented in Algorithm 2. For each node v at period t, LPC fixes the actions on the boundaries of available information and then solves for the optimal actions inside the boundaries. Specifically, for an instance $p = (\mathcal{G}, H, x_0, \{f_t^v, c_t^v, s_t^e, D_t^v\}_{t \in [H], v \in \mathcal{V}, e \in \mathcal{E}})$ of the Networked OCO problem, define $\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u \mid u \in N_v^r\}, \{z_\tau^u \mid (\tau, u) \in \partial N_{(t,v)}^{(k,r)}\}\right)$ as the optimal solution of the problem [3]

$$\begin{aligned} \min \sum_{\tau=t}^{t+k-1} & \left(f_\tau^{(N_v^{r-1})}\left(x_\tau^{(N_v^r)}\right) + c_\tau^{(N_v^r)}\left(x_\tau^{(N_v^r)}, x_{\tau-1}^{(N_v^r)}\right)\right) \\ \text{s.t. } & x_{t-1}^u = y_{t-1}^u, \forall u \in N_v^r, \\ & x_\tau^u = z_\tau^u, \forall (\tau, u) \in \partial N_{(t,v)}^{(k,r)}, \\ & x_\tau^u \in D_\tau^u, \forall (\tau, u) \in N_{(t,v)}^{(k-1,r-1)}, \end{aligned} \tag{3.14}$$

where the partial hitting cost and partial switching cost f_τ^S and c_τ^S for a subset S of nodes were defined in (3.13). When the context is clear, we use the shorthand $\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \{z_\tau^u\}\right)$. Note that $\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \{z_\tau^u\}\right)$ is a matrix of actions (in $\mathbb{R}^n$) indexed by $(\tau, u) \in N_{p,(t,v)}^{(k-1,r-1)}$. Once the parameters $\{y_{t-1}^u\}$ and $\{z_\tau^u\}$ are fixed, the node v can leverage its local predictions to solve the optimization problem in (3.14).

LPC fixes the parameters $\{y_{t-1}^u\}$ to be $\{x_{t-1}^u\}$, which are the previous local actions in N_v^r, and fixes the parameters $\{z_\tau^u\}$ to be the minimizers of the predicted local node cost functions at nodes in $\partial N_{(t,v)}^{(k,r)}$. The selection of the parameters at nodes in $\partial N_{(t,v)}^{(k,r)}$ plays a similar role as the terminal cost of classical MPC in single-node settings.

For a single-node system, MPC-style algorithms are perhaps the most prominent approach for optimization-based control (García, Prett, and Morari, 1989) because of their simplicity and excellent performance in practice. LPC extends the ideas of MPC to a decentralized setting in a networked system by leveraging available

[3] To simplify notation, in cases when the prediction horizon exceeds the whole horizon length H, we adopt the convention that $f_t^v(x_t^v) = \frac{\mu}{2}\|x_t^v\|^2$, $c_t^v \equiv s_t^e \equiv 0$ and $D_t^v = \mathbb{R}^n$ for $t > H$, where μ is the strongly convexity coefficient defined in Assumption 3.3.1. These extended definitions do not affect our original problem with horizon H. Note that every node has access to exact predictions of the local cost functions with $t > H$ with this convention.

predictions in both the temporal and spatial dimensions, whereas classical MPC focuses only on the temporal dimension. This change makes our algorithm simple and practical for applications including multi-product pricing but also leads to significant technical challenges in the analysis. For ease of presentation, we first study the case when all the predictions are *exact* and discuss how to generalize the results to include inexact predictions.

Algorithm 2: Localized Predictive Control (for node v)

```
Parameters: Prediction horizon $k$ and observation radius $r$.
for $t = 1$ to $H$ do
    Receive information $\{x_{t-1}^u \mid u \in N_v^r\}$ and the predictions
        $\{\{(f_{\tau|t}^u, c_{\tau|t}^u) \mid u \in N_v^r\}, \{s_{\tau|t}^e \mid e \in \mathcal{E}(N_v^r)\}\}_{t \le \tau < t+k}.$    (3.15)
    Solve the optimization problem (3.14) with the predicted local cost
      functions in (3.15):
        $\psi_{(t,v)}^{(k,r)}\left(\{x_{t-1}^u \,|\, u \in N_v^r\}, \left\{\theta_\tau^u \,|\, (\tau, u) \in \partial N_{(t,v)}^{(k,r)}\right\}\right),$
      where $\theta_\tau^u := \arg\min_{y \in D_\tau^u} f_{\tau|t}^u(y)$.
    Choose local action $x_t^v$ to be the $(t, v)$-th element in the solution.
end
```

Our analysis is based on standard smoothness and convexity assumptions on the local cost functions:

Assumption 3.3.1. *For $\mu > 0, \ell_f < \infty, \ell_T < \infty, \ell_S < \infty$, the local cost functions and feasible sets in an instance $(\mathcal{G}, H, x_0, \{f_t^v, c_t^v, s_t^e, D_t^v\}_{t \in [H], v \in \mathcal{V}, e \in \mathcal{E}})$ of the Networked OCO problem satisfy:*

- *$f_t^v : \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ is μ-strongly convex, ℓ_f-smooth, and in C^2;*
- *$c_t^v : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ is convex, ℓ_T-smooth, and in C^2;*
- *$s_t^e : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\ge 0}$ is convex, ℓ_S-smooth, and in C^2;*
- *$D_t^v \subseteq \mathbb{R}^n$ satisfies $int(D_t^v) \neq \emptyset$ and can be written as $D_t^v := \{x_t^v \in \mathbb{R}^n \mid (g_t^v)_i(x_t^v) \le 0, \forall 1 \le i \le m_t^v\}$, where each $(g_t^v)_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function in C^2.*

Intuitively, under Assumption 3.3.1, the Networked OCO problem becomes easier as the coefficient μ increases. This is because a larger μ encourages each node to choose its local action closer to the minimizer of the node cost function f_t^v, which makes every node more "independent" from each other. In contrast, the problem becomes more challenging as the coefficient ℓ_T increases, because that strengthens

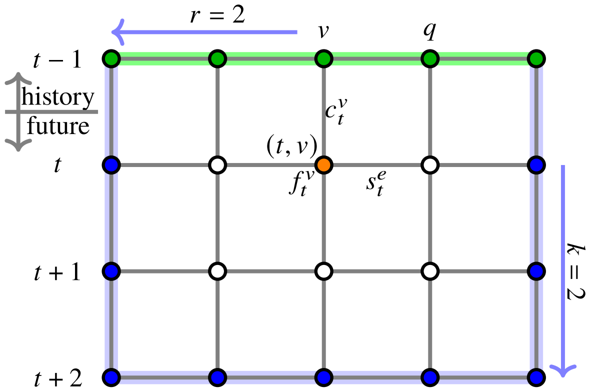

Figure 3.6: Illustration of LPC with $k = 3, r = 2$ on a line graph (the underlying graph is replicated over the time dimension). The orange node marks the decision variable at (t, v). The green part denotes the actions in N_v^r at period $(t-1)$. The blue "U" shape denotes the boundary of available predictions for node v at period t.

the need to "coordinate" between the decisions at different periods. Similarly, increasing the coefficient ℓ_S or the maximum degree Δ of each node also makes the problem harder by requiring more coordination between different nodes. Therefore, the competitive ratio bounds that we derive depend on the quantities ℓ_T/μ and $\Delta\ell_S/\mu$, which characterize the difficulty of a class of Networked OCO problems.

We define the set of possible configurations for Networked OCO Υ as

$$\{(\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \mid \mu \in \mathbb{R}_{>0}, \ell_f \in \mathbb{R}_{>0}, \ell_T \in \mathbb{R}_{\geq 0}, \ell_S \in \mathbb{R}_{\geq 0}, \Delta \in \mathbb{N}, h : \mathbb{N} \to \mathbb{N}\}.$$

Each configuration tuple in Υ specifies a class of Networked OCO problems for which we study the competitive ratio of LPC, and this relationship is defined formally in Definition 3.3.3.

Definition 3.3.3. *For any configuration tuple $(\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$, we define $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$ as the set of problem instances of Networked OCO $p = (\mathcal{G}, H, x_0, \{f_t^v, c_t^v, s_t^e, D_t^v\}_{t\in[H], v\in\mathcal{V}, e\in\mathcal{E}})$ that satisfy:*

1. *Assumption 3.3.1 with $(\mu, \ell_f, \ell_T, \ell_S)$;*
2. *degree$(v) \leq \Delta$ for every node $v \in \mathcal{V}$;*
3. *$\left|\partial N_v^r\right| \leq h(r)$ for every node $v \in \mathcal{V}$ and $r \in \mathbb{N}$.*

Before presenting our main results in the most general form, we first provide two examples that instantiate our competitive ratio bounds for LPC under specific parameters. The first example, Corollary 3.3.1, is a special instance of our main result

(Theorem 3.3.5) when ℓ_T/μ and $\Delta\ell_S/\mu$ are bounded by some specific constants. A formal proof can be found in Section 3.D.

Corollary 3.3.1. *For any tuple $\upsilon = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$ that satisfies $\Delta \geq 3$, $\ell_T/\mu \leq 0.88$, $\Delta\ell_S/\mu \leq 0.28$, and $h(\gamma) \leq C \cdot 2^{\gamma/2}$ for some constant $C < \infty$, if the prediction horizon k and observation radius r are sufficiently large such that $\left(14 \cdot 2^{-3r} + 234 \cdot 2^{-4k}\right) C^2 \leq \frac{1}{2}$, then the competitive ratio of LPC for the problem class $\mathcal{P}(\nu)$, denoted as $CR_{\mathcal{P}(\upsilon)}(LPC)$, is bounded above by*

$$1 + \left(1 + C_1' \cdot \frac{\ell_f + \Delta\ell_S + 2\ell_T}{\mu} \cdot C^2\right) \cdot 2^{-r} + \left(2 + C_2' \cdot \frac{\ell_f + \Delta\ell_S + 2\ell_T}{\mu} \cdot C^2\right) \cdot 2^{-k},$$

where $C_1' < \infty$ and $C_2' < \infty$ are numerical constants.

Corollary 3.3.1 shows that, if we increase the prediction horizon k and observation radius r simultaneously, the competitive ratio of LPC improves exponentially to 1 with a decay factor of $\frac{1}{2}$. Besides the constant upper bounds on ℓ_T/μ and $\Delta\ell_S/\mu$, this corollary also requires the boundary of any γ-hop neighborhood to not grow too fast (i.e., $h(\gamma) = O(2^{\gamma/2})$). Note that any graph such that $h(\gamma) = \text{poly}(\gamma)$ satisfies this assumption. Corollary 3.3.1 is a special case of our general result, Theorem 3.3.5, which holds for any $\ell_T/\mu, \Delta\ell_S/\mu$, and $h(\gamma)$ (without the constraints), with decay factors which (instead of 1/2) are functions of ℓ_T/μ and $\Delta\ell_S/\mu$. We note that the decay factor is smaller than $\frac{1}{2}$ for small enough ℓ_T/μ and $\Delta\ell_S/\mu$.

Our second example, Corollary 3.3.2, studies the dependence of the exponential decay factor on the quantities ℓ_T/μ and $\Delta\ell_S/\mu$ as they approach zero. A formal proof can be found in Section 3.D.

Corollary 3.3.2. *For any tuple $\upsilon = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$ that satisfies $\Delta \geq 3$, $\ell_T/\mu \leq \frac{1}{16}$, $\ell_S/\mu \leq \Delta^{-7}$, if the prediction horizon k and observation radius r are sufficiently large that $\Delta^{-6r} + 4 \cdot 2^{-12k} \leq \frac{1}{256}$, then the competitive ratio of LPC for the problem class $\mathcal{P}(\nu)$, denoted as $CR_{\mathcal{P}(\upsilon)}(LPC)$, is bounded above by*

$$1 + \left(1 + C_3' \cdot \frac{\ell_f + \Delta\ell_S + 2\ell_T}{\mu}\right) \cdot \left(\frac{\Delta^3\ell_S}{\mu}\right)^{\frac{r}{2}} + \left(2 + C_4' \cdot \frac{\ell_f + \Delta\ell_S + 2\ell_T}{\mu}\right) \cdot \left(\frac{8\ell_T}{\mu}\right)^k,$$

where C_3' and C_4' are numerical constants.

Corollary 3.3.2 shows that the constraint on $h(\gamma)$ can be relaxed when if the quantities ℓ_T/μ and $\Delta\ell_S/\mu$ are sufficiently small. Further, the decay factor of the competitive ratio bound tends to zero if Δ is a fixed constant and both ℓ_T/μ and ℓ_S/μ tend to zero. While we simplify the expression in Corollary 3.3.2 by adopting the worst-case upper bound on $h(\gamma)$, i.e., $h(\gamma) \leq \Delta^{\gamma}$, our general result in Theorem 3.3.4 considers general $h(\gamma)$, which can give tighter bounds for the decay factors.

Perturbation Analysis for Networked OCO

The key idea underlying our analysis of LPC is that the impact of perturbations to the actions at the boundaries of the available predictions of a node decays quickly, in fact exponentially fast, in the distance of the boundary from the node. This quick decay means that small errors cannot build up to hurt algorithm performance.

In this section, we formally study such perturbations by deriving several new results that generalize perturbation bounds for Networked OCO. Our bounds capture both the effect of temporal interactions as well as spatial interactions between node actions, which is a more challenging problem compared to previous literature that only considers either temporal interactions (Lin, Hu, Shi, et al., 2021) or spatial interactions (Shin, Anitescu, and Zavala, 2022) but not both simultaneously.

More specifically, recall that for each node v at period t, LPC solves an optimization problem $\psi_{p,(t,v)}^{(k,r)}$ where actions on the boundaries of available predictions (i.e., $\{t - 1\} \times N_v^r$ and $\partial N_{(t,v)}^{(k,r)}$) are fixed. By the principle of optimality, we know that if the actions on the boundaries are selected to be identical to the offline optimal actions, one can decide the optimal current action for a node by solving $\psi_{p,(t,v)}^{(k,r)}$. However, due to the limits on the prediction horizon and observation radius, LPC can only approximate the offline optimal actions on the boundaries (we do this by using the minimizer of node cost functions). The key idea to our analysis of the optimality gap of LPC is by first asking: *If we perturb the parameters of $\psi_{p,(t,v)}^{(k,r)}$, i.e., the fixed actions on the prediction boundaries $\partial N_{(t,v)}^{(k,r)}$, how large is the resulting change on the local action x_t^v in the optimal solution to* (3.14)*, which corresponds to the decision of LPC?*

Ideally, we would like the above impact to decay exponentially fast with respect to either the prediction horizon k or the observation radius r. We formalize this goal as *exponentially decaying local perturbation bound* in Definition 3.3.4. We then show in Theorems 3.3.3 and 3.3.4 that such bounds hold for the class of Networked OCO problems in Definition 3.3.3.

Definition 3.3.4. *We say an **exponentially decaying local perturbation bound** holds* *for the problem class $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$ if there exists non-negative constants*

$$C_1 = C_1(\ell_T/\mu, (\Delta\ell_S)/\mu) < \infty\,,\ \rho_T = \rho_T(\ell_T/\mu) < 1\,,\ \text{and}\ \rho_S = \rho_S((\Delta\ell_S)/\mu) < 1\,,$$

such that for any $p \in \mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$ and arbitrary, $\{(y_{t-1}^u)\}, \{(z_\tau^u)\}$, and $\{(z_\tau^u)'\}$, we have:

$$\begin{aligned}&\left\|\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \{z_\tau^u\}\right)_{(t,v)} - \psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \{(z_\tau^u)'\}\right)_{(t,v)}\right\| \\ \le{}& C_1 \sum_{(u,\tau)\in\partial N_{(t,v)}^{(k,r)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)}\left\|z_\tau^u - (z_\tau^u)'\right\|.\end{aligned}$$

Note that the exponentially decaying local perturbation bound in Definition 3.3.4 only consider the perturbations on $\{(z_\tau^u)\}$, which are caused by the limited prediction/observation power. We ignore the perturbations on $\{(y_{t-1}^u)\}$ because we quantify the error of LPC at each period by comparing its decision against the clairvoyant optimal trajectory that starts from the *same* previous actions $\{(x_{t-1}^u)\}$. The form of exponentially decaying local perturbation bound in Definition 3.3.4 is already sufficient to bound this per-period error, and the accumulation of past per-period errors can be handled separately by exponentially decaying perturbation bounds on the global scale, which previous works have established (Lin, Hu, Shi, et al., 2021; Lin, Hu, Qu, et al., 2022).

We illustrate two important consequences of the exponentially decaying local perturbation bound before proving it holds for Networked OCO in the following section. The first consequence is that when a unit of perturbation is applied to a node on the prediction boundary, the magnitude of impact on the optimal solution decays exponentially with respect to the temporal (or spatial) distance when the spatial (or temporal) distance is fixed (see Figure 3.7). The second is that when we apply a unit of perturbation at every node on the decision boundary, the impact on the decision decays quickly when k and r are increased simultaneously. Only increasing either one of k or r while the other is fixed cannot decrease the magnitude of the impact significantly. We illustrate this effect in Figure 3.8.

As we discuss in Section 3.2, perturbation bounds are important because they guarantee that (a) the online decision of predictive control is close to the clairvoyant optimal action and (b) the past "error" does not accumulate over time.

Exponentially Decaying Perturbation Bounds

The exponentially decaying local perturbation bound defined above is similar in spirit to two recent results, i.e., Lin, Hu, Shi, et al. (2021) derives a similar perturbation

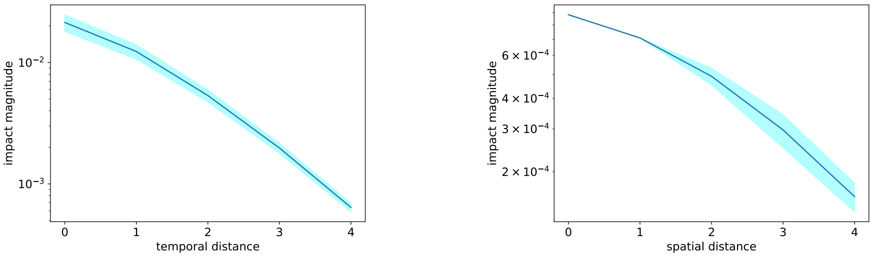

Figure 3.7: Simulation of exponentially decaying local perturbation bound. We study the optimal solution to (3.14) when $k = 6, r = 5$. In the left figure, we perturb the constraint at a random node in $\{(t, u) \mid u \in \partial N_v^r\}$ for $t = 0, 1, \ldots, 4$ and study the impact at node $(0, v)$. In the right figure, we perturb the constraint at a random node in $\{(k-1, u) \mid u \in \partial N_v^j\}$ for $j = 0, 1, \ldots, 4$ and study the impact at node $(0, v)$.

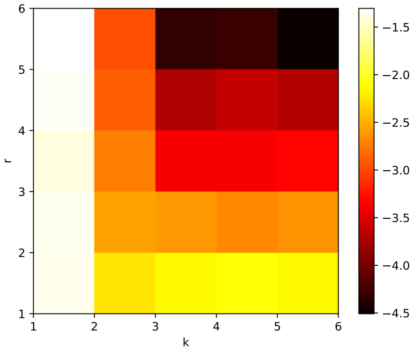

Figure 3.8: The aggregated impact of boundary perturbations when the prediction horizon and observation radius are (k, r). For each (k, r) pair in $\{1, \ldots, 5\}^2$, we generate a uniform random perturbation on every boundary constraint at the nodes in $\partial N_{(t,v)}^{(k,r)}$ and study the impact on the decision at node (t, v). We take log on the impact magnitude, and lighter color means the impact is stronger. From the figure, we see that one need to balance k and r to reduce the impact of prediction errors on the decision boundary.

bound for line graphs and Shin, Anitescu, and Zavala (2022) for general graphs with local perturbations. In fact, one may attempt to derive such a bound by applying these results directly; however, a major weakness of the direct approach is that it will yield $\rho_T = \rho_S$, i.e., it cannot distinguish between spatial and temporal dependencies, and the bound deteriorates as $\max\{\ell_T/\mu, \ell_S/\mu\}$ increases. For instance, even if the temporal interactions are weak (i.e., $\ell_T/\mu \approx 0$), $\rho_T = \rho_S$ can still be close to 1 if ℓ_S/μ is large, leading to a large slack in the perturbation bound for small prediction horizons k.

We overcome this limitation by redefining the action variables. Specifically, to focus on the temporal decay effect, we regroup all local actions in $\{\tau\} \times N_v^r$ as a "large" decision variable for period τ (in Figure 3.5 we would group each horizontal blue plane in N_v^r to create a new variable). After regrouping, we have $(k+1)$ "large" decision variables located on a line graph, where the strength of the interactions between consecutive variables is upper bounded by ℓ_T. On the other hand, to focus on spatial decay, we regroup all local actions in $\{\tau \mid t-1 \leq \tau < t+k\} \times \{v\}$ as a decision variable (in Figure 3.5 we would group each vertical orange line connecting from $t-1$ to $t+k-1$ to create a new variable). After regrouping, we have $|\mathcal{V}|$ "large" decision variables located on $\mathcal{G}$, where the strength of the interactions between two neighbors is upper bounded by ℓ_S. Averaging over the two perturbation bounds (since we have two valid bounds, their average is also a valid bound) provides the following exponentially decaying local perturbation bound (see Section 3.D for details of the proof).

Theorem 3.3.3. *For any tuple $(\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$, the exponentially decaying local perturbation bound (Definition 3.3.4) holds for the problem class $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$ with $C_1 = \frac{2\sqrt{\Delta \ell_S \ell_T}}{\mu}$, and*

$$\rho_T = \sqrt{1 - 2\left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1}}\ , \ \rho_S = \sqrt{1 - 2\left(\sqrt{1 + (\Delta \ell_S/\mu)} + 1\right)^{-1}}\ .$$

Note that, as ℓ_T/μ (respectively ℓ_S/μ) tends to zero, ρ_T (respectively ρ_S) in Theorem 3.3.3 also tends to zero with the scaling $\rho_T = \Theta(\sqrt{\ell_T/\mu})$ (resp. $\rho_S = \Theta(\sqrt{\ell_S/\mu})$). One may wonder if it is possible to derive a tighter bound on the decay factor. For example, if we can show $\rho_T = \Theta(\ell_T/\mu)$ and $\rho_S = \Theta(\ell_S/\mu)$, the lower bound on k and r to achieve a target competitive ratio can be decreased by half.

Next, we provide a tighter bound (through a refined analysis) for the regime where μ is much larger than ℓ_T, ℓ_S. Specifically, we establish a bound with the scaling

$\rho_T = \Theta(\ell_T/\mu)$ and $\rho_S = \Theta(\ell_S/\mu)$. Again, it is not possible to obtain this result from previous perturbation bounds in the literature.

Theorem 3.3.4. *Consider a tuple* $(\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$. *Given any* $b_1, b_2 > 0$, *define* $a = \sum_{\gamma \geq 0}(\frac{1+b_1}{1+b_1+b_2})^\gamma h(\gamma)$, $\tilde{a} = \sum_{\gamma \geq 0}(\frac{1}{1+b_1})^\gamma h(\gamma)$ *and* $\gamma_S = \frac{\sqrt{1+\Delta\ell_S/\mu}-1}{\sqrt{1+\Delta\ell_S/\mu}+1}$. *Suppose* $a, \tilde{a} < \infty$ *and* $\mu \geq \max\{8\tilde{a}\ell_T, \Delta\ell_S(b_1+b_2)/4\}$. *Then the exponentially decaying local perturbation bound (Definition 3.3.4) holds for* $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$ *with* $C_1 = \max\{\frac{a^2}{2\tilde{a}(1-4\tilde{a}\ell_T/\mu)}, \frac{2a^2\Delta\ell_S/\mu}{\gamma_S(1+b_1+b_2)(1-4\tilde{a}\ell_T/\mu)}\}$

$$\rho_T = \frac{4\tilde{a}\ell_T}{\mu}, \quad \rho_S = (1+b_1+b_2)\gamma_S.$$

Note that $\rho_T, \rho_S < 1$ follow from the condition on μ. Also observe that $\gamma_S = \Theta(\ell_S/\mu)$ as $\ell_S/\mu \to 0$.

The main difference between this result and Theorem 3.3.3 is, instead of dividing and redefining the action variables, we explicitly write down the perturbations along spatial edges and along temporal edges in the original temporal-spatial graph. We observe that per-period spatial interactions are characterized by a banded matrix and that the inverse of the banded matrix exhibits exponential correlation decay, which implies the exponentially decaying local perturbation bounds holds if the perturbed boundary action and the impacted local action we consider are at the same the time step. However, for a multi-period problem, to characterize the impact at a local action at some time step due to perturbation at a boundary action at a different time step is a difficult problem. To address this difficulty, the main technical contribution of our proof is to establish that a product of exponentially decaying matrices still satisfies exponential decay under the conditions in Theorem 3.3.4.

Our condition on $a, \tilde{a} < \infty$ and $\mu > \max\{8\tilde{a}\ell_T, \Delta\ell_S(b_1+b_2)/4\}$ characterizes a tradeoff between the allowable neighborhood boundary sizes $h(\gamma)$, and how large μ needs to be compared to the interaction cost parameters ℓ_T, ℓ_S. At one extreme, if $h(\gamma) = \Delta^\gamma$, then by setting $b_1 = 2\Delta - 1$ and $b_2 = 4\Delta^2 - 2\Delta$, we obtain $a = \tilde{a} = 2$ but must make a strong requirement on μ, namely, $\mu > \max\{16\ell_T, \Delta^3\ell_S(1-\frac{1}{4\Delta^2})\}$ (as we discussed in Corollary 3.3.2). At the other extreme, if $h(\gamma) \leq O(poly(\gamma))$ (as is the case if $\mathcal{G}$ is a grid), then $a, \tilde{a} < \infty$ holds for any $b_1, b_2 > 0$ and we can impose a weaker requirement on μ: for example, taking $b_1 = b_2 = 1$ yields a requirement $\mu > \max\{8\tilde{a}\ell_T, \Delta\ell_S/2\}$ (where $\tilde{a} = \sum_{\gamma \geq 0}(\frac{1}{2})^\gamma h(\gamma)$); which grows only linearly in Δ, and compares favorably with the $\mu > \Omega(\Delta^3)$ requirement which arose earlier.

From Perturbations to Competitive Bounds

We now present our main result, which bounds the competitive ratio of LPC using the exponentially decaying local perturbation bounds defined in the previous section.

Before presenting the result, we first provide some intuition as to why the perturbation bounds are useful for deriving the competitive ratio bound. Specifically, to bound the competitive ratio requires bounding the gap between LPC's trajectory and the offline optimal trajectory. This gap comes from the following two sources: (i) the per-period error made by LPC due to its limited prediction horizon and observation radius; and (ii) the cumulative impact of all per-period errors made in the past. Intuitively, the local perturbation bounds allow us to bound the per-period error made jointly by all nodes in LPC. Then, we use the perturbation bounds from Lin, Hu, Shi, et al. (2021) to help us bound the second type of cumulative errors: Although a per-period error is incurred at every period, the impact of past errors decays exponentially fast, so their accumulative effect does not grow with respect to time.

We present our main result in the following theorem. A formal proof can be found in Section 3.D.

Theorem 3.3.5. *For any tuple $v = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$, suppose the exponentially decaying local perturbation bound (Definition 3.3.4) holds with the decay factors ρ_T and ρ_S. Define*

$$\rho_G := 1 - 2 \cdot \left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1}, \text{ and } C_3(r) := \sum_{\gamma=0}^{r} h(\gamma) \cdot \rho_S^{\gamma}, \text{ for all } r \in \mathbb{N}.$$

If the prediction horizon r and observation radius k are large enough such that $h(r)^2 \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2k} \cdot \rho_G^{2k} \leq c_1 = c_1(\ell_f/\mu, \ell_T/\mu, (\Delta\ell_S)/\mu)$, then the competitive ratio of LPC for the problem class $\mathcal{P}(v)$ is bounded above as

$$CR_{\mathcal{P}(v)}(LPC) = 1 + O\left(h(r)^2 \cdot \rho_S^r\right) + O\left(C_3(r)^2 \cdot \rho_T^k\right).$$

Here the $\Omega(\cdot)$ and $O(\cdot)$ notations hide factors that depend polynomially on ℓ_f/μ, ℓ_T/μ, and $(\Delta\ell_S)/\mu$; see Appendix 3.D.

Recall that $h(r)$ denotes the size of the largest r-hop boundary in $\mathcal{G}$. The bound in Theorem 3.3.5 implies that if $h(r)$ can be upper bounded by $poly(r) \cdot \rho_S^{-\frac{(1-\iota)r}{2}}$ for some constant $\iota > 0$, the competitive ratio of LPC can be upper bounded by $1 + O(\rho_S^{\iota r}) + O(\rho_T^k)$, because $C_3(r)$ can be upper bounded by some constant that depends on ι in this case. Therefore, the competitive ratio improves exponentially with respect to the prediction horizon k and observation radius r (see Corollary 3.3.1 for an example). Note that the assumption $h(r) \leq poly(r) \cdot \rho_S^{-\frac{(1-\iota)r}{2}}$ is not particularly restrictive: For commonly seen graphs like an m-dimensional grid, $h(r)$ is polynomial in r, so $\iota = 1$ works.

More generally, note that ρ_S will converge to zero as ℓ_S tends to 0. Thus, for graphs with bounded degree $\Delta < \infty$, there exists $\delta = \delta(\Delta) > 0$ such that, when $\ell_S/\mu \leq \delta$, the spatial decay factor ρ_S (from either Theorem 3.3.3 or Theorem 3.3.4) will be small enough that, e.g., $h(r) \leq \Delta^r = O(\rho_S^{-\frac{r}{4}})$; i.e., $\iota = 1/2$ works. Therefore, we can eliminate the dependence on $h(r)$ and $C_3(r)$ in the competitive ratio by making additional assumptions on ℓ_S/μ, and a concrete example has been discussed in Corollary 3.3.2.

A Lower Bound

In this section, we show the competitive ratio in Theorem 3.3.5 is order-optimal by deriving a lower bound on the competitive ratio of any decentralized online algorithm with prediction horizon k and observation radius r. The specific constants and a proof of Theorem 3.3.6 can be found in Section 3.D.

Theorem 3.3.6. *Consider any tuple $v = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$ that satisfies $\Delta \geq 3$. The competitive ratio of any decentralized online algorithm ALG with prediction horizon k and observation radius r for the problem class $\mathcal{P}(v)$ is bounded below as*

$$CR_{\mathcal{P}(v)}(\mathsf{ALG}) = 1 + \Omega(\lambda_T^k) + \Omega(\lambda_S^r).$$

Here, the decay factor λ_T is given by $\lambda_T = \left(1 - 2\left(\sqrt{1 + (4\ell_T/\mu)} + 1\right)^{-1}\right)^2$. The decay factor λ_S is given by $\lambda_S = \frac{(\Delta\ell_S/\mu)}{3+3(\Delta\ell_S/\mu)}$ if $\Delta\ell_S/\mu < c_0$; $\lambda_S = \left(1 - 4\sqrt{3} \cdot (\Delta\ell_S/\mu)^{-\frac{1}{2}}\right)^2$ otherwise, where $c_0 \approx 267.3$ is a numerical constant. The $\Omega(\cdot)$ notation hides factors that depend polynomially on $1/\mu$, ℓ_T, and ℓ_S.

While Theorem 3.3.6 highlights that Theorem 3.3.5 is order-optimal, the decay factors λ_T, λ_S in the lower bound differ from their counterparts ρ_T, ρ_S in the upper bound for LPC. To understand the magnitude of the difference, we compare the bounds on graphs with bounded degree Δ. The decay factors are a function of the interaction strengths, which are measured by ℓ_S/μ and ℓ_T/μ. Our lower bound on the temporal decay factor λ_T and upper bound ρ_T only differ by a constant factor in the log-scale, and the same holds for the lower/upper bound in terms of the spatial decay factor.

To formalize this comparison, we derive a resource augmentation bound that bounds the additional "resources" that LPC needs to outperform the optimal decentralized online algorithm.[4] Here the prediction horizon k and the observation radius r can be viewed as the "resources" available to a decentralized online algorithm in our setting. We ask *how large do k and r given to LPC need to be, to ensure that it beats the optimal decentralized online algorithm given an observation radius r^* and prediction horizon k^*?*

We formally state our result in the following corollary and provide a proof in Section 3.D.

Corollary 3.3.7. *Consider any tuple $\nu = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$ that satisfies $\Delta \geq 3$ and $h(\gamma) = poly(\gamma) \cdot \rho_S^{-\gamma/4}$, where the $\tilde{O}$ notation hides a factor that depends polynomially on γ. Suppose the optimal decentralized online algorithm achieves a competitive ratio of $c(k^*, r^*)$ on the problem class $\mathcal{P}(\nu)$ with prediction horizon k^* and observation radius r^*. There exists a mapping $(k^*, r^*) \to (k, r)$ such that LPC with prediction horizon k and observation radius r is at least as good as that of the optimal decentralized online algorithm on class $\mathcal{P}(\nu)$ and* $\sup_{r^*} \limsup_{k^* \to \infty} k/k^* \leq 4$ *and* $\sup_{k^*} \limsup_{r^* \to \infty} r/r^* \leq 32$.

Note that we establish Corollary 3.3.7 based on the local perturbation bound in Theorem 3.3.3 rather than Theorem 3.3.4. This approach does not require assumptions on the relationship among μ, ℓ_T, and ℓ_S. In contrast, Theorem 3.3.4 can give better resource augmentation bounds under stronger assumptions on μ, ℓ_T, and ℓ_S, which we state formally below and provide a proof in Section 3.D.

Corollary 3.3.8. *Consider any tuple $\nu = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$ that satisfies $\Delta \geq 3$ and $\ell_T/\mu \leq \frac{1}{16}, \ell_S/\mu \leq \Delta^{-7}$. Suppose the optimal decentralized online algorithm achieves a competitive ratio of $c(k^*, r^*)$ on the problem class $\mathcal{P}(\nu)$ with prediction horizon k^* and observation radius r^*. There exists a mapping $(k^*, r^*) \to (k, r)$ and a positive constant $C = C(\Delta) < \infty$ such that LPC with prediction horizon k and observation radius r is at least as good as that of the optimal decentralized online algorithm on class $\mathcal{P}(\nu)$ and* $\sup_{r^*} \limsup_{k^* \to \infty} k/k^* \leq 1 + C/\log(\mu/\ell_T)$ *and* $\sup_{k^*} \limsup_{r^* \to \infty} r/r^* \leq 2 + C/\log(\mu/\ell_S)$.

[4] See, e.g., Roughgarden (2020), for an introduction to this flavor of bounds for expressing the near-optimality of an algorithm.

In the extreme that the ratios ℓ_T/μ and ℓ_S/μ tends to 0, to match the competitive ratios of the optimal decentralized algorithm, the resources required by LPC satisfies $k/k^* \to 1$ and $r/r^* \to 2$. Here, r/r^* does not converge to 1 because we do not make additional assumptions about the function h, which characterizes how fast the γ-hop neighborhood in $\mathcal{G}$ grows. Thus, a part of the spatial decay factor ρ_S is diverted to handle the exponentially growing h. To improve the convergence limit to $r/r^* \to 1$, we need to additionally assume the boundary size function h grows polynomially, i.e., $h(\gamma) = \text{poly}(\gamma)$.

Inexact Predictions

In this section, we discuss how to generalize the performance bounds for LPC (Algorithm 2) to the case when the predictions of future cost function are inexact, i.e., the predicted functions $\{f^u_{\tau \mid t}, c^u_{\tau \mid t}, s^e_{\tau \mid t}\}$ are different from the true functions $\{f^u_\tau, c^u_\tau, s^e_\tau\}$. Here, we use the subscript $\tau \mid t$ to denote the prediction for period τ ($\tau \geq t$) made at period t. Such cases can arise naturally in applications, for example, when we predict the future demand functions in multi-product pricing.

Under inexact predictions, the degradation of the performance depends on the magnitude of the prediction errors. To quantify such errors, we make a structural assumption by assuming that the uncertainty on each local cost function comes from an *uncertainty parameter* and the prediction error of this function is the result of noisy estimation of its uncertainty parameter. Specifically, we introduce the notations of generalized local cost functions

$$\tilde{f}^u_t : \mathbb{R}^d \times \Omega_u \to \mathbb{R}_{\geq 0},\ \tilde{c}^u_t : \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{A}_u \to \mathbb{R}_{\geq 0},\ \tilde{s}^e_t : \mathbb{R}^d \times \mathbb{R}^d \times \mathcal{B}_e \to \mathbb{R}_{\geq 0},$$

which can help to put both the true and the predicted cost functions under a unified framework. Here, $\{\Omega_u\}_{u \in \mathcal{V}}$, $\{\mathcal{A}_u\}_{u \in \mathcal{V}}$, and $\{\mathcal{B}_e\}_{e \in \mathcal{E}}$ are convex compact subsets of a finite-dimension Euclidean space $\mathbb{R}^m$. The relationship between the generalized local cost functions and the true/predicted local cost functions are summarized in the table below:

	True cost function at period τ	Prediction made at period $t (t \leq \tau < t + k)$
Node	$f^u_\tau(x^u_\tau) \coloneqq \tilde{f}^u_\tau(x^u_\tau; (\omega^v_\tau)^*)$	$f^u_{\tau \mid t}(x^u_{\tau \mid t}) \coloneqq \tilde{f}^u_\tau(x^u_{\tau \mid t}; \omega^v_{\tau \mid t})$
Temporal	$c^v_\tau(x^v_\tau, x^v_{\tau-1}) \coloneqq \tilde{c}^v_\tau(x^v_\tau, x^v_{\tau-1}; (\alpha^v_\tau)^*)$	$c^v_{\tau \mid t}(x^v_\tau, x^v_{\tau-1}) \coloneqq \tilde{c}^v_\tau(x^v_\tau, x^v_{\tau-1}; \alpha^v_{\tau \mid t})$
Spatial	$s^e_\tau(x^u_\tau, x^v_\tau) \coloneqq \tilde{s}^e_\tau(x^u_\tau, x^v_\tau; (\beta^e_\tau)^*)$	$s^e_{\tau \mid t}(x^u_\tau, x^v_\tau) \coloneqq \tilde{s}^e_\tau(x^u_\tau, x^v_\tau; \beta^e_{\tau \mid t})$.

Previous works (Chen, Agarwal, et al., 2015; Chen, Comden, et al., 2016; Lin, Hu, Qu, et al., 2022) also use similar ways to parameterize the prediction errors. With the notation of generalized local cost functions, we can measure the prediction errors, i.e., the error for predicting τ periods into the future, by the total distance between true/predicted uncertainty parameters (see, e.g., Γ_τ defined in Theorem 3.3.10). An instance of *Networked OCO with inexact predictions* problem is characterized by the generalized cost function, the ground truth uncertainty parameters, and the predicted uncertainty parameters. We provide the formal definition in Definition 3.3.5.

Definition 3.3.5. *An instance of the Networked OCO with inexact predictions is characterized by a tuple with 7 entries,*

$$(\mathcal{G}, H, x_0, \{\tilde{f}_t^v, \tilde{c}_t^v, \tilde{s}_t^e, D_t^v\}_{t\in[H],v\in\mathcal{V},e\in\mathcal{E}}, \{\Omega_v, \mathcal{A}_v, \mathcal{B}_e\}_{v\in\mathcal{V},e\in\mathcal{E}}, \{\xi_t^*\}_{t\in[H]}, \{\hat{\xi}_t\}_{t\in[H]}),$$

that contains the graph, the number of periods, the initial decisions, the set of all cost functions/constraint sets, the sets of uncertainty parameters, the ground true uncertainty parameters, and the predicted uncertainty parameters. Here, for every $t \in [H]$, we use the notations

$$\xi_t^* = \{(\omega_t^v)^*, (\alpha_t^v)^*, (\beta_t^e)^*\}_{v\in\mathcal{V},e\in\mathcal{E}}, \text{ and } \hat{\xi}_t = \{\omega_{\tau|t}^v, \alpha_{\tau|t}^v, \beta_{\tau|t}^e\}_{t\le\tau<t+k,v\in\mathcal{V},e\in\mathcal{E}}.$$

Based on the results we have derived for LPC with exact prediction, a critical step in our proof is to bound the difference between the decisions of LPC with inexact predictions and its counterpart with exact predictions in terms of the magnitude of the prediction errors. To achieve this, we show a generalized local exponentially decaying perturbation bound that also considers the perturbations on uncertainty parameters. The proof of the generalized perturbation bound requires an additional assumption that the gradients of the generalized local cost functions with respect to decision variables are ℓ_w-Lipschitz in uncertainty parameters.

Assumption 3.3.2. *For $\mu > 0, \ell_f < \infty, \ell_T < \infty, \ell_S < \infty, \ell_w < \infty$, the local cost functions and feasible sets for all $t \in [H], v \in \mathcal{V}, e \in \mathcal{E}$ satisfy:*

- *For every $v \in \mathcal{V}$ and $e \in \mathcal{E}$, $\Omega_v, \mathcal{A}_v$, and $\mathcal{B}_e$ are convex compact subsets of $\mathbb{R}^m$.*
- *$D_t^v \subseteq \mathbb{R}^n$ satisfies $int(D_t^v) \neq \emptyset$ and can be written as $D_t^v := \{x_t^v \in \mathbb{R}^n \mid (g_t^v)_i(x_t^v) \le 0, \forall 1 \le i \le m_t^v\}$, where each $(g_t^v)_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function in C^2.*
- *$\tilde{f}_t^v : \mathbb{R}^n \times \Omega_v \to \mathbb{R}_{\ge 0}$ is in C^2. It also satisfies that*

$$\mu I_n \preceq \nabla_{x_t^v}^2 \tilde{f}_t^v(x_t^v; \omega_t^v) \preceq \ell_f I_n, \forall x_t^v \in \mathbb{R}^n, \omega_t^v \in \Omega_v,$$
$$\left\|\nabla_{\omega_t^v} \nabla_{x_t^v} \tilde{f}_t^v(x_t^v; \omega_t^v)\right\| \le \ell_w, \forall x_t^v \in D_t^v, \omega_t^v \in \Omega_v.$$

- *$\tilde{c}_t^v : \mathbb{R}^n \times \mathbb{R}^n \times \mathcal{A}_v \to \mathbb{R}_{\ge 0}$ is in C^2. It also satisfies that*

$$0 \preceq \nabla_{(x_t^v, x_{t-1}^v)}^2 \tilde{c}_t^v(x_t^v, x_{t-1}^v; \alpha_t^v) \preceq \ell_T I_{2n}, \forall x_t^v, x_{t-1}^v \in \mathbb{R}^n, \alpha_t^v \in \mathcal{A}_v,$$
$$\left\|\nabla_{\alpha_t^v} \nabla_{(x_t^v, x_{t-1}^v)} \tilde{c}_t^v(x_t^v, x_{t-1}^v; \alpha_t^v)\right\| \le \ell_w, \forall x_t^v, x_{t-1}^v \in D_t^v, \alpha_t^v \in \mathcal{A}_v.$$

- $\tilde{s}_t^e : \mathbb{R}^n \times \mathbb{R}^n \times \mathcal{B}_e \to \mathbb{R}_{\geq 0}$ *is in* C^2. *Let* $e = (u, v)$. *It also satisfies that*

$$0 \preceq \nabla^2_{(x_t^u, x_t^v)} s_t^e(x_t^u, x_t^v; \beta_t^e) \preceq \ell_S I_{2n}, \forall x_t^u, x_t^v \in \mathbb{R}^n, \beta_t^e \in \mathcal{B}_e,$$
$$\left\| \nabla_{\beta_t^e} \nabla_{(x_t^u, x_t^v)} s_t^e(x_t^u, x_t^v; \beta_t^e) \right\| \leq \ell_w, \forall x_t^u, x_t^v \in D_t^e, \beta_t^e \in \mathcal{B}_e.$$

Intuitively, the Lipschitzness assumption we make on the generalized local cost functions guarantees that the impact of inexact predictions of the uncertainty parameters on the optimal solution can be bounded, and the Networked OCO with inexact problem gets more challenging as ℓ_w increases. We define the set of possible configurations for Networked OCO with inexact predictions as

$$\tilde{\Upsilon} := \{(\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h) \mid \mu \in \mathbb{R}_{>0}, \ell_f \in \mathbb{R}_{>0}, \ell_T \in \mathbb{R}_{\geq 0}, \ell_S \in \mathbb{R}_{\geq 0}, \ell_w \in \mathbb{R}_{\geq 0},$$
$$\Delta \in \mathbb{N}, h : \mathbb{N} \to \mathbb{N}\}.$$

Each configuration tuple in $\tilde{\Upsilon}$ specifies a problem class of Networked OCO with inexact predictions. We define this relationship in Definition 3.3.6.

Definition 3.3.6. *For any configuration tuple* $v = (\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h) \in \tilde{\Upsilon}$, *we define* $\tilde{\mathcal{P}}(v)$ *as the set of problem instances of Networked OCO with inexact predictions*

$$(\mathcal{G}, H, x_0, \{\tilde{f}_t^v, \tilde{c}_t^v, \tilde{s}_t^e, D_t^v\}_{t \in [H], v \in \mathcal{V}, e \in \mathcal{E}}, \{\Omega_v, \mathcal{A}_v, \mathcal{B}_e\}_{v \in \mathcal{V}, e \in \mathcal{E}}, \{\xi_t^*\}_{t \in [H]}, \{\hat{\xi}_t\}_{t \in [H]}),$$

that satisfy:

1. *Assumption 3.3.2 with* $(\mu, \ell_f, \ell_T, \ell_S, \ell_w)$;
2. *degree*$(v) \leq \Delta$ *for every node* $v \in \mathcal{V}$;
3. $|\partial N_v^r| \leq h(r)$ *for every node* $v \in \mathcal{V}$ *and* $r \in \mathbb{N}$.

Before picking a local action x_t^v at time t, agent v can observe k periods of future node costs, temporal interaction costs, and spatial interaction costs within its r-hop neighborhood but with noisy prediction of the uncertainty parameters, $\{\{(f_\tau^u, \omega_{\tau|t}^u), (c_\tau^u, \alpha_{\tau|t}^u) \mid u \in N_v^r\}, \{(s_\tau^e, \beta_{\tau|t}^e) \mid e \in \mathcal{E}(N_v^r)\}\}_{t \leq \tau < t+k}$, and the previous local actions in N_v^r: $\{x_{t-1}^u \mid u \in N_v^r\}$. To simplify the notations, we define

$$\hat{\xi}_{p,(t,v)}^{(k,r)} := \{\{\omega_{\tau|t}^u, \alpha_{\tau|t}^u \mid u \in N_v^r\}, \{\beta_{\tau|t}^e \mid e \in \mathcal{E}(N_v^r)\}\}_{t \leq \tau < t+k},$$

which are the predicted parameters at period t;

$$\left(\xi_{p,(t,v)}^{(k,r)}\right)^* := \{\{(\omega_\tau^u)^*, (\alpha_\tau^u)^* \mid u \in N_v^r\}, \{(\beta_\tau^e)^* \mid e \in \mathcal{E}(N_v^r)\}\}_{t \leq \tau < t+k},$$

which are the ground true parameters. We define

$$\tilde{\psi}_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u \mid u \in N_v^r\}, \{z_\tau^u \mid (\tau,u) \in \partial N_{(t,v)}^{(k,r)}\}; \xi_{(t,v)}^{(k,r)}\right)$$

as the optimal solution of the problem

$$\begin{aligned}
\min \sum_{\tau=t}^{t+k-1} \Bigg(& \sum_{u \in N_v^{r-1}} f_\tau^u(x_\tau^u; \omega_\tau^u) + \sum_{u \in N_v^{r-1}} c_\tau^u(x_\tau^u, x_{\tau-1}^u; \alpha_\tau^u) \\
& + \sum_{(u,q) \in \mathcal{E}(N_v^{r-1})} s_\tau^{(u,q)}(x_\tau^u, x_\tau^q; \beta_\tau^{(u,q)}) \Bigg) \\
\text{s.t. } & x_{t-1}^u = y_{t-1}^u, \forall u \in N_v^r, \\
& x_\tau^u = z_\tau^u, \forall (\tau, u) \in \partial N_{(t,v)}^{(k,r)}, \\
& x_\tau^u \in D_\tau^u, \forall (\tau, u) \in N_{(t,v)}^{(k-1,r-1)}.
\end{aligned} \tag{3.16}$$

When applied to Networked OCO with inexact predictions, LPC first construct the local cost functions within the prediction horizon and observation radius by treating the predicted uncertainty parameters as the true ones. Then, it follows the same procedure as LPC with exact predictions to decide the current local action. The pseudocode of LPC for Networked OCO with inexact predictions is given in Algorithm 3.

```
**Algorithm 3:** Localized Predictive Control with Inexact Predictions (for node $v$)
**Parameters:** Prediction horizon $k$ and observation radius $r$. **for** $t = 1$ ***to*** $H$ **do**
    Receive information $\{x_{t-1}^u \mid u \in N_v^r\}$ and observe
        $$\{\{(f_\tau^u, \omega_{\tau|t}^u), (c_\tau^u, \alpha_{\tau|t}^u) \mid u \in N_v^r\}, \{(s_\tau^e, \beta_{\tau|t}^e) \mid e \in \mathcal{E}(N_v^r)\}\}_{t \le \tau < t+k}.$$
    Choose local action $x_t^v$ to be the $(t, v)$-th element in
        $$\psi_{p,(t,v)}^{(k,r)}\Big(\{x_{t-1}^u \,|\, u \in N_v^r\}, \big\{\theta_\tau^u \,|\, (\tau, u) \in \partial N_{(t,v)}^{(k,r)}\big\}; \hat{\xi}_{p,(t,v)}^{(k,r)}\Big)$$
     the solution of (3.14), where $\theta_{\tau|t}^u \coloneqq \arg\min_{y \in D_\tau^u} f_\tau^u(y; \omega_{\tau|t}^u)$.
**end**
```

We generalize the exponentially decaying local perturbation bound in Definition 3.3.4 to also include perturbations on the predicted uncertainty parameters. This generalization is necessary to study how the magnitude of the prediction errors affect the performance of LPC. We provide the formal definition the generalized in Definition 3.3.7.

Definition 3.3.7. *We say the* ***generalized exponentially decaying local perturbation bound*** *holds for the problem class $\tilde{P}(\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h)$ if for non-negative constants*

$$C_1 = C_1(\ell_T/\mu, (\Delta\ell_S)/\mu) < \infty, C_2 = C_2(\ell_w/\mu) < \infty,$$
$$\rho_T = \rho_T(\ell_T/\mu) < 1, \rho_S = \rho_S((\Delta\ell_S)/\mu) < 1,$$

such that for any $p \in \tilde{P}(\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h)$ and arbitrary

$$\{y_{t-1}^u\}, \{(z_\tau^u)'\}, \{(z_\tau^u)\}, \xi_{(t,v)}^{(k,r)}, (\xi_{(t,v)}^{(k,r)})',$$

we have:

$$\left\| \psi_{p,(t,v)}^{(k,r)} \left(\{y_{t-1}^u\}, \{z_\tau^u\}; \xi_{(t,v)}^{(k,r)} \right)_{(t,v)} - \psi_{p,(t,v)}^{(k,r)} \left(\{y_{t-1}^u\}, \{(z_\tau^u)'\}; (\xi_{(t,v)}^{(k,r)})' \right)_{(t,v)} \right\|$$
$$\leq C_1 \sum_{(u,\tau)\in\partial N_{(t,v)}^{(k,r)}} \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,u)} \left\| z_\tau^u - (z_\tau^u)' \right\| + C_2 \cdot \mathsf{dist}_p \left(\xi_{(t,v)}^{(k,r)}, (\xi_{(t,v)}^{(k,r)})' \right),$$

where $\mathsf{dist}_p \left(\xi_{(t,v)}^{(k,r)}, (\xi_{(t,v)}^{(k,r)})' \right)$ is defined as

$$\sum_{(u,\tau)\in N_{(t,v)}^{(k-1,r-1)}} \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,u)} \left\| \omega_\tau^u - (\omega_\tau^u)' \right\| + \sum_{(u,\tau)\in N_{(t,v)}^{(k,r)}} \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,u)} \left\| \alpha_\tau^u - (\alpha_\tau^u)' \right\|$$
$$+ \sum_{\tau=t}^{t+k} \sum_{e\in\mathcal{E}(N_v^r)} \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,e)} \left\| \beta_\tau^e - (\beta_\tau^e)' \right\|.$$

Using a similar approach with our proof of Theorem 3.3.3, we state the following generalized exponentially decaying local perturbation bound for Networked OCO with inexact predictions and defer its proof to Section 3.D.

Theorem 3.3.9. *For any tuple $v = (\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h) \in \tilde{\Upsilon}$, the generalized exponentially decaying local perturbation bound (Definition 3.3.7) holds for the problem class $\tilde{P}(v)$ with $C_1 = \frac{2\sqrt{\Delta\ell_S\ell_T}}{\mu}, C_2 = \frac{2\ell_w}{\mu}$, and*

$$\rho_T = \sqrt{1 - 2\left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1}}, \rho_S = \sqrt{1 - 2\left(\sqrt{1 + (\Delta\ell_S/\mu)} + 1\right)^{-1}}.$$

We present our main result for LPC with inexact predictions in Theorem 3.3.10.

Theorem 3.3.10. *For any tuple $v = (\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h) \in \tilde{\Upsilon}$, suppose the generalized exponentially decaying local perturbation bound (Definition 3.3.7) holds with the decay factors ρ_T and ρ_S. Define*

$$\rho_G := 1 - 2 \cdot \left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1}, \text{ and } C_3(r) := \sum_{\gamma=0}^{r} h(\gamma) \cdot \rho_S^\gamma, \text{ for all } r \in \mathbb{N}.$$

If the prediction horizon r and observation radius k are large enough such that $h(r)^2 \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2k} \cdot \rho_G^{2k} \le c_2 = c_2(\ell_f/\mu, \ell_T/\mu, (\Delta\ell_S)/\mu)$, then, for any problem instance p in the class $\tilde{\mathcal{P}}(\nu)$, the total cost of LPC satisfies

$$\begin{aligned}&\text{cost}_p(LPC) - \text{cost}_p(\mathsf{OPT})\\ \le\ & O\left(h(r)^2 \cdot \rho_S^r + C_3(r)^2 \cdot \rho_T^k\right) \cdot \text{cost}_p(\mathsf{OPT}) + O\left(C_3(r)^2 \sum_{\tau=0}^{k-1} \rho_T^\tau \Gamma_\tau(p)\right),\end{aligned}$$

where $\Gamma_\tau(p)$ denotes the total error for predicting τ time steps into the future on p, i.e.,

$$\begin{aligned}\Gamma_\tau(p) = \sum_{t=1}^{H-\tau}\Bigg(&\sum_{v\in\mathcal{V}} \left\|\omega_{t+\tau|t}^v - (\omega_{t+\tau}^v)^*\right\|^2 + \sum_{v\in\mathcal{V}} \left\|\alpha_{t+\tau|t}^v - (\alpha_{t+\tau}^v)^*\right\|^2\\ &+ \sum_{e\in\mathcal{E}} \left\|\beta_{t+\tau|t}^e - (\beta_{t+\tau}^e)^*\right\|^2\Bigg).\end{aligned}$$

Here the $O(\cdot)$ notation hides factors that depend polynomially on ℓ_f/μ, ℓ_T/μ, $(\Delta\ell_S)/\mu$, and ℓ_w/μ.

To interpret Theorem 3.3.10, note that the first (multiplicative) term that involves cost(OPT) is the same up to a constant factor as the competitive ratio bound derived in Theorem 3.3.5 for LPC with exact predictions. The second (additive) term is a weighted sum of the total squared prediction errors $\Gamma_\tau(p)$ in the entire horizon. The exponentially decaying coefficients (ρ_T^τ) indicate that LPC can tolerate more error in predictions of the more distant future. This also suggests that given limited resources to improve predictions, improving predictions of the near future may be more valuable than improving predictions of the distant future. Specifically, for an instance p of Networked OCO with inexact predictions, recall that $\Gamma_\tau(p)$ denotes the total error of predicting τ periods into the future on p. Depending on how $\Gamma_\tau(p)$ grows with τ, there may be an optimal prediction horizon k^* under which the cost of LPC is minimized. The problem of finding the optimal prediction horizon for an MPC-based algorithm has been studied in the single-node setting (Lin, Preiss, Anand, et al., 2023; Li, Preiss, et al., 2023), and generalizing existing results to Networked OCO with inexact predictions is an interesting topic of future research.

3.4 Application: Adaptive Video Streaming

Adaptive bitrate streaming (ABR) addresses the challenge of delivering consistent high-quality video under volatile network conditions with the goal to balance three critical factors: maximizing visual fidelity, minimizing rebuffering events, and avoiding abrupt bitrate switches. Although existing works make attempts to apply model predictive control (MPC) for ABR, a straightforward application often suffers in environments with fluctuating or uncertain throughput predictions. By drawing on recent advances in perturbation analysis, one can show that MPC becomes robust to these prediction errors if the underlying optimal control formulation satisfies exponentially decaying perturbation properties. To achieve such properties in the setting of video streaming, we introduce a time-based ABR formulation that explicitly encodes a buffer cost. The buffer cost follows two intuitions: Empirically, it penalizes the agent for letting the buffer level get close to the constraint boundaries even if it does not violate the constraint; Theoretically, it brings in a strongly convex stage cost on the state that is critical for establishing the exponentially decaying perturbation properties.

A Time-based ABR Formulation

Our time-based ABR formulation treats a video stream as a *continuous flow* rather than a discrete sequence of segments. Consider a streaming session that consists of N time intervals with fixed duration Δt in terms of *clock time* (not video time). The controller's task is to select a bitrate for each time interval from a set of available bitrates $\mathcal{R} \subset [r_{\min}, r_{\max}]$ to optimize for a combination of higher video quality, shorter rebuffering time, and less frequent bitrate switching.

Let ω_n denote the average throughput during the n^{th} time interval, r_n the selected bitrate for that time interval, and x_n the buffer level immediately after that time interval. Our objective is to minimize the overall cost given as a linear combination of the three QoE components:

$$\sum_{n=1}^{N} \left(v(r_n) \cdot \frac{\omega_n \Delta t}{r_n} + \beta \cdot b(x_n) + \gamma \cdot c(r_n, r_{n-1}) \right), \tag{3.17}$$

where

- $v(r_n)$ is the **distortion cost**, which should be a positive, strictly decreasing, and convex function that models the encoding distortion, e.g., $v(r_n) = 1/r_n$. It is then weighted by the amount of video downloaded during that time interval, i.e., $\omega_n \Delta t / r_n$ because the controller downloads a *variable* amount of video during each fixed time interval.

- $b(x_n)$ is the **buffer cost**, which aims to stabilize the buffer level around a target level $\bar{x}$, i.e.,

  $$b(x_n) = \begin{cases} (\bar{x} - x_n)^2 & x_n \leq \bar{x} \\ \epsilon(x_n - \bar{x})^2 & x_n > \bar{x} \end{cases},$$

  where $\epsilon < 1$ is a small constant. Note that we purposely do not model the rebuffering time explicitly to avoid the pitfalls encountered by `RobustMPC` and as we show later, this helps `SODA` achieve theoretical performance guarantees.

- $c(r_n, r_{n-1})$ is the **switching cost** from the previous bitrate to the current bitrate, e.g., $c(r_n, r_{n-1}) = (v(r_n) - v(r_{n-1}))^2$.

Coefficients β and γ are positive weights for the buffer and the switching cost, respectively, based on user preferences. The choices for the distortion and switching cost functions are flexible.

The *time-based* buffer dynamics are introduced into the optimization problem through the following constraint:

$$x_n = x_{n-1} + \frac{\omega_n \Delta t}{r_n} - \Delta t \in [0, x_{\max}],$$

where $\omega_n \Delta t / r_n$ accounts for the variable amount of video downloaded during a time interval and Δt accounts for the fixed amount of buffer drained during the same time interval. Note that we do not allow the controller to violate the buffer range constraint during the *optimization phase* when determining the bitrate. Of course, due to throughput prediction errors, this may sometimes be inevitable during the *execution phase* when applying the bitrate decision.

Why a Time-Based Formulation? The time-based formulation allows a cleaner theoretical analysis over a given throughput sequence $(\omega_1, \ldots, \omega_N)$. For example, consider the throughput function shown in Figure 3.9. In the time-based formulation, we naturally have $\omega_1 = 4$, $\omega_2 = 1$, and $\omega_3 = \omega_4 = \{2\} Mb/s$ given $\Delta t = \{1\} s$. By contrast, in the segment-based formulation, the throughput sequence becomes dependent on the bitrate sequence. Assuming the segment duration is also $L = \{1\} s$, if the controller chooses $r_1 = \{2.0\} Mb/s$ and $r_2 = \{2.5\} Mb/s$, then it takes 0.5 and 1 s to download the first and second segments, respectively, resulting in $\omega_1 = 4$ and $\omega_2 = \{2.5\} Mb/s$. As such, the segment based formulation gets causally biased due to bitrate selection $r_1, ..., r_N$, which in turn makes it difficult to theoretically analyze the design (Bothra et al., 2023).

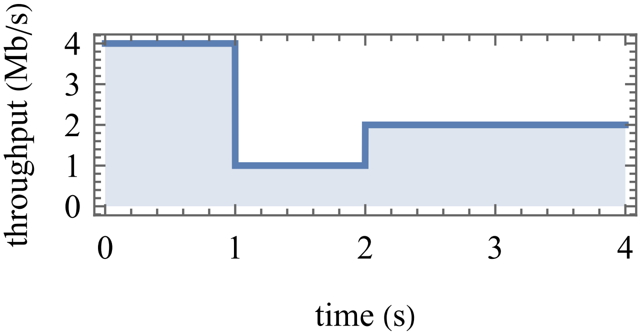

Figure 3.9: Sample throughput function illustrates why time-based formulation is better for future prediction.

Incorporating Throughput Predictions

In addition to facilitating theoretical analysis, our time-based formulation is crucial to ensuring the validity of throughput predictions over the prediction horizon. An important observation is that *bitrate decisions have no causal impact on how long the throughput predictions are valid for*. However, segment-based controllers such as MPC (Yin et al., 2015) and Fugu (Yan et al., 2020) intertwine throughput predictions and bitrate decisions in non-causal ways. In these designs, for a given available bandwidth, the throughput prediction horizon spans shorter periods of clock time when low bitrate is selected compared to when high bitrate is selected. In fact, their underlying assumption about the validity of the throughput prediction horizon can vary by $r_{\max}/r_{\min}$.

By contrast, the way we incorporate throughput predictions into SODA *does not* suffer from this issue. Specifically, just before each time interval, the controller is given access to a (not necessarily accurate) throughput prediction for the next K time intervals from a black-box throughput predictor. It is always assumed that the validity of the throughput prediction is $K\Delta t$, a fixed value. In general, a throughput predictor may output a different value for each of the next K time intervals, i.e., $\hat{\omega}_{n|n-1}, \hat{\omega}_{n+1|n-1}, \ldots, \hat{\omega}_{n+K-1|n-1}$, where $\hat{\omega}_{m|n-1}$ $(m \geq n)$ is the throughput prediction for the m^{th} time interval given previous download information up until the $(n-1)^{\text{th}}$ time interval. In other words, a throughput predictor can output a piecewise constant throughput function for the next $K\Delta t$ time. In practice, though, a typical throughput predictor outputs a single value that corresponds to a constant throughput function.

Control Mechanism

Inspired by the model predictive control framework, SODA selects a bitrate for each time interval by optimizing over the next K time intervals and then committing to the bitrate decision for the immediate next time interval, i.e., minimizing the following objective function

$$\sum_{m=n}^{n+K-1}\left(v(r_m)\cdot\frac{\hat{\omega}_{m|n-1}\Delta t}{r_m}+\beta\cdot b(x_m)+\gamma\cdot c(r_m,r_{m-1})\right)$$

$$\begin{aligned}\text{subject to}\quad & x_m = x_{m-1}+\frac{\hat{\omega}_{m|n-1}\Delta t}{r_m}-\Delta t,\\ & x_m\in[0,x_{\max}],\quad r_m\in\mathcal{R},\end{aligned}$$

with respect to variables $r_n,\dots,r_{n+K-1}$ and then committing to only the first bitrate decision r_n.

Solving this optimization problem is computationally expensive, furthermore, it is unclear what prediction horizon should be used and how accurate throughput predictions must be in order for SODA to perform well.

Theoretical Design Insights

Our design of SODA is motivated by recent theoretical advances at the interface of learning and control (Lin, Hu, Shi, et al., 2021; Hazan and Singh, 2022; Agarwal et al., 2019; Dean et al., 2020) and smoothed online convex optimization (Chen, Goel, and Wierman, 2018; Goel, Lin, et al., 2019; Lin, Gan, et al., 2022). In particular, we design SODA to satisfy an *exponentially decaying perturbation property* that has been shown to ensure efficient and robust use of predictions in model predictive control policies (Lin, Hu, Shi, et al., 2021; Lin, Hu, Qu, et al., 2022). Intuitively, this property describes the behavior of the solution to the optimization problem defining SODA as a function of problem parameters, including bandwidth predictions $\{\hat{\omega}_{m|n-1}\}_{n\leq m<n+K}$ and the previous buffer level/bitrate. When this property holds, the impact of perturbing a prediction $\hat{\omega}_{m|n-1}$ on the current bitrate decision r_n decays exponentially with respect to their temporal distance $(m-n)$. The formal definition of exponentially decaying perturbation generalizes the intuition above to consider the optimal trajectory and other parameters.

Definition 3.4.1 (Exponentially Decaying Perturbation Bound for ABR)**.** *We say the exponentially decaying perturbation bound holds if there exists uniform constants* $C>0,\rho\in(0,1)$ *such that the following inequalities hold:*

$$\left|\psi_t^{t+p}\left(\xi_{[t,t+p]};\mathbf{0}\right)_{x_\tau}-\psi_t^{t+p}\left(\xi'_{[t,t+p]};\mathbf{0}\right)_{x_\tau}\right|$$

$$\leq C\rho^{\tau-t+1}\left(\left|\sigma_{t-1}-\sigma'_{t-1}\right|+\left|\nu_{t-1}-\nu'_{t-1}\right|\right)+C\sum_{j=t}^{t+p}\rho^{|\tau-j|}\left|\hat{\omega}_j-\hat{\omega}'_j\right|, \tag{3.18}$$

$$\begin{aligned}&\left|\psi_t^{t+p}\left(\xi_{[t,t+p]};\mathbb{I}\right)_{x_\tau}-\psi_t^{t+p}\left(\xi'_{[t,t+p]};\mathbb{I}\right)_{x_\tau}\right|\\ &\leq C\rho^{\tau-t+1}\left(\left|\sigma_{t-1}-\sigma'_{t-1}\right|+\left|\nu_{t-1}-\nu'_{t-1}\right|\right)+C\sum_{j=t}^{t+p}\rho^{|\tau-j|}\left|\hat{\omega}_j-\hat{\omega}'_j\right|\\ &\quad+C\rho^{t+p-\tau}\left(\left|\sigma_{t+p}-\sigma'_{t+p}\right|+\left|\nu_{t+p+1}-\nu'_{t+p+1}\right|\right),\end{aligned} \tag{3.19}$$

where

$$\begin{aligned}\xi_{[t,t+p]}&\coloneqq\left((\sigma_{t-1},\nu_{t-1});\hat{\omega}_{t:t+p};(\sigma_{t+p},\nu_{t+p+1})\right),\\ \xi'_{[t,t+p]}&\coloneqq\left((\sigma'_{t-1},\nu'_{t-1});\hat{\omega}'_{t:t+p};(\sigma'_{t+p},\nu'_{t+p+1})\right).\end{aligned}$$

Two metrics that we use to measure SODA's performance theoretically are *dynamic regret* and *competitive ratio*, which are standard in the literature of online optimization Lin, Hu, Shi, et al., 2021; Hazan and Singh, 2022; Agarwal et al., 2019; Chen, Goel, and Wierman, 2018; Goel, Lin, et al., 2019. Specifically, let cost(ALG) denote the total cost incurred by an online algorithm ALG and cost(OPT) denote the offline optimal cost (3.17) an agent can incur if it has exact knowledge of all future bandwidth at the beginning. We say ALG achieves a dynamic regret of R if cost(ALG) $-$ cost(OPT) $\leq R$ always holds, and ALG achieves a competitive ratio of C if cost(ALG) $\leq C\cdot$ cost(OPT) always holds.

The key idea of our theoretical analysis is leveraging the exponentially decaying property to bound the per-step error of SODA against the hindsight optimal decision and the aggregations of such errors over time. To prove the exponentially decaying perturbation, we require a technical assumption that guarantees the controller can "reach" any desired buffer level by choosing the largest/smallest bitrate (see Assumption 3.4.1).

Assumption 3.4.1. *There exists uniform constants $\omega_{max}>\omega_{min}>0$ such that for any time step t, we have that $\omega_{min}\leq\omega_t\leq\omega_{max}$ holds. We also assume that $\omega_{min}/r_{min}\geq x_{max}$, and $\omega_{max}/r_{max}-1\leq-\delta$ holds for a fixed constant $\delta>0$.*

Intuitively, Assumption 3.4.1 guarantees that the controller can always fill up the buffer at the cost of choosing the smallest bitrate or decrease the buffer level by choosing the largest bitrate. This assumption is used to eliminate extreme boundary cases in the analysis, but SODA empirically performs well even when Assumption 3.4.1 is not strictly satisfied. Using this assumption, we show the exponentially decaying perturbation property holds for the video streaming problem in Theorem 3.4.1.

Theorem 3.4.1. *Under Assumption 3.4.1, the exponentially decaying perturbation bound holds with constants*

$$\rho = \left(1 - \frac{2}{1 + \sqrt{1 + \frac{\max\{6\omega_{min}(\omega_{min}+3), 4x_{max}(\omega_{min}+8\gamma)\}}{\omega_{min}^3 \epsilon \beta}}} \right)^{\frac{1}{3(3+\lceil x_{max}/\delta \rceil)}}$$

and

$$C = \frac{(1+\omega_{max}) \left(3\beta\omega_{min}^3 + \max\{6\omega_{min}(\omega_{min}+3), 4x_{max}(\omega_{min}+8\gamma)\}\right)}{\omega_{min}^3 \rho^{3+\lceil x_{max}/\delta \rceil}}.$$

Exact Predictions

When the bandwidth predictions are accurate, a small prediction horizon is sufficient for SODA to achieve near-optimal performance. In practice, it is desirable to use a relatively small prediction horizon for a predictive controller like SODA because prediction errors grow dramatically as we predict further into the future. Fortunately, the exponential decay property that ensures good performance with only a few predictions. More formally, we present a theorem showing that a small prediction horizon is sufficient for SODA to achieve near-optimal performance when the predictions within this window are accurate (i.e., $\hat{\omega}_{m|n-1} = \omega_m$ for $m = n, \dots, n+K-1$).

Theorem 3.4.2. *[Informal] When the predictions of the bandwidth in future K steps are exact (i.e., $\hat{\omega}_{m|n-1} = \omega_m$ for $m = n, \dots, n+K-1$) and the prediction horizon $K \geq O(1)$, SODA achieves a dynamic regret of $O(\rho^K N)$ and a competitive ratio of $1 + O(\rho^K)$, where $\rho < 1$ is the decay factor of the exponentially decaying perturbation property.*

The formal statement of Theorem 3.4.2 is given in Section 3.E. This result implies that SODA's performance approaches that of the optimal sequence of decisions *exponentially fast* in the prediction horizon size K; thus, only a small prediction horizon length is necessary to obtain good performance.

Inexact Predictions

We now relax the exact prediction assumption to prove SODA's robustness to a certain level of prediction errors thanks to its exponentially decaying perturbation property.

Theorem 3.4.3. *[Informal] Suppose the prediction error at each step is bounded above. The buffer level of* SODA *will never hit the constraint boundary, i.e.,* $0 < x_n < x_{max}$*. Further, define* $\mathcal{E} = \rho^{2K} N + \sum_{\kappa=1}^{K} \rho^{\kappa} E_{\kappa}$*, where* E_{κ} *is the total squared error for predicting* κ *steps into the future.* SODA *achieves a dynamic regret of* $O(\sqrt{\mathcal{E} N} + \mathcal{E})$*.*

The formal statement of Theorem 3.4.3 is given in Section 3.E. Theorem 3.4.3 shows that, if the buffer costs are "steep" and the prediction errors on the bandwidth are relatively small, SODA can achieve a sequence of buffer levels that stay safely away from the boundaries of buffer constraint $[0, x_{\max}]$. The dynamic regret of SODA depends on the magnitude of the prediction errors and the regret improves when the errors become smaller. SODA acquires this guarantee thanks to its maintenance of the buffer near a target level $\bar{x}$. In contrast, RobustMPC (Yin et al., 2015) does not offer the same performance guarantee, thus even small bandwidth prediction errors can cause the video to rebuffer if the buffer level is near zero.

Computational Efficiency

Solving the predictive optimization problem to determine the exact optimal solution can be unrealistic in the application of adaptive bitrate streaming, where each decision needs to be made in the minimum possible time. A critical observation underlying the implementation of SODA is that it is sufficient to search only for bitrate sequences that are increasing or decreasing monotonically. We provide a theoretical justification in the following theorem.

Theorem 3.4.4. *[Informal] Suppose* SODA *is given the predictions that satisfy* $\hat{\omega}_{n|n-1} = \cdots = \hat{\omega}_{n+K-1|n-1}$ *at an intermediate time step n. Then, the bitrate trajectory solved by* SODA *can be approximated by a feasible monotonic bitrate trajectory with an error of* $O\left(\frac{K}{\sqrt{\gamma}}\right)$*.*

The formal statement of Theorem 3.4.4 is given in Section 3.E. Theorem 3.4.4 shows that the true optimal solution becomes closer to monotonic as the weight γ of switching costs increases. While the theoretical bound can be conservative, we find that even with moderate γ, the (discrete) decision made under the monotonic heuristic is usually identical to the true optimal solution on a real trajectory (see Figure 3.10).

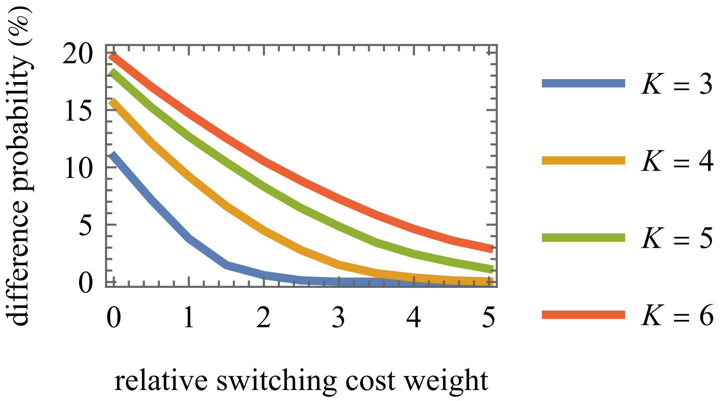

Figure 3.10: The probability that the bitrate decision produced by the approximate solver is different from that produced by the brute-force solver quickly converges to 0 as switching cost weight increases.

3.A Proofs for the Perturbation Analysis

Proof of Theorem 3.1.2

In the next lemma we will use the notation A_{S_R,S_C} to denote the submatrix obtained by selecting the blocks indexed by some set $S_R \times S_C$ while preserving their relative order. Specifically, consider a matrix $A \in \mathbb{R}^{\omega n \times \omega n}$ formed by $\omega \times \omega$ blocks $A_{i,j} \in \mathbb{R}^{n\times n}$. Let $i_1 < \cdots < i_{|S_R|}$ be the elements in $S_R \subseteq \{1, \ldots, \omega\}$, and $j_1 < \cdots < j_{|S_C|}$ be the elements in $S_C \subseteq \{1, \ldots, \omega\}$, both in ascending order. Then $A_{S_R,S_C} \in \mathbb{R}^{|S_R|n\times|S_C|n}$ is defined as a block matrix

$$A_{S_R,S_C} := \begin{bmatrix} A_{i_1,j_1} & A_{i_1,j_2} & \cdots & A_{i_1,j_{|S_C|}} \\ A_{i_2,j_1} & A_{i_2,j_2} & \cdots & A_{i_2,j_{|S_C|}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{i_{|S_R|},j_1} & A_{i_{|S_R|},j_2} & \cdots & A_{i_{|S_R|},j_{|S_C|}} \end{bmatrix}.$$

For a diagonal block matrix $D = \operatorname{diag}(D_1, \ldots, D_\omega)$ and a set $S \subseteq \{1, \ldots, \omega\}$, we use the shorthand notation $D_S := \operatorname{diag}\left(D_{i_1}, D_{i_2}, \ldots, D_{i_{|S|}}\right)$, where $i_1 < \ldots < i_{|S|}$ are the elements in S.

Lemma 3.A.1. *Suppose A is a positive definite matrix in $\mathbb{S}^{\omega n}$ formed by $\omega \times \omega$ blocks $A_{i,j} \in \mathbb{R}^{n\times n}$. Assume that A is q-banded for an even positive integer q, i.e.,*

$$A_{i,j} = 0, \forall |i - j| > q/2.$$

Let $[a_0, b_0]$ ($b_0 > a_0 > 0$) be the smallest interval containing the spectrum $\sigma(A)$. Suppose $D = \operatorname{diag}(D_1, \ldots, D_\omega)$, where $D_i \in \mathbb{S}^n$ is positive semi-definite. Let $M = \left((A + D)^{-1}\right)_{S_R,S_C}$ as defined above, where $S_R, S_C \subseteq \{1, \ldots, \omega\}$. Then we have

$\|M\| \leq C\gamma^{\hat{d}}$, *where the coefficient* C, *the decay factor* γ, *and the distance* $\hat{d}$ *are given by*

$$C = \frac{2}{a_0}, \gamma = \left(\frac{\sqrt{\operatorname{cond}(A)} - 1}{\sqrt{\operatorname{cond}(A)} + 1}\right)^{2/q}, \hat{d} = \min_{i \in S_R, j \in S_c} |i - j|.$$

Here $\operatorname{cond}(A) = b_0/a_0$ *denotes the condition number of matrix* A.

Proof of Lemma 3.A.1. We first prove the lemma for the the special case where $D = 0$.

For the case $\hat{d} \neq 0$, write $\hat{d} = \upsilon q/2 + \kappa$ for integers υ, κ satisfying $\upsilon \geq 0, 1 \leq \kappa \leq q/2$. Following the same approach as the proof of Proposition 2.2 in Demko, Moss, and Smith, 1984, we see that there exists a polynomial p_υ of degree υ, where

$$\left\|A^{-1} - p_\upsilon(A)\right\| \leq \frac{1}{a_0} \cdot \frac{\left(1 + \sqrt{\operatorname{cond}(A)}\right)^2}{2\operatorname{cond}(A)} \gamma^{\hat{d}} \leq C\gamma^{\hat{d}},$$

where the last inequality holds because $\operatorname{cond}(A) \geq 1$.

Since p_v has degree $v < \frac{2\hat{d}}{q}$ and A is q-banded, the matrix $p_\upsilon(A)$ satisfies $(p_\upsilon(A))_{i,j} = 0$ for any $i \in S_R$ and $j \in S_C$. We then obtain

$$\|P\| = \left\|\left(A^{-1}\right)_{S_R, S_C}\right\| = \left\|\left(A^{-1} - p_\upsilon(A)\right)_{S_R, S_C}\right\| \leq \left\|A^{-1} - p_\upsilon(A)\right\| \leq C\gamma^{\hat{d}},$$

because 2-norm of a submatrix cannot be larger than that of the original matrix.

For the case $\hat{d} = 0$, as $\|P\| = \left\|(A^{-1})_{S_R, S_C}\right\| \leq \left\|A^{-1}\right\| = \frac{1}{a_0} \leq C$, the result trivially holds.

Now we show the general case (where $D_i \succeq 0$ for $1 \leq i \leq n$) through a reduction to the special case. Define a positive definite matrix $N := (a_0 I + D) \in \mathbb{S}^{n\omega}$, and then define matrix $H \in \mathbb{S}^{n\omega}$ as follows,

$$H = N^{-\frac{1}{2}}(A + D)N^{-\frac{1}{2}}.$$

We start by showing that $I \preceq H \preceq \frac{b_0}{a_0} \cdot I$. For any $x \in \mathbb{R}^{n\omega}$, we observe

$$\begin{aligned} x^\top Hx &= x^\top N^{-\frac{1}{2}} A N^{-\frac{1}{2}} x + x^\top N^{-\frac{1}{2}} D N^{-\frac{1}{2}} x \\ &\geq x^\top N^{-\frac{1}{2}} a_0 I N^{-\frac{1}{2}} x + x^\top N^{-\frac{1}{2}} D N^{-\frac{1}{2}} x \\ &= x^\top N^{-\frac{1}{2}} (a_0 I + D) N^{-\frac{1}{2}} x \\ &= \|x\|^2. \end{aligned}$$

For the other inequality, we have

$$\begin{aligned}
x^\top Hx &= x^\top N^{-\frac{1}{2}}AN^{-\frac{1}{2}}x + x^\top N^{-\frac{1}{2}}DN^{-\frac{1}{2}}x \\
&\le x^\top N^{-\frac{1}{2}}b_0IN^{-\frac{1}{2}}x + x^\top N^{-\frac{1}{2}}DN^{-\frac{1}{2}}x \\
&= x^\top N^{-\frac{1}{2}}(a_0I + D)N^{-\frac{1}{2}}x + (b_0 - a_0)x^\top N^{-1}x \\
&\le \|x\|^2 + \frac{b_0 - a_0}{a_0} \cdot \|x\|^2 \\
&= \frac{b_0}{a_0} \cdot \|x\|^2.
\end{aligned}$$

Thus $I \preceq H \preceq \frac{b_0}{a_0} \cdot I$, which gives $\operatorname{cond}(H) \le \frac{b_0}{a_0} = \operatorname{cond}(A)$. Note that H is also q-banded, so we can apply the result of the special case ($D_i = 0, i = 1, \cdots, n$) to obtain that

$$\left\|(H^{-1})_{S_R,S_C}\right\| \le 2\gamma_H^{\hat{d}} \le 2\gamma^{\hat{d}},$$

where $\gamma_H = \left(\frac{\sqrt{\operatorname{cond}(H)}-1}{\sqrt{\operatorname{cond}(H)}+1}\right)^{2/q} \le \gamma$. Using this inequality, we conclude that

$$\begin{aligned}
\|P\| = \left\|((A + D)^{-1})_{S_R,S_C}\right\| &= \left\|\left(N^{-\frac{1}{2}}H^{-1}N^{-\frac{1}{2}}\right)_{S_R,S_C}\right\| \\
&\le \left\|(a_0I + D_{S_R})^{-\frac{1}{2}}\right\| \cdot \left\|(H^{-1})_{S_R,S_C}\right\| \cdot \left\|(a_0I + D_{S_C})^{-\frac{1}{2}}\right\| \\
&\le \frac{1}{a_0}\left\|(H^{-1})_{S_R,S_C}\right\| \\
&\le C\gamma^{\hat{d}}.
\end{aligned}$$

Here we apply the fact that $\left\|(a_0I + D_S)^{-\frac{1}{2}}\right\| \le \frac{1}{\sqrt{a_0}}$ since $D_S \succeq 0$. □

Now we return to the proof of Theorem 3.1.2

Proof of Theorem 3.1.2. Let $e = (e_0^\top, \mu^\top, e_p^\top)^\top$ be a vector where $e_0, e_p \in \mathbb{R}^n$ and

$$\mu = [\mu_0, \mu_1, \ldots, \mu_{p-1}],$$

for $\mu_i \in \mathbb{R}^r, i = 0, 1, \ldots, p-1$. Let θ be an arbitrary real number. Define function $\hat{h} : \mathbb{R}^{(p-1)\times n} \times \mathbb{R}^n \times \mathbb{R}^{p\times r} \times \mathbb{R}^n \to \mathbb{R}_+$ as

$$\hat{h}(\hat{x}_{1:p-1}, \hat{x}_0, \hat{w}_{0:p-1}, \hat{x}_p) = \sum_{\tau=1}^{p-1} \hat{f}_\tau(\hat{x}_\tau) + \sum_{\tau=1}^{p} \hat{c}_\tau(\hat{x}_\tau, \hat{x}_{\tau-1}; \hat{w}_{\tau-1}).$$

To simplify the notation, we use $\hat{\zeta}$ to denote the tuple of system parameters, i.e.,

$$\hat{\zeta} := (\hat{x}_0, \hat{w}_{0:p-1}, \hat{x}_p).$$

From out construction, we know that $\hat{h}$ is μ-strongly convex in $\hat{x}_{1:p-1}$, so we use the decomposition $\hat{h} = \hat{h}_a + \hat{h}_b$, where

$$\hat{h}_a(\hat{x}_{1:p-1}, \hat{\zeta}) = \sum_{\tau=1}^{p-1} \frac{\mu}{2}\|\hat{x}_\tau\|^2 + \sum_{\tau=1}^{p} \hat{c}_\tau(\hat{x}_\tau, \hat{x}_{\tau-1}; \hat{w}_{\tau-1}),$$
$$\hat{h}_b(\hat{x}_{1:p-1}, \hat{\zeta}) = \sum_{\tau=1}^{p-1} \left(\hat{f}_\tau(\hat{x}_\tau) - \frac{\mu}{2}\|\hat{x}_\tau\|^2\right).$$

Since $\hat{\psi}(\hat{\zeta} + \theta e)$ is the minimizer of convex function $\hat{h}(\cdot, \hat{\zeta} + \theta e)$, we see that

$$\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e) = 0.$$

Taking the derivative with respect to θ gives that

$$\begin{aligned}
&\nabla^2_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e) \frac{d}{d\theta}\hat{\psi}(\hat{\zeta} + \theta e) \\
= &- \nabla_{\hat{x}_0}\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e)e_0 - \nabla_{\hat{x}_p}\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e)e_p \\
&- \sum_{\tau=0}^{p-1} \nabla_{w_\tau}\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e)\mu_\tau.
\end{aligned}$$

To simplify the notation, we define

$$\begin{aligned}
&M := \nabla^2_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e), \text{which is a } (p-1) \times (p-1) \text{ block matrix},\\
&R^{(0)} := -\nabla_{\hat{x}_0}\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e), \text{which is a } (p-1) \times 1 \text{ block matrix},\\
&R^{(p)} := -\nabla_{\hat{x}_p}\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e), \text{which is a } (p-1) \times 1 \text{ block matrix},\\
&K^{(\tau)} := -\nabla_{w_\tau}\nabla_{\hat{x}_{1:p-1}} \hat{h}(\hat{\psi}(\hat{\zeta} + \theta e), \hat{\zeta} + \theta e), \forall 0 \le \tau \le p-1, \text{which are}\\
&(p-1) \times 1 \text{ block matrices},
\end{aligned}$$

where in M, $R^{(0)}$, $R^{(p)}$, the block size is $n \times n$; in $K^{(\tau)}$, the block size is $n \times r$. Hence we can write

$$\frac{d}{d\theta}\hat{\psi}(\hat{\zeta} + \theta e) = M^{-1}\left(R^{(0)}e_0 + R^{(p)}e_p + \sum_{\tau=0}^{p-1} K^{(\tau)}\mu_\tau\right).$$

Recall that $R^{(0)}$, $R^{(p)}$ are $(p-1)\times 1$ block matrices with block size $n\times n$. $\{K^{(\tau)}\}_{0\le\tau\le p-1}$ are $(p-1) \times 1$ block matrices with block size $n \times r$. For $R^{(0)}$ and $K^{(0)}$, only the $(1,1)$-th blocks are non-zero. For $R^{(p)}$ and $K^{(p-1)}$, only the $(p-1,1)$-th blocks are non-zero. For $K^{(\tau)}, \tau = 1, \ldots, p-2$, only the $(\tau, 1)$-th and $(\tau+1, 1)$-th blocks are non-zero. Hence we see that

$$\frac{d}{d\theta}\hat{\psi}(\hat{\zeta} + \theta e)_h = (M^{-1})_{h,1}R^{(0)}_{1,1}e_0 + (M^{-1})_{h,p-1}R^{(p)}_{p-1,1}e_p$$

$$+ (M^{-1})_{h,1} K^{(0)}_{1,1} \mu_0 + (M^{-1})_{h,p-1} K^{(p-1)}_{p-1,1} \mu_{p-1}$$
$$+ \sum_{\tau=1}^{p-2} (M^{-1})_{h,\tau:\tau+1} K^{(\tau)}_{\tau:\tau+1,1} \mu_\tau.$$

Since the switching costs $c_\tau(\cdot,\cdot,\cdot), \tau = 1, \ldots, p$ are ℓ-strongly smooth, we know that the norms of

$$R^{(0)}_{1,1}, R^{(p)}_{p-1,1}, K^{(0)}_{1,1}, K^{(p-1)}_{p-1,1}, \text{ and } \{K^{(\tau)}_{\tau:\tau+1,1}\}_{1\le\tau\le p-2}$$

are all upper bounded by ℓ. Taking norm on both sides gives that

$$\begin{aligned}\left\|\frac{d}{d\theta}\hat{\psi}(\hat{\zeta}+\theta e)_h\right\| \le{}& \ell\left\|(M^{-1})_{h,1}\right\|\|e_0\| + \ell\left\|(M^{-1})_{h,p-1}\right\|\|e_p\| \\ &+ \ell\left\|(M^{-1})_{h,1}\right\|\|\mu_0\| + \ell\left\|(M^{-1})_{h,p-1}\right\|\|\mu_{p-1}\| \\ &+ \sum_{\tau=1}^{p-2} \ell\left\|(M^{-1})_{h,\tau:\tau+1}\right\|\|\mu_\tau\|. \end{aligned} \tag{3.20}$$

Note that M can be decomposed as $M = M_a + M_b$, where

$$M_a \coloneqq \nabla^2_{1:p-1}\hat{h}_a(\hat{\psi}(\hat{\zeta}+\theta e), \hat{\zeta}+\theta e),$$
$$M_b \coloneqq \nabla^2_{1:p-1}\hat{h}_b(\hat{\psi}(\hat{\zeta}+\theta e), \hat{\zeta}+\theta e).$$

Since M_a is block tri-diagonal and satisfies $(\mu + 2\ell)I \succeq M_a \succeq \mu I$, and M_b is block diagonal and satisfies $M_b \succeq 0$, we obtain the following with Lemma 3.A.1:

$$\left\|(M^{-1})_{h,1}\right\| \le \frac{2}{\mu}\lambda_0^{h-1}, \left\|(M^{-1})_{h,p-1}\right\| \le \frac{2}{\mu}\lambda_0^{p-h-1}, \text{ and } \left\|(M^{-1})_{h,\tau:\tau+1}\right\| \le \frac{2}{\mu}\lambda_0^{|h-\tau|-1},$$

where $\lambda_0 := (\sqrt{\mathrm{cond}(M_a)} - 1)/(\sqrt{\mathrm{cond}(M_a)} + 1) = 1 - 2\cdot\left(\sqrt{1+(2\ell/\mu)}+1\right)^{-1}$.

Substituting this into (3.20), we see that

$$\left\|\frac{d}{d\theta}\hat{\psi}(\hat{\zeta}+\theta e)_h\right\| \le C_0\left(\lambda_0^{h-1}\|e_0\| + \sum_{\tau=0}^{p-1}\lambda_0^{|h-\tau|-1}\|\mu_\tau\| + \lambda_0^{p-h-1}\|e_p\|\right),$$

where $C_0 = (2\ell)/\mu$.

Hence we obtain

$$\begin{aligned}\left\|\hat{\psi}(\hat{\zeta})_h - \hat{\psi}(\hat{\zeta}+e)_h\right\| ={}& \left\|\int_0^1 \frac{d}{d\theta}\hat{\psi}(\hat{\zeta}+\theta e)_h d\theta\right\| \\ \le{}& \int_0^1 \left\|\frac{d}{d\theta}\hat{\psi}(\hat{\zeta}+\theta e)_h\right\| d\theta \\ \le{}& C_0\left(\lambda_0^{h-1}\|e_0\| + \sum_{\tau=0}^{p-1}\lambda_0^{|h-\tau|-1}\|\mu_\tau\| + \lambda_0^{p-h-1}\|e_p\|\right).\end{aligned}$$

This finishes the proof. □

Proof of Lemma 3.1.3

Before we prove Lemma 3.1.3, we first show two technical lemmas. The first is about the properties of the optimal value/solution of an optimization problem.

Lemma 3.A.2. *Suppose function $f(x,y)$ is convex and L-strongly smooth in (x,y), μ-strongly convex in y, and continuously differentiable. Define functions $y^*(x) := \arg\min_y f(x,y)$ and $g(x) := \min_y f(x,y)$. Then, function y^* is $\frac{L}{\mu}$-Lipschitz and function g is $\left(L + \frac{L^2}{\mu}\right)$-strongly smooth.*

Proof of Lemma 3.A.2. Let $y^*(x) = \arg\min_y f(x,y)$. This function is well-defined since the strong convexity of $f(x,y)$ in y guarantees that $y^*(x)$ is unique. We see that for all x, x',

$$\nabla_y f(x, y^*(x)) = 0 \text{ and } \nabla_y f(x', y^*(x')) = 0.$$

Using these equalities, we obtain

$$\begin{aligned} 0 &= \langle y^*(x) - y^*(x'), \nabla_y f(x, y^*(x)) - \nabla_y f(x', y^*(x'))\rangle \\ &= \langle y^*(x) - y^*(x'), \nabla_y f(x, y^*(x)) - \nabla_y f(x, y^*(x'))\rangle \\ &\quad + \langle y^*(x) - y^*(x'), \nabla_y f(x, y^*(x')) - \nabla_y f(x', y^*(x'))\rangle \\ &\geq \mu \|y^*(x) - y^*(x')\|^2 - \|y^*(x) - y^*(x')\| \cdot \left\|\nabla_y f(x, y^*(x')) - \nabla_y f(x', y^*(x'))\right\|, \end{aligned}$$

where we used the fact that a μ-strongly convex function h satisfies

$$\langle a - b, \nabla h(a) - \nabla h(b)\rangle \geq \mu \|a - b\|^2, \forall a, b$$

and the Cauchy-Schwartz inequality in the last inequality. Since f is L-strongly smooth, we see that

$$\|y^*(x) - y^*(x')\| \leq \frac{1}{\mu}\left\|\nabla_y f(x, y^*(x')) - \nabla_y f(x', y^*(x'))\right\| \leq \frac{L}{\mu}\|x - x'\|,$$

which implies function y^* is $\frac{L}{\mu}$-Lipschitz.

Note that the gradient of g is given by

$$\nabla g(x) = \nabla_x f(x, y^*(x)) + \nabla_y f(x, y^*(x)) \frac{\partial y^*(x)}{\partial x} = \nabla_x f(x, y^*(x)),$$

because $\nabla_y f(x, y^*(x)) = 0$. Hence we obtain

$$\|\nabla g(x) - \nabla g(x')\|$$

$$
\begin{aligned}
&\le \|\nabla_x f(x, y^*(x)) - \nabla_x f(x', y^*(x'))\| \\
&\le \|\nabla_x f(x, y^*(x)) - \nabla_x f(x', y^*(x))\| + \|\nabla_x f(x', y^*(x)) - \nabla_x f(x', y^*(x'))\| \\
&\le L\|x - x'\| + L\|y^*(x) - y^*(x')\| \\
&\le \left(L + \frac{L^2}{\mu}\right) \|x - x'\|.
\end{aligned}
$$

□

The second technical lemma connects the induced 2-norm of a block matrix with the 2-norms of individual blocks.

Lemma 3.A.3. *Suppose A is a $\omega_1 \times \omega_2$ block matrix. Let A_{ij} denote the (i, j) th block of A, $1 \le i \le \omega_1, 1 \le j \le \omega_2$. The induced 2-norm of A is upper bounded by*

$$
\|A\| \le \left(\sum_{i=1}^{\omega_1} \sum_{j=1}^{\omega_2} \|A_{ij}\|^2 \right)^{\frac{1}{2}}.
$$

Proof of Lemma 3.A.3. For unit vector x, we have the following:

$$
\begin{aligned}
\|Ax\|^2 &= \sum_{i=1}^{\omega_1} \left\| \sum_{j=1}^{\omega_2} A_{ij} x_j \right\|^2 \\
&\le \sum_{i=1}^{\omega_1} \left(\sum_{j=1}^{\omega_2} \|A_{ij}\| \cdot \|x_j\| \right)^2 \\
&\le \sum_{i=1}^{\omega_1} \left(\sum_{j=1}^{\omega_2} \|A_{ij}\|^2 \right) \left(\sum_{j=1}^{\omega_2} \|x_j\|^2 \right) \\
&= \sum_{i=1}^{\omega_1} \sum_{j=1}^{\omega_2} \|A_{ij}\|^2,
\end{aligned}
$$

where we used the definition of the induced 2-norm in the first inequality and the Cauchy-Schwarz inequality in the second inequality. □

Now we come back to the proof of Lemma 3.1.3.

Proof of Lemma 3.1.3. To simplify the notation, we define the stacked state vector y, control vector v, and disturbance vector ζ as

$$
y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_p \end{bmatrix}, v = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{p-1} \end{bmatrix}, \zeta = \begin{bmatrix} \zeta_0 \\ \zeta_1 \\ \vdots \\ \zeta_{p-1} \end{bmatrix}.
$$

Recall that the transition matrix $\Phi(t_2, t_1)$ is defined as

$$\Phi(t_2, t_1) := \begin{cases} A_{t_2-1}A_{t_2-2}\cdots A_{t_1} & \text{if } t_2 > t_1 \\ I & \text{if } t_2 \leq t_1 \end{cases}.$$

Using this, we can express the state vector y as an affine function of initial state x, control v, and disturbance ζ:

$$y = S^x x + S^v v + S^\zeta \zeta, \tag{3.21}$$

where

$$S^\zeta := \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \Phi(t+1, t+1) & 0 & \cdots & 0 \\ \Phi(t+2, t+1) & \Phi(t+2, t+2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \Phi(t+p, t+1) & \Phi(t+p, t+2) & \cdots & \Phi(t+p, t+p) \end{bmatrix}, S^x = \begin{bmatrix} \Phi(t,t) \\ \Phi(t+1, t) \\ \Phi(t+2, t) \\ \vdots \\ \Phi(t+p, t) \end{bmatrix},$$

and $S^v = S^\zeta \cdot diag(B_t, \ldots, B_{t+p-1})$.

To simplify the notation, we use the shorthand $M := M(t, p)$ for the controllability matrix and

$$R^\zeta := [\Phi(t+p, t+1), \Phi(t+p, t+2), \ldots, \Phi(t+p, t+p)]$$

throughout the proof. Since p is greater than the controllability index d, we know M has full row rank. The dynamical constraints for (3.4) can be written as

$$Mv = z - \Phi(t+p, t)x - R^\zeta \zeta.$$

Because M has full row rank, we let $M^\dagger = M^\top \left(MM^\top\right)^{-1}$ be the Moore-Penrose pseudo-inverse of M. Let $V \in \mathbb{R}^{(mp)\times(mp-n)}$ be a matrix whose columns constitute an orthonormal basis of $ker(M)$. Then, we can express all feasible control vector v as

$$v = M^\dagger \left(z - \Phi(t+p, t)x - R^\zeta \zeta\right) + Vr, \tag{3.22}$$

where r is a free variable that can take any value in $\mathbb{R}^{mp-n}$.

Let F denote the objective function of ξ_t^p, i.e.,

$$F(y, v) := \left(\sum_{\tau=1}^{p-1} f_{t+\tau}(y_\tau) + c_{t+\tau}(v_{\tau-1})\right) + c_{t+p}(v_{p-1}).$$

Since we can express the state vector y and control vector v as linear functions of x, z, ζ and r, we can write the switching cost (3.4) as an unconstrained optimization problem

$$\min_{r \in \mathbb{R}^{mp-n}} F(y(x,z,\zeta,r), v(x,z,\zeta,r)), \tag{3.23}$$

where functions $y(x,z,\zeta,r)$ and $v(x,z,\zeta,r)$ are determined by

$$\begin{bmatrix} y \\ v \end{bmatrix} = \begin{bmatrix} S^x - S^v M^\dagger \Phi(t+p,t) & S^v M^\dagger & S^\zeta - S^v M^\dagger R^\zeta & S^v V \\ -M^\dagger \Phi(t+p,t) & M^\dagger & -M^\dagger R^\zeta & V \end{bmatrix} \cdot \begin{bmatrix} x \\ z \\ \zeta \\ r \end{bmatrix}. \tag{3.24}$$

Note that if $a \neq 1$, the following is due to Lemma 3.A.3:

$$\begin{aligned} \|S^\zeta\| &\leq \left(\sum_{i=1}^{p} \sum_{j=1}^{i} \|\phi(t+i,t+j)\|^2 \right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{p} \sum_{j=1}^{i} a^{2(i-j)} \right)^{\frac{1}{2}} \\ &= \frac{\sqrt{a^{2p+2} - (p+1)a^2 + p}}{\left|a^2 - 1\right|}. \end{aligned}$$

By Lemma 3.A.3, we also have

$$\|S^x\| \leq \sqrt{\frac{a^{2p+2} - 1}{a^2 - 1}}, \ \left\|M^\dagger\right\| \leq \frac{b}{\sigma^2} \cdot \sqrt{\frac{a^{2p} - 1}{a^2 - 1}}, \ \|S^v\| \leq b\|S^\zeta\|, \text{ and}$$
$$\|R^\zeta\| \leq \sqrt{\frac{a^{2p} - 1}{a^2 - 1}} \leq \frac{a^p - 1}{a - 1}.$$

Since the norm of a block matrix is upper bounded by the sum of norms of each block, we see that

$$\left\| \begin{bmatrix} S^x - S^v M^\dagger \Phi(t+p,t) & S^v M^\dagger & S^\zeta - S^v M^\dagger R^\zeta & S^v V \\ -M^\dagger \Phi(t+p,t) & M^\dagger & -M^\dagger R^\zeta & V \end{bmatrix} \right\| \leq C(p). \tag{3.25}$$

When $a \neq 1$, $C(p)$ is given by

$$\begin{aligned} C(p) = &\left(\frac{b(a^{p+1} + a - 2)}{\sigma^2 (a-1)} \cdot \sqrt{\frac{a^{2p} - 1}{a^2 - 1}} + \frac{1+b}{b} \right) \left(\frac{b\sqrt{\left(a^{2p+2} - (p+1)a^2 + p\right)}}{\left|a^2 - 1\right|} + 1 \right) \\ &+ \sqrt{\frac{a^{2p+2} - 1}{a^2 - 1}} - \frac{1}{b}. \end{aligned}$$

If $a = 1$, by Lemma 3.A.3, we see that

$$\|S^\zeta\| \leq \left(\sum_{i=1}^{p} \sum_{j=1}^{i} \|\phi(t+i,t+j)\|^2 \right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{p} \sum_{j=1}^{i} a^{2(i-j)} \right)^{\frac{1}{2}} = \sqrt{\frac{p(p+1)}{2}}.$$

By Lemma 3.A.3, we also see that

$$\|S^x\| \le \sqrt{p+1}, \left\|M^\dagger\right\| \le \frac{b}{\sigma^2} \cdot \sqrt{p}, \|S^v\| \le b\|S^\zeta\|, \|R^\zeta\| \le \sqrt{p}.$$

Therefore, for (3.25) to hold when $a = 1$, we need to set

$$C(p) = \left(\frac{b\sqrt{p}}{\sigma^2}\left(\sqrt{p}+2\right)+1\right)\left(1+b\sqrt{\frac{p(p+1)}{2}}\right)+\sqrt{p+1}\cdot\left(1+\sqrt{\frac{p}{2}}\right).$$

Since F is convex and strongly smooth in (x, u), and both x, u are affine functions of (y, z, r), $F(x(y, z, r), u(y, z, r))$ is convex and $\ell \cdot C(p)^2$-strongly smooth in (y, z, r). Since $F(x, u)$ is m_c-strongly convex in u, by (3.22), we have

$$\begin{aligned}\nabla_r^2 F(x(y,z,w,r), u(y,z,w,r)) &\succeq V^\intercal \nabla_u^2 F(x,u) V \\ &\succeq m_c I,\end{aligned}$$

where we used that $\|Vv\|_2 = \|v\|_2, \forall v \in \mathbb{R}^{mp-n}$ because the columns of V are orthonormal in the last inequality. Therefore, by Lemma 3.A.2, we know that (3.23) is convex and $L_2(p)$-strongly smooth in (y, z), where

$$L_2(p) := \ell \cdot C(p)^2 + \frac{\ell^2 \cdot C(p)^4}{m_c}.$$

By Lemma 3.A.2, we also know that the optimal solution of (3.23):

$$r^*(x, z, \zeta) := \arg\min_{r \in \mathbb{R}^{mp-n}} F(y(x, z, \zeta, r), v(x, z, \zeta, r))$$

is $\ell \cdot C(p)^2/m_c$-Lipschitz. By (3.24) and (3.25), we see that

$$\psi_t^p(x, \zeta, z) = \begin{bmatrix} S^x - S^v M^\dagger \Phi(t+p, t) & S^v M^\dagger & S^\zeta - S^v M^\dagger R^\zeta & S^v V \\ -M^\dagger \Phi(t+p, t) & M^\dagger & -M^\dagger R^\zeta & V \end{bmatrix} \cdot \begin{bmatrix} x \\ z \\ \zeta \\ r^*(x, z, \zeta) \end{bmatrix}$$

is $L_1(p)$-Lipschitz, where

$$L_1(p) = C(p)(1 + \ell \cdot C(p)^2/m_c).$$

Proof of Theorem 3.1.4

The proof of Theorem 3.1.4 is based on the decision-point transformation method that we introduce before.

Recall that we use d to the controllability index as defined in Definition 3.1.2. Without any loss of the generality, we only need to show the perturbation bound of $\psi_t^{t+p}((x_t, w_{t:t+p-1}, x_{t+p}))_{x_{t+h}}$ holds for $t = 0$, and the proof generalizes to other time t. Suppose h and p satisfy $ud \leq h < (u+1)d$ and $p = vd + r$, where $\delta, v, r \in \mathbb{N}$ and $0 \leq r < d$. Now we shall select the decision points as

$$x_0, x_d, \cdots, x_{(\delta-1)d}, x_h, x_{(\delta+2)d}, \cdots, x_{(v-1)d}, x_p,$$

which are also denoted by $x_{i_0}, \cdots, x_{i_{v-1}}$ for simplicity. Since the distance of any consecutive decision points falls in $[d, 2d)$, we can apply Lemma 3.1.3 to bound the strong smoothness of switching costs. In the transformed SOCO problem, the disturbance input of the $(\tau-1)$-th time period is a vector $\hat{w}_{\tau-1} \coloneqq w_{i_{\tau-1}:i_\tau - 1} \in \mathbb{R}^{n\times(i_\tau - i_{\tau-1})}$. Each stage cost $\Upsilon_{i_{\tau-1}}^{i_\tau}((x_{i_{\tau-1}}, \hat{w}_{\tau-1}, x_{i_\tau}))$ is convex and $L_2(i_\tau - i_{\tau-1})$-strongly smooth by Lemma 3.1.3, and is thus L_0-strongly smooth by definition. Recall that the solution of the transformed SOCO problem is denoted by $\hat{\psi}(x_0, \hat{w}, x_p)$. Then by Theorem 3.1.2 we have

$$\begin{aligned}
&\left\|\psi_0^p((x_0, w_{0:p-1}, x_p))_{x_h} - \psi_0^p((x_0', w_{0:p-1}', x_p')_{x_h}\right\| \\
&= \left\|\hat{\psi}(x_0, \hat{w}, x_p)_\delta - \hat{\psi}(x_0', \hat{w}', x_p')_\delta\right\| \\
&\leq C_0 \left(\lambda_0^{\delta-1}\left\|x_0 - x_0'\right\|_2 + \sum_{\tau=0}^{v-2} \lambda_0^{|\delta-\tau|-1}\left\|\hat{w}_\tau - \hat{w}_\tau'\right\|_2 + \lambda_0^{(v-1)-\delta-1}\left\|x_p - x_p'\right\|_2\right) \\
&\leq C_0 \left(\lambda_0^{\delta-1}\left\|x_0 - x_0'\right\|_2 + \sum_{\tau=0}^{v-2} \lambda_0^{|\delta-\tau|-1} \sum_{j=i_\tau}^{i_{\tau+1}-1} \left\|w_j - w_j'\right\|_2 + \lambda_0^{(v-1)-\delta-1}\left\|x_p - x_p'\right\|_2\right) \\
&\leq \frac{C_0}{\lambda_0} \left(\lambda^{i_\delta - i_0}\left\|x_0 - x_0'\right\|_2 + \sum_{\tau=0}^{v-2} \sum_{j=i_\tau}^{i_{\tau+1}-1} \lambda^{|j-i_\delta|}\left\|w_j - w_j'\right\|_2 + \lambda^{i_{v-1}-i_\delta}\left\|x_p - x_p'\right\|_2\right) \\
&= \bar{C}\left(\lambda^h\left\|x_0 - x_0'\right\| + \sum_{\tau=0}^{p-1} \lambda^{|h-\tau|}\left\|w_\tau - w_\tau'\right\| + \lambda^{p-h}\left\|x_p - x_p'\right\|\right). \qquad (3.26)
\end{aligned}$$

The last inequality holds because each interval is of length at most $(2d-1)$. Here the constants are

$$\begin{aligned}
&C_0 = \frac{2L_0}{m_c}, \lambda_0 = 1 - 2\cdot\left(\sqrt{1 + (2L_0/m_c)} + 1\right)^{-1}, \\
&\bar{C} = C_0/\lambda_0 = \frac{2L_0}{m_c}\left(1 - 2\cdot\left(\sqrt{1 + (2L_0/m_c)} + 1\right)^{-1}\right)^{-1}, \\
&\lambda = \left(1 - 2\left(\sqrt{1 + (2L_0/m_c)} + 1\right)^{-1}\right)^{\frac{1}{2d-1}}.
\end{aligned}$$

The proof of the perturbation bound of $\psi_t^{t+p}((x_t, w_{t:t+p-1}, \zeta_{t+p}); F)_{x_{t+h}}$ is quite similar. The only difference lies in the terminal cost, which can be addressed by a two-step approach.

Again, we set $t = 0$ without any loss of the generality. First, let ζ_p be fixed. We append $x_{\text{aux}} = 0$ to the end of the decision point sequence, and define a zero transition cost to the auxiliary state $\hat{c}_v(x_{t+p}, \hat{w}_v, x_{\text{aux}}) \equiv 0$ (note that $\hat{c}_v$ is trivially convex and L_0-strongly smooth). Denote the solution of the modified version of transformed SOCO problem by $\hat{\psi}'(x_t, \hat{w}, x_{\text{aux}})$, then by the same argument as above, we have

$$\begin{aligned}
&\left\|\psi_0^p((x_0, w_{0:p-1}, \zeta_p); F)_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \zeta_p); F)_{x_h}\right\| \\
&= \left\|\hat{\psi}'(x_0, \hat{w}, 0)_\delta - \hat{\psi}'(x_0', \hat{w}', 0)_\delta\right\| \\
&\leq \cdots \leq \bar{C}\left(\lambda^h \left\|x_0 - x_0'\right\| + \sum_{\tau=0}^{p-1} \lambda^{|h-\tau|}\left\|w_\tau - w_\tau'\right\|\right), \qquad (3.27)
\end{aligned}$$

where the constants are the same as previously defined. Now, we go on to bound the distance

$$\left\|\psi_0^p((x_0', w_{0:p-1}', \zeta_p); F)_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \zeta_p'); F)_{x_h}\right\|.$$

To simplify the notations, we define

$$\bar{x}_p \coloneqq \psi_0^p((x_0', w_{0:p-1}', \zeta_p); F)_{x_p}, \text{ and } \bar{x}_p' \coloneqq \psi_0^p((x_0', w_{0:p-1}', \zeta_p'); F)_{x_p}.$$

By the first-order optimality condition, we see that

$$\nabla_{x_p} \iota_0^p((x_0', w_{0:p-1}', \bar{x}_p)) + \nabla_x F(\bar{x}_p; \zeta_p) = 0, \qquad (3.28a)$$

$$\nabla_{x_p} \iota_0^p((x_0', w_{0:p-1}', \bar{x}_p')) + \nabla_x F(\bar{x}_p'; \zeta_p') = 0. \qquad (3.28b)$$

Note that the function

$$H(x_p) \coloneqq \iota_0^p((x_0', w_{0:p-1}', x_p)) + F(x_p; \zeta_p)$$

is a m_F-strongly convex in x_p. Thus, we see that

$$\begin{aligned}
m_F\left\|\bar{x}_p - \bar{x}_p'\right\|^2 &\leq \langle \nabla H(\bar{x}_p) - H(\bar{x}_p'), \bar{x}_p - \bar{x}_p'\rangle && (3.29a)\\
&= \langle \nabla_x F(\bar{x}_p'; \zeta_p') - \nabla_x F(\bar{x}_p'; \zeta_p), \bar{x}_p - \bar{x}_p'\rangle && (3.29b)\\
&\leq \left\|\nabla_x F(\bar{x}_p'; \zeta_p') - \nabla_x F(\bar{x}_p'; \zeta_p)\right\| \cdot \left\|\bar{x}_p - \bar{x}_p'\right\| && (3.29c)\\
&\leq \ell_F\left\|\zeta_p - \zeta_p'\right\| \cdot \left\|\bar{x}_p - \bar{x}_p'\right\|, && (3.29d)
\end{aligned}$$

where we use the properties of strongly convex functions in (3.29a); we use (3.28) in (3.29b); we use Cauchy-Schwartz inequality in (3.29c); we use Assumption 3.1.1 in (3.29d). Thus, we have shown that

$$\left\|\bar{x}_p - \bar{x}_p'\right\| \leq \frac{\ell_F}{m_F}\left\|\zeta_p - \zeta_p'\right\|. \tag{3.30}$$

Therefore, we obtain that

$$\begin{aligned}
&\left\|\psi_0^p((x_0', w_{0:p-1}', \zeta_p); F)_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \zeta_p'); F)_{x_h}\right\| \\
&= \left\|\psi_0^p((x_0', w_{0:p-1}', \bar{x}_p))_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \bar{x}_p'))_{x_h}\right\| && \text{(3.31a)}\\
&\leq \bar{C} \cdot \lambda^{p-h}\left\|\bar{x}_p - \bar{x}_p'\right\| && \text{(3.31b)}\\
&\leq \frac{\ell_F}{m_F} \cdot \bar{C} \cdot \lambda^{p-h}\left\|\zeta_p - \zeta_p'\right\|, && \text{(3.31c)}
\end{aligned}$$

where we use the principle of optimality in (3.31a); we use the perturbation bound with terminal constraint (3.26) in (3.31b); we use (3.30) in (3.31c).

Combining (3.31) with (3.27) by the triangle inequality gives

$$\begin{aligned}
&\left\|\psi_0^p((x_0, w_{0:p-1}, \zeta_p); F)_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \zeta_p'); F)_{x_h}\right\| \\
&\leq \left\|\psi_0^p((x_0, w_{0:p-1}, \zeta_p); F)_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \zeta_p); F)_{x_h}\right\| \\
&\quad + \left\|\psi_0^p((x_0', w_{0:p-1}', \zeta_p); F)_{x_h} - \psi_0^p((x_0', w_{0:p-1}', \zeta_p'); F)_{x_h}\right\| \\
&\leq C\left(\lambda^h\left\|x_0 - x_0'\right\| + \sum_{\tau=0}^{p-1}\lambda^{|h-\tau|}\left\|w_\tau - w_\tau'\right\| + \lambda^{p-h}\left\|\zeta_p - \zeta_p'\right\|\right).
\end{aligned}$$

Recall that $C := \max\left\{1, \frac{\ell_F}{m_F}\right\}$. This finishes the proof of Theorem 3.1.4.

3.B Proofs for the Perturbation Pipeline

Proof of Lemma 3.2.1

We have already shown (3.8) holds for all time step $t < T - k$ in the main body. For $t \geq T - k$, we see that

$$\begin{aligned}
e_t &= \left\|\psi_t^T\left(x_t, w_{t:T|t}; F_T\right) - \psi_t^T\left(x_t, w_{t:T}^*; F_T\right)\right\| && \text{(3.32a)}\\
&\leq \sum_{\tau=0}^{k}\left(\|x_t\| \cdot q_1(\tau) + q_2(\tau)\right)\rho_{t,\tau} && \text{(3.32b)}\\
&\leq \sum_{\tau=0}^{k}\left(\left(\frac{R}{C_3} + D_{x^*}\right) \cdot q_1(\tau) + q_2(\tau)\right)\rho_{t,\tau}, && \text{(3.32c)}
\end{aligned}$$

where we used the definition of per-step error e_t in (3.32a); we used the perturbation bound (3.6) specified by Property 3.2.1 in (3.32b); we used the assumption $x_t \in \mathcal{B}\left(x_t^*, \frac{R}{C_3}\right)$, $\left\|x_t^*\right\| \le D_{x^*}$, and the convention $\rho_{t,\tau} := 0$ if $t + \tau > T$ in (3.32c). Thus e_t also satisfies (3.8) for $t \ge T - k$.

Proof of Lemma 3.2.2

To simplify the notation, we will use $\psi_t^T(z_t)$ as a shorthand notation of $\psi_t^T(z_t, w_{t:T}^*; F_T)$ in the proof of Lemma 3.2.2, since the proof only relies on the perturbation bound (3.7).

Note that for any time step $t + 1$, by Lipschitzness of the dynamics we have

$$\begin{aligned}
\left\|x_{t+1} - \psi_t^T(x_t)_{x_{t+1}}\right\| &= \left\|g_t(x_t, u_t, w_t) - g_t\left(x_t, \psi_t^T(x_t)_{u_t}, w_t\right)\right\| \\
&\le L_g \left\|u_t - \psi_t^T(x_t)_{u_t}\right\| \\
&\le L_g e_t. \qquad (3.33)
\end{aligned}$$

Therefore, we can show the statement that $x_t \in \mathcal{B}\left(x_t^*, \frac{R}{C_3}\right)$ holds if $e_\tau \le R/(C_3^2 L_g), \forall \tau < t$ by induction. Note that this statement clearly holds for $t = 0$ since $x_0^* = x_0$. Suppose it holds for $0, 1, \ldots, t-1$. Then, we see that

$$\begin{aligned}
\left\|x_t - x_t^*\right\| &= \left\|x_t - \psi_0^T(x_0)_{x_t}\right\| \\
&\le \left\|x_t - \psi_{t-1}^T(x_{t-1})_{x_t}\right\| + \sum_{i=1}^{t-1} \left\|\psi_{t-i}^T(x_{t-i})_{x_t} - \psi_{t-i-1}^T(x_{t-i-1})_{x_t}\right\| \\
&\le \left\|x_t - \psi_{t-1}^T(x_{t-1})_{x_t}\right\| + \sum_{i=1}^{t-1} q_3(i) \left\|x_{t-i} - \psi_{t-i-1}^T(x_{t-i-1})_{x_{t-i}}\right\| \qquad (3.34a) \\
&\le \sum_{i=0}^{t-1} q_3(i) \left\|x_{t-i} - \psi_{t-i-1}^T(x_{t-i-1})_{x_{t-i}}\right\| \qquad (3.34b) \\
&\le L_g \sum_{i=0}^{t-1} q_3(i) e_{t-i-1}, \qquad (3.34c)
\end{aligned}$$

where in (3.34a), we apply the perturbation bound (3.7) specified by Property 3.2.1. To see why it can be applied, note that for $i \in [1, t-1]$, x_{t-i-1} satisfies $x_{t-i-1} \in \mathcal{B}\left(x_{t-i-1}^*, \frac{R}{C_3}\right)$ by the induction assumption, thus we have

$$\psi_{t-i-1}^T(x_{t-i-1})_{x_{t-i}} \in \mathcal{B}\left(x_{t-i}^*, R\right)$$

because $q_3(1) \le \sum_{\tau=0}^{\infty} q_3(\tau) \le C_3$. Therefore, we can apply the perturbation bound (3.7) specified by Property 3.2.1 to compare the optimization solution vectors

$\psi_{t-i}^T(x_{t-i})$ and $\psi_{t-i}^T\left(\psi_{t-i-1}^T(x_{t-i-1})_{x_{t-i}}\right)$, and by the principle of optimality, we see that

$$\psi_{t-i}^T\left(\psi_{t-i-1}^T(x_{t-i-1})_{x_{t-i}}\right)_{x_t} = \psi_{t-i-1}^T(x_{t-i-1})_{x_t}.$$

We also used $q_3(0) \geq 1$ in (3.34b) and (3.33) in (3.34c). Recall that we assume $e_{t-i} \leq \frac{R}{C_3^2 L_g}$. Substituting this into (3.34) gives that

$$\left\|x_t - x_t^*\right\| \leq L_g \cdot \frac{R}{C_3^2 L_g} \sum_{i=0}^{t-1} q_3(i) \leq \frac{R}{C_3}.$$

Hence we have shown $x_t \in \mathcal{B}\left(x_t^*, \frac{R}{C_3}\right)$ holds if $e_\tau \leq R/(C_3^2 L_g), \forall \tau < t$ by induction. An implication of this result is that $x_t \in \mathcal{B}\left(x_t^*, \frac{R}{C_3}\right)$ holds for all $t \leq T$ if $e_t \leq R/(C_3^2 L_g)$ holds for all $t < T$.

Similar with (3.34), we see the following inequality holds for all $t < T$ if $e_t \leq R/(C_3^2 L_g), \forall t < T$:

$$\begin{aligned}
\left\|u_t - u_t^*\right\| &= \left\|u_t - \psi_0^T(x_0)_{u_t}\right\| \\
&\leq \left\|u_t - \psi_t^T(x_t)_{u_t}\right\| + \sum_{i=0}^{t-1} \left\|\psi_{t-i}^T(x_{t-i})_{u_t} - \psi_{t-i-1}^T(x_{t-i-1})_{u_t}\right\| \\
&\leq \left\|u_t - \psi_t^T(x_t)_{u_t}\right\| + \sum_{i=0}^{t-1} q_3(i)\left\|x_{t-i} - \psi_{t-i-1}^T(x_{t-i-1})_{x_{t-i}}\right\| \\
&\leq e_t + L_g \sum_{i=0}^{t-1} q_3(i) e_{t-i-1}, \qquad (3.35)
\end{aligned}$$

where the second inequality holds for the same reason as (3.34a).

By (3.34), we see that

$$\begin{aligned}
\left\|x_t - x_t^*\right\|^2 &\leq L_g^2 \left(\sum_{i=0}^{t-1} q_3(i) e_{t-i-1}\right)^2 \\
&\leq L_g^2 \left(\sum_{i=0}^{t-1} q_3(i)\right) \cdot \left(\sum_{i=0}^{t-1} q_3(i) e_{t-i-1}^2\right) \qquad (3.36a) \\
&\leq C_3 L_g^2 \left(\sum_{i=0}^{t-1} q_3(i) e_{t-i-1}^2\right), \qquad (3.36b)
\end{aligned}$$

where we use the Cauchy-Schwarz inequality in (3.36a), and $\sum_{i=0}^{t-1} q_3(i) \leq C_3$ in (3.36b).

Similarly, by (3.35), we see that

$$\begin{aligned}
\left\|u_t - u_t^*\right\|^2 &\leq \left(e_t + L_g \sum_{i=0}^{t-1} q_3(i) e_{t-i-1}\right)^2 \\
&\leq \left(1 + L_g^2 \sum_{i=0}^{t-1} q_3(i)\right) \cdot \left(e_t^2 + \sum_{i=0}^{t-1} q_3(i) e_{t-i-1}^2\right) \qquad (3.37a)\\
&\leq \left(1 + C_3 L_g^2\right) \cdot \left(e_t^2 + \sum_{i=0}^{t-1} q_3(i) e_{t-i-1}^2\right), \qquad (3.37b)
\end{aligned}$$

where we use the Cauchy-Schwarz inequality in (3.37a), and we use $\sum_{i=0}^{t-1} q_3(i) \leq C_3$ in (3.37b).

Summing (3.36) and (3.37) over time steps t gives that

$$\begin{aligned}
&\sum_{t=1}^{T} \left\|x_t - x_t^*\right\|^2 + \sum_{t=0}^{T-1} \left\|u_t - u_t^*\right\|^2 \\
&\leq C_3 L_g^2 \sum_{t=1}^{T} \left(\sum_{i=0}^{t-1} q_3(i) e_{t-i-1}^2\right) + \left(1 + C_3 L_g^2\right) \cdot \sum_{t=0}^{T-1} \left(e_t^2 + \sum_{i=0}^{t-1} q_3(i) e_{t-i-1}^2\right) \\
&\leq \left(1 + 2C_3 L_g^2\right) \cdot (1 + C_3) \cdot \sum_{t=0}^{T-1} e_t^2, \qquad (3.38)
\end{aligned}$$

where we rearrange the terms and use $\sum_{j=0}^{\infty} q_3(j) \leq C_3$ in the last inequality.

Since the cost function $f_t(\cdot, \cdot; w_t^*)$ and $F_T(\cdot; w_T^*)$ are nonnegative, convex, and ℓ-smooth in their inputs, by Lemma F.2 in Lin, Hu, Shi, et al., 2021, we see that the following inequality holds for arbitrary $\eta > 0$:

$$\begin{aligned}
&\text{cost}(\mathsf{ALG}) - \text{cost}(\mathsf{OPT}) \\
&\leq \left(\sum_{t=0}^{T-1} f_t(x_t, u_t; w_t^*) + F_T(x_T; w_T^*)\right) - \left(\sum_{t=0}^{T-1} f_t(x_t^*, u_t^*; w_t^*) + F_T(x_T^*; w_T^*)\right) \\
&\leq \eta \left(\sum_{t=0}^{T-1} f_t(x_t^*, u_t^*; w_t^*) + F_T(x_T^*; w_T^*)\right) \\
&\quad + \frac{\ell}{2}\left(1 + \frac{1}{\eta}\right)\left(\sum_{t=1}^{T} \left\|x_t - x_t^*\right\|^2 + \sum_{t=0}^{T-1} \left\|u_t - u_t^*\right\|^2\right) \qquad (3.39a)\\
&\leq \eta \cdot \text{cost}(\mathsf{OPT}) + \left(1 + \frac{1}{\eta}\right) \cdot \frac{\ell}{2} \cdot \left(1 + 2C_3 L_g^2\right) \cdot (1 + C_3) \cdot \sum_{t=0}^{T-1} e_t^2 \qquad (3.39b)\\
&= \eta \cdot \text{cost}(\mathsf{OPT}) + \frac{1}{\eta} \cdot \frac{\ell}{2} \cdot \left(1 + 2C_3 L_g^2\right) \cdot (1 + C_3) \cdot \sum_{t=0}^{T-1} e_t^2
\end{aligned}$$

$$+\frac{\ell}{2}\cdot\left(1+2C_3L_g^2\right)\cdot(1+C_3)\cdot\sum_{t=0}^{T-1}e_t^2, \tag{3.39c}$$

where we apply Lemma F.2 in Lin, Hu, Shi, et al., 2021 in (3.39a), and we use (3.38) in (3.39b). Setting the tunable weight η in (3.39c) to be

$$\eta=\left(\frac{\frac{\ell}{2}\cdot\left(1+2C_3L_g^2\right)\cdot(1+C_3)\cdot\sum_{t=0}^{T-1}e_t^2}{\mathrm{cost}(\mathsf{OPT})}\right)^{\frac{1}{2}}$$

gives that

$$\begin{aligned}&\mathrm{cost}(\mathsf{ALG})-\mathrm{cost}(\mathsf{OPT})\\ &\le\sqrt{\left(\frac{\ell}{2}\cdot\left(1+2C_3L_g^2\right)\cdot(1+C_3)\right)\cdot\mathrm{cost}(\mathsf{OPT})\cdot\sum_{t=0}^{T-1}e_t^2}\\ &+\frac{\ell}{2}\cdot\left(1+2C_3L_g^2\right)\cdot(1+C_3)\cdot\sum_{t=0}^{T-1}e_t^2.\end{aligned} \tag{3.40}$$

This finishes the proof of Lemma 3.2.2.

Proof of Theorem 3.2.3

We first use induction to show that the following two conditions holds for all time steps $t < T$:

$$x_t\in\mathcal{B}\left(x_t^*,\frac{R}{C_3}\right), \tag{3.41a}$$

$$e_t\le\sum_{\tau=0}^{k}\left(\left(\frac{R}{C_3}+D_{x^*}\right)\cdot q_1(\tau)+q_2(\tau)\right)\rho_{t,\tau}+2R\left(\left(\frac{R}{C_3}+D_{x^*}\right)\cdot q_1(k)+q_2(k)\right). \tag{3.41b}$$

At time step 0, (3.41a) holds because $x_0 = x_0^*$, and (3.41b) holds by Lemma 3.2.1 and the assumption on the terminal cost F_k of MPC_k.

Suppose (3.41a) and (3.41b) hold for all time steps $\tau < t$. For time step t, by the assumption on the prediction errors $\rho_{t,\tau}$ and prediction horizon k in Theorem 3.2.3, we know that $e_\tau \le \frac{R}{C_3^2L_g}$ holds for all $\tau < t$ because (3.41b) holds for all $\tau < t$. Thus, we know that (3.41a) holds for time step t by Lemma 3.2.2. Then, since (3.41a) holds for time step t, and the terminal cost F_{t+k} of MPC_k is set to be the indicator function of some state $\bar{x}(w_{t+k|t})$ that satisfies $\bar{x}(w_{t+k|t})\in\mathcal{B}(x_{t+k}^*,R)$ if $t < T-k$, we know (3.41b) also holds for time step t by Lemma 3.2.1. This finishes the induction proof of (3.41).

To simplify the notation, let $R_0 \coloneqq \frac{R}{C_3} + D_{x^*}$. Note that (3.41b) implies that

$$\begin{aligned} e_t^2 \leq\ & \left(\sum_{\tau=0}^{k} \left(R_0 \cdot q_1(\tau) + q_2(\tau)\right) + 2R\left(R_0 + 1\right)\right) \\ & \cdot \left(\sum_{\tau=0}^{k} \left(R_0 \cdot q_1(\tau) + q_2(\tau)\right)\rho_{t,\tau}^2 + 2R\left(R_0 \cdot q_1(k)^2 + q_2(k)^2\right)\right) \qquad \text{(3.42a)} \\ \leq\ & \left(R_0 C_1 + C_2 + 2R(R_0 + 1)\right) \\ & \cdot \left(\sum_{\tau=0}^{k-1} \left(R_0 \cdot q_1(\tau) + q_2(\tau)\right)\rho_{t,\tau}^2 + (2R+1)\left(R_0 \cdot q_1(k)^2 + q_2(k)^2\right)\right), \qquad \text{(3.42b)} \end{aligned}$$

where we use the Cauchy-Schwarz inequality in (3.42a); we use the assumptions that $\sum_{\tau=0}^{k} q_1(\tau) \leq C_1$, $\sum_{\tau=0}^{k} q_2(\tau) \leq C_2$, and $\rho_{t,\tau} \leq 1$ in (3.42b).

Since (3.41) and (3.42) holds for all time steps $t < T$, we can apply Lemma 3.2.2 to obtain that

$$\mathrm{cost}(\mathsf{MPC}_k) - \mathrm{cost}(\mathsf{OPT}) \leq \sqrt{\mathrm{cost}(\mathsf{OPT}) \cdot E_0} + E_0,$$

where

$$\begin{aligned} E_0 \coloneqq\ & \left(R_0 C_1 + C_2 + 2R(R_0 + 1)\right) \\ & \cdot \left(\sum_{\tau=0}^{k-1} \left(R_0 \cdot q_1(\tau) + q_2(\tau)\right)P(\tau) + (2R+1)\left(R_0 \cdot q_1(k)^2 + q_2(k)^2\right)T\right). \end{aligned}$$

This finishes the proof of Theorem 3.2.3.

Proof of Theorem 3.2.4

By Theorem 3.3 in Lin, Hu, Shi, et al., 2021, we know Property 3.2.1 holds under Assumption 3.2.1 for arbitrary R and $q_1(t) = 0, q_2(t) = H_1\lambda_1^t$, and $q_3(t) = H_1\lambda_1^t$, where $H_1 = H_1(\mu, \ell, d, \sigma, a, b, b') > 0$ is some constant, and

$$\lambda_1 = \lambda_1(\mu, \ell, d, \sigma, a, b, b') \in (0, 1)$$

is the decay rate. Here, H_1 corresponds to C and λ_1 corresponds to λ in Theorem 3.3 in Lin, Hu, Shi, et al., 2021.

By setting $R \coloneqq \max\left\{D_{x^*}, \frac{2L_g H_1^3}{(1-\lambda_1)^3}\right\}$, we guarantee that the terminal state 0 of MPC_k is always in the closed ball $\mathcal{B}(x_{t+k}^*, R)$, and the condition

$$\sum_{\tau=0}^{k}\left(\left(\frac{R}{C_3} + D_{x^*}\right)\cdot q_1(\tau) + q_2(\tau)\right)\rho_{t,\tau} + 2R\left(\left(\frac{R}{C_3} + D_{x^*}\right)\cdot q_1(k) + q_2(k)\right) \le \frac{R}{C_3^2 L_g}$$

holds once $k \ge \ln\left(\frac{4H_1^3 L_g}{(1-\lambda_1)^2}\right)/\ln(1/\lambda_1)$ because $\rho_{t,\tau} \le 1$. Therefore, we can apply Theorem 3.2.3 to finish the proof of Theorem 3.2.4.

3.C Proof Outline for Networked Online Convex Optimization

In this section, we outline the major novelties in our proofs for the tighter exponentially decaying local perturbation bound in Theorem 3.3.4 and the main competitive ratio bound for LPC in Theorem 3.3.5. The full details of the proofs of these results are deferred to Section 3.D.

Refined Analysis of Perturbation Bounds

We begin by outlining the four-step structure we use to prove Theorem 3.3.4. Our goal is to highlight the main ideas, while deferring a detailed proof to Section 3.D. Throughout this section, we consider a Networked OCO problem instance $p \in \mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$.

Step 1. Establishing first order equations

The exponentially decaying local perturbation bounds (Definition 3.3.4) study how the optimal solution $\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \{z_\tau^u\}\right)$ reacts as we change the fixed actions on the boundary $\{z_\tau^u | (\tau, u) \in \partial N_{(t,v)}^{(k,r)}\}$. We combine these fixed actions into a single vector and denote it as

$$\zeta \coloneqq \{z_\tau^u | (\tau, u) \in \partial N_{(t,v)}^{(k,r)}\}.$$

Since we do not perturb the previous actions $\{y_{t-1}^u\}_{u \in N_v^r}$ in the local exponentially decaying perturbation bound (Definition 3.3.4), we introduce the shorthand notation $\psi(\zeta) \coloneqq \psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \zeta\right)$ to simplify the notations when the context is clear. The objective function of the corresponding optimization problem

$$\sum_{\tau=t}^{t+k-1}\left(f_\tau^{(N_v^{r-1})}\left(x_\tau^{(N_v^r)}\right) + c_\tau^{(N_v^r)}\left(x_\tau^{(N_v^r)}, x_{\tau-1}^{(N_v^r)}\right)\right), \text{ where}$$

$$x_{t-1}^u = y_{t-1}^u, \forall u \in N_v^r,\ x_\tau^u = z_\tau^u, \forall (\tau, u) \in \partial N_{(t,v)}^{(k,r)},$$

is a function of the previous actions $\{y_{t-1}^u\}_{u\in N_v^r}$, the fixed actions on the boundaries ζ, and the free variables $\{x_\tau^u | (\tau, u) \in N_{(t,v)}^{(k-1,r-1)}\}$. Since we do not perturb the previous actions, we express the objective function as $\hat{h}\left(\{x_\tau^u | (\tau, u) \in \partial N_{(t,v)}^{(k-1,r-1)}\}, \zeta\right)$. To avoid writing the period index t repeatedly, for arbitrary $i \in \{0, 1, \ldots, k\}$, we use $\hat{x}_i^u$ to denote the decision variable x_{t-1+i}^u for $u \in N_v^{r-1}$ at period $t-1+i$ and introduce the notations

$$\hat{z}_i^u = \begin{cases} z_{t-1+i}^u & \text{for } u \in \partial N_v^r \text{ if } i \in \{1, \ldots, k-1\}, \\ z_{t+k-1}^u & \text{for } u \in N_v^r \text{ if } i = k. \end{cases}$$

To simplify the notations, we also use the shorthand

$$\hat{x}_i := \hat{x}_i^{(N_v^{r-1})},\ \hat{z}_i := \hat{z}_i^{(\partial N_v^r)},\ \text{for } i \in \{1, \ldots, k-1\},$$
$$\hat{z}_k := \hat{z}_k^{(N_v^r)}.$$

Using these notations, we can rewrite the objective function as $\hat{h}(\hat{x}_{1:k-1}, \hat{z}_{1:k})$. The main lemma for this step is the following.

Lemma 3.C.1. *Let $e = (e_1, \ldots, e_k)$ be the perturbation vector such that e_i shares the same dimensions as $\hat{z}_i$ for $i \in \{1, 2, \ldots, k\}$. Given $\theta \in \mathbb{R}$, optimization parameter ζ and perturbation vector e, we have*

$$\frac{d}{d\theta}\psi(\zeta + \theta e) = M^{-1}\left(R^{(k)} e_k + \sum_{\tau=1}^{k-1} K^{(\tau)} e_\tau\right), \text{ where}$$

$$\begin{aligned} M &:= \nabla^2_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e), \\ R^{(k)} &:= -\nabla_{\hat{z}_k} \nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e), \\ K^{(\tau)} &:= -\nabla_{\hat{z}_\tau} \nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e), \text{ for } \tau \in \{1, \ldots, k-1\}. \end{aligned}$$

The proof for Lemma 3.C.1 using first order conditions at the global optimal solution for the convex function $\hat{h}(\cdot, \zeta + \theta e)$ and then takes derivatives with respect to to θ. See Appendix 3.D for a proof.

Step 2: Decomposing M^{-1} as infinite series

M is a hierarchical block matrix with the first level of dimension $(k-1) \times (k-1)$. When fixing the first level indices (i.e., period indices) in M, the lower level matrices are non-zero only if their difference in the period indices is ≤ 1. Hence we decompose M to $M = D + A$, where D is a block diagonal matrix and A is a tri-diagonal block matrix with zero matrix on the diagonal. Each diagonal block in D is a graph-induced banded matrix, which captures the Hessian of $\hat{h}$ in a single period. Denote each diagonal block as $D_{i,i}$ for $1 \leq i \leq k-1$. Further, for $2 \leq i \leq k-1$, $A_{i,i-1}$ (similarly $A_{i,i+1}$) captures the temporal correlation of an individual's action between consecutive periods. Under this decomposition, we know that each $D_{i,i}$ is invertible because $f_{t-1+i}^{(N_v^r)}$ is strictly strongly convex for $i = 1, \dots, k-1$. Thus, we know that $D^{-1} = \operatorname{diag}(\{D_{i,i}^{-1}\}_{1 \leq i \leq k-1})$. And M^{-1} can be expressed as

$$M^{-1} = (D + A)^{-1} = D^{-1}(I + AD^{-1})^{-1}.$$

For the ease of notation, we denote $I + AD^{-1}$ by P. Note that P is not necessarily a symmetric matrix. Nevertheless, under technical conditions on P's eigenvalues, we have the following power series expansion (Shin, Zavala, and Anitescu, 2020). The details are presented in the Lemma 3.C.2 in Section 3.D.

Lemma 3.C.2. *Let $\rho(\cdot)$ denote the spectral radius of a matrix. Under the condition $\mu > 2\ell_T$, we have $\rho(I - P) < 1$, and*

$$P^{-1} = \sum_{\tau \geq 0} (I - P)^{\tau}. \tag{3.43}$$

To understand the power series in (3.43), consider the special case where each block $A_{i,i+1} = A_{i+1,i} = \ell_T \cdot I$ for $i = 1, \dots, k-2$, and $D_{i,i} = Q$. Denote $J := P - I = AD^{-1}$. Then, we have $J_{i,i} = 0$, $J_{i,i-1} = J_{i,i+1} = \ell_T Q^{-1}$, $J_{i,j} = 0$ when $|i - j| > 1$. Intuitively, J captures the "correlation over actions" after one period. More generally, for $q \geq 0$ and any two period indices τ', τ,

$$(J^q)_{\tau',\tau} = \ell_T^q Q^{-q} b(q, \tau, \tau'),$$

where $b(q, \tau, \tau')$ is a coefficient that is upper bounded by 2^q and it equals to zero if $q < |\tau - \tau'|$.

Given that Q is a graph-induced banded matrix, Q^{-1} satisfies exponential-decay properties, which makes it plausible that $Q^{-\tau}$ is an exponential decay matrix with a slower rate.

For the general case where diagonal blocks $D_{i,i}$ are not identical and $A_{i,i+1}, A_{i+1,i}$ are not equal to $\ell_T \cdot I$, we need to bound terms such as

$$\left\| (A_{i_0,i_1} D_{i_1,i_1}^{-1} A_{i_1,i_2} D_{i_2,i_2}^{-1} \cdots A_{i_{q-1},i_q} D_{i_q,i_q}^{-1})_{u,v} \right\|$$

where $i_0, i_1, i_2, \cdots, i_q \in \{1, \dots, k-1\}$. This is the goal of Step 3.

Step 3: Showing exponential-decay properties are preserved through matrix multiplications

The goal of this step is to establish that, given an underlying graph, a product of a general class of exponential decay matrices still exhibits exponential decay property under technical conditions.

Lemma 3.C.3. *Given any graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ and integers $d, q \geq 1$, suppose a sequence of block matrices $\{A_i \in \mathbb{R}^{|\mathcal{V}'|d \times |\mathcal{V}'|d}\}_{1 \leq i \leq q}$ all satisfy exponential decay properties with respect to $\mathcal{G}'$, i.e. there exists $C_i \geq 0$, and $0 \leq \lambda < 1$, such that*

$$\left\|(A_i)_{u,v}\right\| \leq C_i \lambda^{d_{\mathcal{G}'}(u,v)} \quad \text{for any nodes } u, v \in \mathcal{V}'.$$

Suppose there exists a constant λ' that satisfies $1 > \lambda' > \lambda$ and

$$\tilde{a} \coloneqq \sum_{k=0}^{\infty} (\frac{\lambda}{\lambda'})^k (\sup_{u \in \mathcal{V}'} |\partial N_u^k|) < \infty.$$

Then, the product matrix $\prod_{i=1}^q A_i$ satisfies exponential decay properties with decay rate λ', i.e.,

$$\left\|(\prod_{i=1}^q A_i)_{u,v}\right\| \leq C'(\lambda')^{d_{\mathcal{G}'}(u,v)} \quad \text{for any nodes } u, v \in \mathcal{V}',$$

where $C' = (\tilde{a})^q \prod_{i=1}^q C_i$.

Intuitively, Lemma 3.C.3 shows that the exponential decay properties of matrices are preserved through matrix multiplications, though the product matrix has a worse decay factor λ'. A proof of Lemma 3.C.3 can be found in the Section 3.D.

Step 4: Establishing exponential decay properties of matrix M^{-1}

The last step of the proof is to study the properties of M. To accomplish this, we first show that, for period indices $i, j \geq 1$, J^ℓ has the following properties:

- $(J^q)_{i,j} = 0$ if $q < |i - j|$ or $q - |i - j|$ is odd.
- $(J^q)_{i,j}$ is a summation of terms $\prod_{k=1}^q A_{i_{k-1},i_k} D_{i_k,i_k}^{-1}$ where $i_0, i_1, i_2, \cdots, i_q \in \{1, \ldots, k-1\}$ and the number of such terms is bounded by $\binom{q}{(q-|i-j|)/2}$.

We formally state and prove the above properties in Section 3.D. We can further use Theorem 3.C.3 on block matrices $A_{i_{k-1},i_k} D_{i_k,i_k}^{-1}$, which gives the following lemma.

Lemma 3.C.4. *Recall* $\gamma_S := \frac{\sqrt{1+(\Delta\ell_S/\mu)}-1}{\sqrt{1+(\Delta\ell_S/\mu)}+1}$. *Suppose there exists a constant* γ_S' *such that* $1 > \gamma_S' > \gamma_S$ *and* $b := \sum_{\gamma=0}^{\infty}(\frac{\gamma_S}{\gamma_S'})^\gamma h(\gamma) < \infty$. *Given positive integers* q, i, j *and* $u, v \in \mathcal{V}$, *we have*

$$\left\|((J^q)_{i,j})_{u,v}\right\| \le \binom{q}{(q-|i-j|)/2}\left(b\frac{2\ell_T}{\mu}\right)^q (\gamma_S')^{d_{\mathcal{G}}(u,v)}.$$

Intuitively speaking, Lemma 3.C.4 bounds the correlations over actions for node u at period $t-1+i$ and action for node q at period $t-1+j$. We present its proof in the Section 3.D.

Recall that, for $1 \le i, j \le k-1$,

$$(M^{-1})_{i,j} = D_{i,i}^{-1}\sum_{q\ge 0}((-J)^q)_{i,j}\,.$$

With the exponential decaying bounds on matrix J^q, we can thus bound the entries that corresponds to node v on the RHS of the first order equations derived in Lemma 3.C.1. We state this result formally in Lemma 3.C.5 and present the proof in Appendix 3.D.

Lemma 3.C.5. *Given* $1 \le i, j \le k-1$, *for any* $e \in \mathbb{R}^{|\partial N_v^r|\times n}$, *we have*

$$\left\|\left((M^{-1})_{i,j}K_j^{(j)}e\right)_v\right\| \le C_1\rho_T^{|i-j|}\sum_{u\in\partial N_v^r}\rho_S^{d_{\mathcal{G}}(v,u)-1}\|e_u\|,$$

and for any $e' \in \mathbb{R}^{|N_v^r|\times n}$,

$$\left\|\left((M^{-1})_{i,k-1})R_{k-1}^{(k)}e'\right)_v\right\| \le C_1\rho_T^{|i-(k-1)|+1}\sum_{u\in N_v^r}\rho_S^{d_{\mathcal{G}}(v,u)}\|e_u'\|,$$

where $\rho_T = \frac{4\tilde{a}\ell_T}{\mu}$ *and* $\rho_S = (1+b_1+b_2)\gamma_S$. *We let*

$$C_1 = \max\{\frac{a^2}{2\tilde{a}(1-4\tilde{a}\ell_T/\mu)}, \frac{2a^2\Delta\ell_S/\mu}{\gamma_S(1+b_1+b_2)(1-4\tilde{a}\ell_T/\mu)}\}.$$

Using the first-order equations derived in Lemma 3.C.1 in Step 1, we can bound the entry in $\frac{d}{d\theta}\psi(\zeta+\theta e)$ that corresponds to node v. Then, we conclude Theorem 3.3.4 by integrating over θ. The detailed proofs of the results we state in this section can be found in Section 3.D.

From Perturbation to Competitive Ratio

We now show how to use the exponentially decaying local perturbation bounds proven in the previous section to prove our competitive ratio bounds in Theorem 3.3.5. Our starting point is the assumption that the exponentially decaying local perturbation bound in Definition 3.3.4 holds for a class of Networked OCO problems, which is established using the proof approach outlined in the Section 3.3.

As we discussed in Section 3.3, our proof contains two key parts: (i) we bound the per-period error of LPC (Lemma 3.C.7); and (ii) show that the per-period error does not accumulate to be unbounded (Theorem 3.C.8).

A key observation that enables the above analysis approach is that the aggregation of the local per-period error made by each agent at x_t^v can be viewed as a global per-period error in the joint global decision x_t. Following this observation, we first introduce a global perturbation bound that focuses on the global decision x_t rather than the local decisions x_t^v. Recall that f_t denotes the global hitting cost (see Section 3.3). Define the optimization problem that solves the optimal global decision trajectory from period $(t-1)$ to period $(t+q-1)$

$$\begin{aligned} \tilde{\psi}_t^q(y, z) = \arg\min_{x_{t:t+q-1}} \sum_{\tau=t}^{t+q-1} \left(f_\tau(x_\tau) + c_\tau(x_\tau, x_{\tau-1})\right) \\ \text{s.t. } x_{t-1} = y, x_{t+q-1} = z, \end{aligned} \tag{3.44}$$

and another one that solves the optimal global decision trajectory from period $(t-1)$ to the end of the game

$$\begin{aligned} \tilde{\psi}_t(y) = \arg\min_{x_{t:T}} \sum_{\tau=t}^{H} \left(f_\tau(x_\tau) + c_\tau(x_\tau, x_{\tau-1})\right) \\ \text{s.t. } x_{t-1} = y. \end{aligned} \tag{3.45}$$

The following global perturbation bound can be derived from Theorem 3.1 in Lin, Hu, Shi, et al. (2021):

Theorem 3.C.6 (Global Perturbation Bound)**.** *For any tuple $(\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$, consider the problem class $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$. The following global perturbation bounds hold for optimization problems* (3.44) *and* (3.45)*: For arbitrary $y \in D_t$ and $z \in D_{t+q-1}$, we have*

$$\left\|\tilde{\psi}_t^q(y, z)_{t_0} - \tilde{\psi}_t^q(y', z')_{t_0}\right\| \le C_G \rho_G^{t_0-t+1} \|y - y'\| + C_G \rho_G^{t+q-1-t_0} \|z - z'\|,$$

for $t_0 \in \{t, \dots, t+q-1\}$ *and*

$$\left\|\tilde{\psi}_t(y)_{t_0} - \tilde{\psi}_t(y')_{t_0}\right\| \leq C_G \rho_G^{t_0-t+1}\|y - y'\|$$

for $t_0 \in \{t, \dots, H-1\}$*. Here,* $\rho_G = 1 - 2 \cdot \left(\sqrt{1 + \frac{2\ell_T}{\mu}} + 1\right)^{-1}$ *and* $C_G = \frac{2\ell_T}{\mu}$*.*

To make the concept of *per-period error* at period t rigorous, we formally define it as the distance between the actual next decision made by LPC and the clairvoyant optimal next decision from previous decision x_{t-1} to the end of the game:

Definition 3.C.1 (Per-period error magnitude)**.** *Consider applying LPC on an instance p of the Networked OCO problem (Definition 3.3.1). At period t, given its previous decision* x_{t-1}*, LPC picks* $x_t \in D_t$*. We define the per-period error magnitude* e_t *as*

$$e_t := \left\|x_t - \tilde{\psi}_t(x_{t-1})_t\right\|. \tag{3.46}$$

Using the exponentially decaying local perturbation bound in Definition 3.3.4, we show the per-period error of LPC decays exponentially with respect to prediction horizon length k and observation radius r. This result is stated formally in Lemma 3.C.7, and the proof can be found in Section 3.D.

Lemma 3.C.7. *For any tuple* $v = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$*, suppose the exponentially decaying local perturbation bound (Definition 3.3.4) holds with the decay factors* ρ_T *and* ρ_S *for* $\mathcal{P}(v)$*. Suppose we apply LPC with prediction horizon k and observation radius r to an instance* $p \in \mathcal{P}(v)$*. Then, the per-period error* e_t *satisfies*

$$\begin{aligned} e_t^2 = {} & O\left(h(r)^2 \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2k}\rho_G^{2k}\right) \cdot \left\|x_{t-1} - x_{t-1}^*\right\|^2 \\ & + O\left(h(r)^2 \cdot \rho_S^{2r}\right) \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} f_\tau(x_\tau^*) + O\left(C_3(r)^2 \cdot \rho_T^{2k}\right) f_{t+k-1}(x_{t+k-1}^*), \end{aligned}$$

where $C_3(r) := \sum_{\gamma=0}^{r} h(\gamma) \cdot \rho_S^\gamma$ *and* ρ_G *is defined in Theorem 3.C.6. We use* $\{x_t^*\}_{t\in[H]}$ *to denote the offline optimal global decision trajectory in the problem instance p.*

Using the global perturbation bound in Theorem 3.C.6, we show the total squared distance between LPC and the offline optimal decision trajectories can be upper bounded by the sum of per-time-step errors of LPC in Theorem 3.C.8. The proof can be found in Section 3.D.

Theorem 3.C.8. *For any tuple $\nu = (\mu, \ell_f, \ell_T, \ell_S, \Delta, h) \in \Upsilon$, consider an instance p of Networked OCO problem in the class $\mathcal{P}(\nu)$. Let $x_0, x_1^*, x_2^*, \ldots, x_H^*$ denote the offline optimal global decision trajectory and $x_0, x_1, x_2, \ldots, x_H$ denote the decision trajectory of LPC. The trajectory of LPC satisfies that*

$$\sum_{t=1}^{H} \left\|x_t - x_t^*\right\|^2 \le \frac{C_0^2}{(1-\rho_G)^2} \sum_{t=1}^{H} e_t^2,$$

where $C_0 := \max\{1, C_G\}$, ρ_G is defined in Theorem 3.C.6, and $\{e_t\}_{t\in[H]}$ are the per-period error magnitudes defined in (3.46).

To understand the bound in Theorem 3.C.8, we can set all per-period error magnitude e_t to be zero except a single period τ. We see the impact of e_τ on the total squared distance $\sum_{t=1}^{T} \left\|x_t - x_t^*\right\|^2$ is up to some constant factor of e_τ. This is because the impact of e_τ on $\left\|x_t - x_t^*\right\|$ decays exponentially as t increases from τ to H.

By substituting the per-period error bound in Lemma 3.C.7 into Theorem 3.C.8, one can bound the total squared distance $\sum_{t=1}^{H} \left\|x_t - x_t^*\right\|^2$ by the offline optimal cost, which can be converted to the competitive ratio bound in Theorem 3.3.5.

Generalization to Inexact Predictions

To show the performance bound in Theorem 3.3.10, we follow a similar procedure with the proof outline for the exact prediction case discussed in Appendix 3.C. The key observation is that Theorem 3.C.8 still applies for the inexact prediction case, and we only need to bound the per-period error magnitude (Definition 3.C.1) of LPC with inexact predictions. We state this bound formally in Lemma 3.C.9 below and defer its proof to Section 3.D.

Lemma 3.C.9. *For any tuple $\nu = (\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h) \in \tilde{\Upsilon}$, consider a problem instance $p \in \tilde{P}(\nu)$. For LPC with inexact predictions (Algorithm 3), the per-period error e_t satisfies*

$$\begin{aligned} e_t^2 &= O\left(h(r)^2 \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2k} \rho_G^{2k}\right) \cdot \left\|x_{t-1} - x_{t-1}^*\right\|^2 \\ &\quad + O\left(h(r)^2 \cdot \rho_S^{2r}\right) \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} f_\tau(x_\tau^*) + O\left(C_3(r)^2 \cdot \rho_T^{2k}\right) f_{t+k-1}(x_{t+k-1}^*) \\ &\quad + O\left((1+\Delta^2) C_3(r)^2\right) \cdot \mathsf{PredictionError}_{p,(t,k)}, \end{aligned}$$

where $C_3(r) := \sum_{\gamma=0}^{r} h(\gamma) \cdot \rho_S^{\gamma}$ and $\mathsf{PredictionError}_{p,(t,k)}$ is defined as

$$\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \left(\sum_{u\in\mathcal{V}} \left\|\omega_{\tau|t}^u - (\omega_\tau^u)^*\right\|^2 + \sum_{u\in\mathcal{V}} \left\|\alpha_{\tau|t}^u - (\alpha_\tau^u)^*\right\|^2 + \sum_{e\in\mathcal{E}} \left\|\beta_{\tau|t}^e - (\beta_\tau^e)^*\right\|^2 \right).$$

By substituting the per-period error bound in Lemma 3.C.9 into Theorem 3.C.8, one can bound the total squared distance $\sum_{t=1}^{H} \left\|x_t - x_t^*\right\|^2$ by the offline optimal cost, which can be converted to the performance bound of LPC with inexact predictions in Theorem 3.3.10.

3.D Proofs for Networked Online Convex Optimization

Proof of Theorem 3.3.3 and Theorem 3.3.9

We begin with a technical lemma. Recall that for any positive integer m, $\mathbb{S}^m$ denotes the set of all symmetric $m \times m$ real matrices.

Lemma 3.D.1. *For a graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$, suppose A is a positive definite matrix in $\mathbb{S}^{\sum_{i \in \mathcal{V}'} p_i}$ formed by $|\mathcal{V}'| \times |\mathcal{V}'|$ blocks, where the (i, j)-th block has dimension $p_i \times p_j$, i.e., $A_{i,j} \in \mathbb{R}^{p_i \times p_j}$. Assume that A is q-banded for an even positive integer q; i.e.,*

$$A_{i,j} = 0, \ \forall d_{\mathcal{G}'}(i, j) > q/2.$$

Let a_0 denote the smallest eigenvalue value of A, and b_0 denote the largest eigenvalue value of A. Assume that $b_0 \geq a_0 > 0$. Suppose $D = diag(D_1, \ldots, D_{|\mathcal{V}'|})$, where $D_i \in \mathbb{S}^{p_i}$ is positive semi-definite. Let $M = \left((A + D)^{-1}\right)_{S_R, S_C}$, where $S_R, S_C \subseteq \{1, \ldots, |\mathcal{V}'|\}$. Then we have $\|M\| \leq C\gamma^{\hat{d}}$, where

$$C = \frac{2}{a_0}, \ \gamma = \left(\frac{\sqrt{cond(A)} - 1}{\sqrt{cond(A)} + 1}\right)^{2/q}, \ \hat{d} = \min_{i \in S_R, j \in S_c} d_{\mathcal{G}'}(i, j).$$

Here $cond(A) = b_0/a_0$ denotes the condition number of matrix A.

We can show Lemma 3.D.1 using the same method as Lemma B.1 in Lin, Hu, Shi, et al. (2021). We only need to note that even when the size of blocks are not identical, the m th power of a q-banded matrix is a qm-banded matrix for any positive integer m.

With the help of Lemma 3.D.1, we can proceed to show a local perturbation bound on a general $\mathcal{G}'$ in Theorem 3.D.2, where $\mathcal{G}'$ can be different from the network $\mathcal{G}$ of agents in Section 3.3. Compared with Theorem 3.1 in Lin, Hu, Shi, et al. (2021), Theorem 3.D.2 is more general because it considers a general network of decision variables while Theorem 3.1 in Lin, Hu, Shi, et al. (2021) only consider the special case of a line graph. Although Theorem 3.D.2 does not consider the temporal dimension which features in the local perturbation bound defined in Definition 3.3.4, we will use it to show Theorems 3.3.3 and 3.3.9 later by redefining the variables from two perspectives.

Theorem 3.D.2. *For a network $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ with undirected edges, suppose that each node $v \in \mathcal{V}'$ is associated with a decision vector $\hat{x}_v \in \mathbb{R}^{p_v}$ and a cost function $\hat{f}_v : \mathbb{R}^{p_v} \times \mathcal{W}_v \to \mathbb{R}_{\geq 0}$, and each edge $e = (u, v) \in \mathcal{E}'$ is associated with an edge cost $\hat{c}_e : \mathbb{R}^{p_v} \times \mathbb{R}^{p_u} \times \mathcal{W}_e \to \mathbb{R}_{\geq 0}$. Here $\mathcal{W} := (\{\mathcal{W}_v\}_{v \in \mathcal{V}'}, \{\mathcal{W}_e\}_{e \in \mathcal{E}'})$ denote the set of all possible disturbances on cost functions, where $\mathcal{W}_v$ and $\mathcal{W}_e$ are convex compact subsets of $\mathbb{R}^q$ for all $v \in \mathcal{V}', e \in \mathcal{E}'$. Assume that: For all $v \in \mathcal{V}'$, $\hat{f}_v(\hat{x}_v; w_v)$ is μ-strongly convex in $\hat{x}_v$ under any fixed disturbance w_v; For all $e = (u, v) \in \mathcal{E}'$, $\hat{c}_e(\hat{x}_u, \hat{x}_v; w_e)$ is convex and ℓ-smooth in $(\hat{x}_u, \hat{x}_v)$ under any fixed disturbance w_e, i.e., for all $v \in \mathcal{V}'$ and $e = (u, v) \in \mathcal{E}'$, we have*

$$\nabla^2_{\hat{x}_v} \hat{f}_v(\hat{x}_v; w_v) \succeq 0, \forall w_v \in \mathcal{W}_v, \text{ and } 0 \preceq \nabla^2_{(\hat{x}_u, \hat{x}_v)} \hat{c}_e(\hat{x}_u, \hat{x}_v; w_e) \preceq \ell I_{p_v + p_u}, \forall w_e \in \mathcal{W}_e.$$

For some subset $S \subset \mathcal{V}'$, define

$$E_0 := \{(u, v) \in \mathcal{E}' \mid u, v \in \mathcal{V}' \setminus S\}, \text{ and } E_1 := \{(u, v) \in \mathcal{E}' \mid u \in \mathcal{V}' \setminus S, v \in S\}.$$

For any disturbance vector $w \in \mathcal{W}$ and boundary vector $y := \{y_v\}_{v \in S} \in \mathcal{Y} := \{\mathcal{Y}_v\}_{v \in S}$ where $\mathcal{Y}_v$ is a convex subset of $\mathbb{R}^{p_v}$ for all $v \in S$, let $\psi(w, y)$ denote the optimal solution defined as

$$\begin{aligned}\psi(w, y) := \underset{\{\hat{x}_v \mid v \in \mathcal{V}' \setminus S\}}{\arg\min} & \sum_{v \in \mathcal{V}' \setminus S} \hat{f}_v(\hat{x}_v; w_v) + \sum_{(u,v) \in E_0} \hat{c}_{(u,v)}(\hat{x}_u, \hat{x}_v; w_{(u,v)}) \\ & + \sum_{(u,v) \in E_1} \hat{c}_{(u,v)}(\hat{x}_u, y_v; w_{(u,v)}).\end{aligned}$$

Let Γ_v be a set that contains all possible v th entry of the optimal solution, i.e., $\Gamma_v \supseteq \{\psi(w, y)_v \mid w \in \mathcal{W}, y \in \mathcal{Y}\}$. We additionally assume that for any $v \in \mathcal{V}'$ and $e = (u, v) \in \mathcal{E}'$, the cost functions satisfies

$$\left\|\nabla_{w_v} \nabla_{\hat{x}_v} \hat{f}_v(\hat{x}_v; w_v)\right\| \leq \ell_w, \text{ if } \hat{x}_v \in \Gamma_v, \tag{3.47}$$

$$\left\|\nabla_{w_e} \nabla_{(\hat{x}_u, \hat{x}_v)} \hat{c}_e(\hat{x}_u, \hat{x}_v; w_e)\right\| \leq \ell_w, \text{ if } \hat{x}_u \in \Gamma_u \text{ and } \hat{x}_v \in \Gamma_v. \tag{3.48}$$

Then, we have that for any vertex $u_0 \in \mathcal{V}' \setminus S$, the following inequality holds for all $y \in \mathcal{Y}, w \in \mathcal{W}$

$$\begin{aligned}&\left\|\psi(w, y)_{u_0} - \psi(w', y')_{u_0}\right\| \\ &\leq C \sum_{v \in S} \lambda^{d_{\mathcal{G}'}(u_0, v) - 1} \left\|y_v - y'_v\right\| \\ &\quad + C_w \left(\sum_{e \in E_0 \cup E_1} \lambda^{d_{\mathcal{G}'}(u_0, e) - 1} \left\|w_e - w'_e\right\| + \sum_{v \in \mathcal{V}' \setminus S} \lambda^{d_{\mathcal{G}'}(u_0, v)} \left\|w_v - w'_v\right\| \right), \text{ where}\end{aligned}$$

$$C := (2\ell\Delta')/\mu, C_w := (2\ell_w)/\mu, \text{ and } \lambda := 1 - 2 \cdot \left(\sqrt{1 + (\Delta'\ell/\mu)} + 1\right)^{-1}.$$

Here, Δ' denote the maximum degree of any vertex $v \in \mathcal{V}'$ in graph $\mathcal{G}'$. For $e = (u, v) \in \mathcal{E}'$, we define $d_{\mathcal{G}'}(u_0, e) := \min\{d_{\mathcal{G}'}(u_0, u), d_{\mathcal{G}'}(u_0, v)\}$.

To show Theorem 3.D.2, we establish the first order equations using the optimality conditions of the optimization problem. The equation shows the way how a small perturbation affects the optimal solution vector relates to the inverse of the Hessian matrix of the objective function. Note that the objective function has a special structure that it is the sum of local costs, and each local cost can couple at most two neighboring decision variables. Thus, the Hessian is a 1-banded matrix with respect to graph $\mathcal{G}'$, so we can leverage Lemma 3.D.1 to show the exponentially decaying property of its entries. Combining the exponentially decaying property of the inverse of the Hessian with the first order equations finishes the proof.

Proof of Theorem 3.D.2. Let $e = [\pi^\top, \epsilon^\top]^\top$ be a perturbation vector where $\epsilon = \{\epsilon_v\}_{v\in S}$ for $\epsilon_v \in \mathbb{R}^{p_v}$ and $\pi = \left(\{\pi_v\}_{v\in\mathcal{V}'\backslash S}, \{\pi_e\}_{e\in E_0\cup E_1}\right)$, for $\pi_e \in \mathbb{R}^q$. Let θ be an arbitrary real number. Define function $\hat{h} : \mathbb{R}^{\sum_{v\in\mathcal{V}'\backslash S} p_v} \times \mathbb{R}^{(|\mathcal{V}'|-|S|+|E_0|+|E_1|)\times q} \times \mathbb{R}^{\sum_{v\in S} p_v} \longrightarrow \mathbb{R}_{\geq 0}$ as

$$\hat{h}(\hat{x}, w, y) = \sum_{v\in\mathcal{V}'\backslash S} \hat{f}_v(\hat{x}_v) + \sum_{(u,v)\in E_0} \hat{c}_{(u,v)}(\hat{x}_u, \hat{x}_v; w_{(u,v)}) + \sum_{(u,v)\in E_1} \hat{c}_{(u,v)}(\hat{x}_u, y_v; w_{(u,v)}).$$

To simplify the notation, we use ζ to denote the tuple of system parameters, i.e.,

$$\zeta := (w, y).$$

From our construction, we know that $\hat{h}$ is μ-strongly convex in x, so we use the decomposition $\hat{h} = \hat{h}_a + \hat{h}_b$, where

$$\begin{aligned}
\hat{h}_a(\hat{x}, \zeta) = &\sum_{v\in\mathcal{V}'\backslash S} \frac{\mu}{2}\|\hat{x}_v\|^2 + \sum_{(u,v)\in E_0} \hat{c}_{(u,v)}(\hat{x}_u, \hat{x}_v; w_{(u,v)}) \\
&+ \sum_{(u,v)\in E_1} \hat{c}_{(u,v)}(\hat{x}_u, y_v; w_{(u,v)}), \\
\hat{h}_b(\hat{x}, \zeta) = &\sum_{v\in\mathcal{V}'\backslash S} \left(\hat{f}_v(\hat{x}_v; w_v) - \frac{\mu}{2}\|\hat{x}_v\|^2\right).
\end{aligned}$$

Since $\psi(\zeta + \theta e)$ is the minimizer of convex function $\hat{h}(\cdot, \zeta + \theta e)$, we see that

$$\nabla_{\hat{x}}\hat{h}(\psi(\zeta + \theta e), \zeta + \theta e) = 0.$$

Taking the derivative with respect to θ gives that

$$\begin{aligned}\nabla_{\hat{x}}^2 \hat{h}(\psi(\zeta+\theta e), \zeta+\theta e)\frac{d}{d\theta}\psi(\zeta+\theta e) = & -\sum_{v\in S}\nabla_{y_v}\nabla_{\hat{x}}\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e)\epsilon_v \\ & -\sum_{e\in E_1\cup E_2}\nabla_{w_e}\nabla_{\hat{x}}\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e)\pi_e \\ & -\sum_{v\in\mathcal{V}'\setminus S}\nabla_{w_v}\nabla_{\hat{x}}\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e)\pi_v.\end{aligned}$$

To simplify the notation, we define

$$\begin{aligned} M &:= \nabla_{\hat{x}}^2\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e), \text{ a } |\mathcal{V}'\setminus S|\times|\mathcal{V}'\setminus S| \text{ block matrix,} \\ R^{(v)} &:= -\nabla_{y_v}\nabla_{\hat{x}}\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e), \forall v\in S, \quad |\mathcal{V}'\setminus S|\times 1 \text{ block matrices,} \\ K^{(e)} &:= -\nabla_{w_e}\nabla_{\hat{x}}\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e), \forall e\in E_0\cup E_1, \quad |\mathcal{V}'\setminus S|\times 1 \text{ block matrices,} \\ Q^{(v)} &:= -\nabla_{w_v}\nabla_{\hat{x}}\hat{h}(\psi(\zeta+\theta e), \zeta+\theta e), \forall v\in \mathcal{V}\setminus S, \quad |\mathcal{V}'\setminus S|\times 1 \text{ block matrices,}\end{aligned}$$

where in M, the block size is $p_u\times p_v, \forall (u,v)\in(\mathcal{V}'\setminus S)^2$; in $R^{(v)}$, the block size is $p_u\times p_v, \forall u\in\mathcal{V}'\setminus S$; in $K^{(e)}$ and $Q^{(v)}$, the block size is $p_u\times q, \forall u\in\mathcal{V}'\setminus S$. Hence we can write

$$\frac{d}{d\theta}\psi(\zeta+\theta e) = M^{-1}\left(\sum_{v\in S}R^{(v)}\epsilon_v + \sum_{e\in E_1\cup E_2}K^{(e)}\pi_e + \sum_{v\in\mathcal{V}'\setminus S}Q^{(v)}\pi_v\right).$$

Recall that $\{R^{(v)}\}_{v\in S}$ are $|\mathcal{V}'\setminus S|\times 1$ block matrices with block size $p_u\times p_v, \forall u\in\mathcal{V}'\setminus S$; $\{K^{(e)}\}_{e\in E_0\cup E_1}$ and $\{Q^{(v)}\}_{v\in\mathcal{V}'\setminus S}$ are $|\mathcal{V}'\setminus S|\times 1$ block matrices with block size $p_u\times q, \forall u\in\mathcal{V}'\setminus S$. Let $N(v)$ denote the set of neighbors of vertex v on $\mathcal{G}'$. For $R^{(v)}, v\in S$, the $(u,1)$-th block can be non-zero only if $u\in(\mathcal{V}'\setminus S)\cap N(v)$. For $K^{(e)}, e\in E_0\cup E_1$, the $(u,1)$-th block can be non-zero only if $u\in e$ and $u\in\mathcal{V}'\setminus S$. For $Q^{(v)}, v\in\mathcal{V}'\setminus S$, the $(u,1)$-th block can be non-zero only if $u=v$. Hence we see that

$$\begin{aligned}\frac{d}{d\theta}\psi(\zeta+\theta e)_{u_0} = & \sum_{v\in S}(M^{-1})_{u_0,(\mathcal{V}'\setminus S)\cap N(v)}R^{(v)}_{(\mathcal{V}'\setminus S)\cap N(v),1}\epsilon_v \\ & + \sum_{e\in E_0\cup E_1}(M^{-1})_{u_0,\{u\in e|u\in\mathcal{V}'\setminus S\}}K^{(e)}_{\{u\in e|u\in\mathcal{V}'\setminus S\},1}\pi_e \\ & + \sum_{v\in\mathcal{V}'\setminus S}(M^{-1})_{u_0,v}Q^{(v)}_v\pi_v.\end{aligned}$$

Since we assume the edge costs $c_e(\hat{x}_u,\hat{x}_v;w_e)$ are ℓ-strongly smooth in $(\hat{x}_u,\hat{x}_v)$, we know that the norms of $\{R^{(v)}_{(\mathcal{V}'\setminus S)\cap N(v),1}\}_{v\in S}$ are upper bounded by $\Delta'\ell$. Similarly, by (3.47) and (3.47), we know that the norms of $\{K^{(\tau)}_{\{u\in e|u\in\mathcal{V}'\setminus S\},1}\}_{e\in E_0\cup E_1}$ and

$\{Q_v^{(v)}\}_{\mathcal{V}'\backslash S}$ are upper bounded by ℓ_w. Taking norms on both sides of the above equation gives that

$$\begin{aligned}\left\|\frac{d}{d\theta}\psi(\zeta+\theta e)_{u_0}\right\| \leq\ & \sum_{v\in S}\ell\left\|(M^{-1})_{u_0,(\mathcal{V}'\backslash S)\cap N(v)}\right\|\|\epsilon_v\| \\ &+\sum_{e\in E_0\cup E_1}\ell_w\left\|(M^{-1})_{u_0,\{u\in e|u\in\mathcal{V}'\backslash S\}}\right\|\|\pi_e\| \\ &+\sum_{v\in\mathcal{V}'\backslash S}\ell_w\left\|(M^{-1})_{u_0,v}\right\|\|\pi_v\|. \end{aligned} \tag{3.49}$$

Note that M can be decomposed as $M = M_a + M_b$, where

$$\begin{aligned} M_a &:= \nabla_{\hat{x}}^2\hat{h}_a(\psi(\zeta+\theta e),\zeta+\theta e),\\ M_b &:= \nabla_{\hat{x}}^2\hat{h}_b(\psi(\zeta+\theta e),\zeta+\theta e).\end{aligned}$$

Since M_a is block tri-diagonal and satisfies $(\mu+\Delta'\ell)I \succeq M_a \succeq \mu I$, and M_b is block diagonal and satisfies $M_b \succeq 0$, we obtain the following using Lemma 3.D.1: $\left\|(M^{-1})_{u_0,(\mathcal{V}'\backslash S)\cap N(v)}\right\| \leq \frac{2}{\mu}\lambda^{d_{\mathcal{G}'}(u_0,v)-1}$, $\left\|(M^{-1})_{u_0,\{u\in e|u\in\mathcal{V}'\backslash S\}}\right\| \leq \frac{2}{\mu}\lambda^{d_{\mathcal{G}'}(u_0,e)-1}$, and $\left\|(M^{-1})_{u_0,v}\right\| \leq \frac{2}{\mu}\lambda^{d_{\mathcal{G}'}(u_0,v)}$.

where $\lambda := (\sqrt{cond(M_a)}-1)/(\sqrt{cond(M_a)}+1) = 1-2\cdot\left(\sqrt{1+(2\ell/\mu)}+1\right)^{-1}$.

Substituting this into (3.49), we see that

$$\begin{aligned}\left\|\frac{d}{d\theta}\psi(\zeta+\theta e)_{u_0}\right\| &\leq C\sum_{v\in S}\lambda^{d_{\mathcal{G}'}(u_0,v)-1}\|\epsilon_v\| \\ &\quad + C_w\left(\sum_{e\in E_0\cup E_1}\lambda^{d_{\mathcal{G}'}(u_0,e)-1}\|\pi_e\|+\sum_{v\in\mathcal{V}'\backslash S}\lambda^{d_{\mathcal{G}'}(u_0,v)}\|\pi_v\|\right),\end{aligned}$$

where $C = (2\ell)/\mu$ and $C_w = (2\ell_w)/\mu$.

Finally, by integration we can complete the proof of Theorem 3.D.2:

$$\begin{aligned}&\left\|\psi(\zeta)_{u_0}-\psi(\zeta+e)_{u_0}\right\| \\ =&\left\|\int_0^1\frac{d}{d\theta}\psi(\zeta+\theta e)_{u_0}d\theta\right\| \\ \leq&\int_0^1\left\|\frac{d}{d\theta}\psi(\zeta+\theta e)_{u_0}\right\|d\theta \\ \leq& C\sum_{v\in S}\lambda^{d_{\mathcal{G}'}(u_0,v)-1}\|\epsilon_v\|+C_w\left(\sum_{e\in E_0\cup E_1}\lambda^{d_{\mathcal{G}'}(u_0,e)-1}\|\pi_e\|+\sum_{v\in\mathcal{V}'\backslash S}\lambda^{d_{\mathcal{G}'}(u_0,v)}\|\pi_v\|\right).\end{aligned}$$

□

Now we return to the proof of Theorem 3.3.9. For simplicity, we temporarily assume the individual decision points are unconstrained, i.e., $D_t^v = \mathbb{R}^n$. We discuss how to relax this assumption later in this section.

We first consider the case when $\left(\{y_{t-1}^u\}, \{z_\tau^u\}; \xi_{(t,v)}^{(k,r)}\right)$ and $\left(\{y_{t-1}^u\}, \{(z_\tau^u)'\}; (\xi_{(t,v)}^{(k,r)})'\right)$ only differ at one entry $z_\tau^u, \omega_\tau^u, \alpha_\tau^u$, or β_τ^e. If the difference is at z_τ^u, by viewing each subset $\{\tau\} \times N_v^r$ for $\tau \in \{t-1, t, \ldots, t+k\}$ in the original problem as a vertex in the new graph $\mathcal{G}'$ and applying Theorem 3.D.2, we obtain that

$$\left\|x_t^v - (x_t^v)'\right\| \leq C_1^0 \cdot (\rho_T^0)^{|t-\tau|} \left\|z_\tau^u - (z_\tau^u)'\right\|, \tag{3.50}$$

where $C_1^0 = (2\ell_T)/\mu$ and $\rho_T^0 = 1 - 2 \cdot \left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1}$. On the other hand, by viewing each subset $\{\tau \mid t-1 \leq \tau < t+k\} \times \{u\}$ for $u \in N_v^r$ in the original problem as a vertex in the new graph $\mathcal{G}'$ and applying Theorem 3.D.2, we obtain that

$$\left\|x_t^v - (x_t^v)'\right\| \leq C_1^1 \cdot (\rho_S^0)^{d_{\mathcal{G}}(u,v)} \left\|z_\tau^u - (z_\tau^u)'\right\|, \tag{3.51}$$

where $C_1^1 = (2\Delta\ell_S)/\mu$ and $\rho_S^0 = 1 - 2 \cdot \left(\sqrt{1 + (2\Delta\ell_S/\mu)}\right)^{-1}$. Combining (3.50) and (3.51) gives that

$$\begin{aligned}
\left\|x_t^v - (x_t^v)'\right\| \leq{}& \min\{C_1^0 \cdot (\rho_T^0)^{|t-\tau|}, C_1^1 \cdot (\rho_S^0)^{d_{\mathcal{G}}(u,v)}\} \cdot \left\|z_\tau^u - (z_\tau^u)'\right\| \\
\leq{}& \sqrt{C_1^0 \cdot C_1^1} \cdot (\rho_T^0)^{|t-\tau|/2} \cdot (\rho_S^0)^{d_{\mathcal{G}}(u,v)/2} \cdot \left\|z_\tau^u - (z_\tau^u)'\right\| \\
\leq{}& C_1 \cdot \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,u)} \left\|z_\tau^u - (z_\tau^u)'\right\|
\end{aligned} \tag{3.52}$$

when $\left(\{y_{t-1}^u\}, \{z_\tau^u\}; \xi_{(t,v)}^{(k,r)}\right)$ and $\left(\{y_{t-1}^u\}, \{(z_\tau^u)'\}; (\xi_{(t,v)}^{(k,r)})'\right)$ only differ at one entry z_τ^u for $(\tau, u) \in \partial N_{(t,v)}^{(k,r)}$. We can use the same method to show that when $\left(\{y_{t-1}^u\}, \{z_\tau^u\}; \xi_{(t,v)}^{(k,r)}\right)$ and $\left(\{y_{t-1}^u\}, \{(z_\tau^u)'\}; (\xi_{(t,v)}^{(k,r)})'\right)$ only differ at another entry at $\omega_\tau^u, \alpha_\tau^u$, or β_τ^e, we have

$$\begin{aligned}
&\left\|x_t^v - (x_t^v)'\right\| \leq C_3 \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,u)} \left\|\omega_\tau^u - (\omega_\tau^u)'\right\|, \text{ if they differ at } \mu_\tau^u, \\
&\left\|x_t^v - (x_t^v)'\right\| \leq C_3 \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v_0,u)} \left\|\alpha_\tau^u - (\alpha_\tau^u)'\right\|, \text{ if they differ at } \alpha_\tau^u, \\
&\left\|x_t^v - (x_t^v)'\right\| \leq C_3 \rho_T^{|t-\tau|} \rho_S^{d_{\mathcal{G}}(v,e)} \left\|\beta_\tau^e - (\beta_\tau^e)'\right\|, \text{ if they differ at } \beta_\tau^e.
\end{aligned} \tag{3.53}$$

In the general case where $\left(\{y_{t-1}^u\}, \{z_\tau^u\}; \xi_{(t,v)}^{(k,r)}\right)$ and $\left(\{(y_{t-1}^u)'\}, \{(z_\tau^u)'\}; (\xi_{(t,v)}^{(k,r)})'\right)$ differ not only at one entry, we can perturb the entries of parameters one at a time and apply the triangle inequality. Then, the conclusion of Theorem 3.3.9 follows from (3.52) and (3.53). One can use the same approach to show Theorem 3.3.3.

Proof of Theorem 3.3.4

Step 1. Establishing first order equations

Given any parameter $\zeta = \{z_\tau^u | (\tau, u) \in \partial N_{(t,v)}^{(k,r)}\}$, the objective function $\hat{h}$ can be written as:

$$\begin{aligned}\hat{h}(\hat{x}_{1:k-1}, \zeta) = & \sum_{i=1}^{k-1} \sum_{u \in N_v^{r-1}} f_{t-1+i}^u(\hat{x}_i^u) + \sum_{i=1}^{k-1} \sum_{(u,u') \in \mathcal{E}(N_v^r)} s_{t-1+i}^{(u,u')}(\hat{x}_i^u, \hat{x}_i^{u'}) \\ & + \sum_{i=1}^{k} \sum_{u \in N_v^r} c_{t-1+i}^u(\hat{x}_i^u, \hat{x}_{i-1}^u), \text{ where} \\ & \hat{x}_0^u = y_{t-1}^u, \ \forall u \in N_v^r; \ \hat{x}_k^u = \hat{z}_k^u, \ \forall u \in N_v^r; \\ & \hat{x}_i^u = \hat{z}_i^u, \ \forall i \in \{1, \ldots, k-1\}, \ u \in \partial N_v^r.\end{aligned}$$

Given $\theta \in \mathbb{R}$, $\psi(\zeta + \theta e)$ is the global minimizer of convex function $\hat{h}(\cdot, \zeta + \theta e)$, and hence we have

$$\nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e) = 0.$$

Taking the derivative with respect to θ, we establish the following set of equations:

$$\begin{aligned} & \nabla_{\hat{x}_{1:k-1}}^2 \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e) \frac{d}{d\theta} \psi(\zeta + \theta e) \\ = & -\nabla_{\hat{z}_k} \nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e) e_k \\ & - \sum_{\tau=1}^{k-1} \nabla_{\hat{z}_\tau} \nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e) e_\tau. \end{aligned} \tag{3.54}$$

We adopt the following short-hand notation:

- $M := \nabla_{\hat{x}_{1:k-1}}^2 \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e)$, which is a hierarchical block matrix with the first level of dimension $(k-1) \times (k-1)$, the second level of dimension $|N_v^{r-1}| \times |N_v^{r-1}|$ and the third level of dimension $n \times n$.
- $R^{(k)} := -\nabla_{\hat{z}_k} \nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e)$, which is also a hierarchical block matrix with the first level of dimension $(k-1) \times 1$, the second level of dimension $|N_v^{r-1}| \times |N_v^r|$ and the third level of dimension $n \times n$.
- For $\tau = 1, \ldots, k-1$, $K^{(\tau)} := -\nabla_{\hat{z}_\tau} \nabla_{\hat{x}_{1:k-1}} \hat{h}(\psi(\zeta + \theta e), \zeta + \theta e)$, which is also a hierarchical block matrix with the first level of dimension $(k-1) \times 1$. the second level of dimension $|N_v^{r-1}| \times |\partial N_v^r|$ and the third level of dimension $n \times n$.

Using the above, we can rewrite (3.54) as follows:

$$\frac{d}{d\theta}\psi(\zeta + \theta e) = M^{-1}\left(R^{(k)}e_k + \sum_{\tau=1}^{k-1} K^{(\tau)}e_\tau\right).$$

Due to the structure of temporal interaction cost functions, for $R^{(k)}$, only when the first level index is $k-1$, the lower level block matrix is non-zero; due to the structure of spatial interaction cost functions, for $K^{(\tau)}$, only when the first level index is τ, the lower level block matrix is non-zero. Hence, for $1 \le \tau' \le k-1$, we have

$$\left(\frac{d}{d\theta}\psi(\zeta + \theta e)\right)_{\tau'} = (M^{-1})_{\tau',k-1}R^{(k)}_{k-1}e_k + \sum_{\tau=1}^{k-1}(M^{-1})_{\tau',\tau}K^{(\tau)}_{\tau}e_\tau, \tag{3.55}$$

where the subscripts on the right hand side denote the first level index of hierarchical block matrices M, $R^{(1)}$, $R^{(k-1)}$ and $K^{(\tau)}$.

Step 2. Decomposing M^{-1} as infinite series

We decompose M to block diagonal matrix D and tri-diagonal block matrix A such that $M = D + A$. We denote each diagonal block in D as $D_{i,i}$ for $1 \le i \le k-1$. Other blocks in D are zero matrices.

$$D \coloneqq \begin{bmatrix}
\ast & 0 & \cdots & \ast & & & & & & & & & \\
0 & \ast & & 0 & & & & & & & & & \\
\vdots & & \ddots & & & & & & & & & & \\
\ast & 0 & \cdots & \ast & & & & & & & & & \\
& & & & \ast & 0 & \cdots & \ast & & & & & \\
& & & & 0 & \ast & & 0 & & & & & \\
& & & & \vdots & & \ddots & & & & & & \\
& & & & \ast & 0 & \cdots & \ast & & & & & \\
& & & & & & & & \ddots & & & & \\
& & & & & & & & & \ast & 0 & \cdots & \ast \\
& & & & & & & & & 0 & \ast & & 0 \\
& & & & & & & & & \vdots & & \ddots & \\
& & & & & & & & & \ast & 0 & \cdots & \ast
\end{bmatrix}$$

Each non-zero block in A is a diagonal block matrix, which captures the Hessian of temporal interaction cost between consecutive time steps. Denote each block as $A_{i,j}$ for $1 \le i, j \le k-1$. The structure of matrix A is given below, where the diagonal blocks are all zeros.

$$A := \begin{bmatrix} & \begin{matrix} * & 0 & \cdots & 0 \\ 0 & * & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & * \end{matrix} & & \\ \begin{matrix} * & 0 & \cdots & 0 \\ 0 & * & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & * \end{matrix} & & \begin{matrix} * & 0 & \cdots & 0 \\ 0 & * & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & * \end{matrix} & \\ & & \ddots & \begin{matrix} * & 0 & \cdots & 0 \\ 0 & * & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & * \end{matrix} \\ & & \begin{matrix} * & 0 & \cdots & 0 \\ 0 & * & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & * \end{matrix} & \end{bmatrix}$$

We rewrite the inverse of M as follows:

$$M^{-1} = (D + A)^{-1} = D^{-1}(I + AD^{-1})^{-1} = D^{-1}P^{-1}.$$

Now we present the proof of Lemma 3.C.2.

Proof of Lemma 3.C.2. Recall that $P = I + AD^{-1} = (D + A)D^{-1}$. We claim that all eigenvalues of P are contained in the set $\{\lambda \in \mathbb{C} | |\lambda - z| \leq R\}$ for some $R \in \mathbb{R}_{>0}$ and $z \in \mathbb{C} \setminus \{0\}$ such that $R < |z|$. We first establish Lemma 3.C.2 based on the claim and then prove the claim.

We follow the argument as in the proof of Thm 4 in Shin, Zavala, and Anitescu (2020). Since any eigenvalue λ of P satisfies $|\lambda - z| \leq R$, $|\lambda/z - 1| \leq R/|z| < 1$. Thus, the eigenvalues of $I - (1/z)P$ lie on $\{\tilde{\lambda} \in \mathbb{C} : |\tilde{\lambda}| \leq R/|z|\}$, which guarantees $\rho(I - (1/z)P) < 1$. Therefore,

$$P^{-1} = \frac{1}{z}\left(I - (I - \frac{1}{z}P)\right)^{-1} = \frac{1}{z}\sum_{q \geq 0}(I - \frac{1}{z}P)^q.$$

Now we show the claim that all eigenvalues of P are contained in the set $\{\lambda \in \mathbb{C} | |\lambda - z| \leq R\}$ holds if we let $z = 1$ and $R = \frac{2\ell_T}{\mu}$. Our proof utilizes Gershgorin circle theorem for block matrices and its implications (see Theorem 1.13.1 and Remark 1.13.2 of Tretter (2008)), which we present in Theorem 3.D.3.

Theorem 3.D.3. *Consider $\mathcal{K} = (K_{ij}) \in \mathbb{R}^{dm\times dm}$ $(d, m \geq 1)$ where $K_{ij} \in \mathbb{R}^{d\times d}$ for $i, j \in \{1, \ldots, m\}$ and K_{ii} is symmetric for $i \in \{1, \ldots, m\}$. Let $\sigma(\cdot)$ denotes the spectrum of a matrix. Define set*

$$G_i \coloneqq \sigma(K_{ii}) \cup \left\{ \cup_{q=1}^{d} B\left(\lambda_q(K_{ii}), \sum_{j\neq i} \left\|K_{ij}\right\|\right)\right\}$$

where $B(\cdot, \cdot)$ denotes a disk $B(c, r) = \{\lambda : \|\lambda - c\| \leq r\}$ and λ_q is the q-th smallest eigenvalues of K_{ii}. Then,

$$\sigma(\mathcal{K}) \in \cup_{i=1}^{n} G_i.$$

Next, we use the above fact to find a superset of $\sigma(P)$. Every diagonal block of P is I. Moreover, $P_{i,j} = 0$ for $|i - j| > 1$, $P_{i,i-1} = A_{i,i-1}D^{-1}_{i-1,i-1}$, $P_{i,i+1} = A_{i,i+1}D^{-1}_{i+1,i+1}$. Hence we have

$$\sum_{j\neq i} \left\|P_{i,j}\right\| \leq \left\|A_{i,i-1}\right\|\left\|D^{-1}_{i-1,i-1}\right\| + \left\|A_{i,i+1}\right\|\left\|D^{-1}_{i+1,i+1}\right\| \leq \frac{2\ell_T}{\mu}.$$

The last inequality holds because the problem instance $p \in \mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$. Therefore, $G_i = B(1, \frac{2\ell_T}{\mu})$. This implies all eigenvalues of P are in $B(1, \frac{2\ell_T}{\mu})$. □

To further simplify the notation in the power series expansion, we define $J \coloneqq AD^{-1} = P - I$. Given any period indices τ' and τ, we have

$$(M^{-1})_{\tau',\tau} = (D^{-1})_{\tau',\tau'}(P^{-1})_{\tau',\tau} = (D^{-1})_{\tau',\tau'} \times \sum_{\ell\geq 0}(-J)^{\ell}_{\tau',\tau}, \tag{3.56}$$

where the first equality is since D^{-1} is a diagonal block matrix, the second equality is due to Lemma 3.C.2.

Step 3: Showing exponential-decay properties are preserved through matrix multiplications

This step simply requires proving Lemma 3.C.3.

Proof of Lemma 3.C.3. Under the assumptions, we see that

$$\begin{aligned}
&\sum_q (\frac{1}{\lambda'})^{d_{\mathcal{M}}(u,q)} \|(A_1 A_2 \cdots A_\ell)_{u,q}\| \\
&= \sum_q (\frac{1}{\lambda'})^{d_{\mathcal{M}}(u,q)} \left\| \sum_{s_1,\cdots,s_{\ell-1}} (A_1)_{u,s_1}(A_2)_{s_1,s_2}\cdots(A_\ell)_{s_{\ell-1},q} \right\| \\
&\le \sum_q (\frac{1}{\lambda'})^{d_{\mathcal{M}}(u,q)} \sum_{s_1,\cdots,s_{\ell-1}} (C_1\lambda^{d_{\mathcal{M}}(u,s_1)})(C_2\lambda^{d_{\mathcal{M}}(s_1,s_2)})\cdots(C_\ell\lambda^{d_{\mathcal{M}}(s_{\ell-1},q)}) \qquad (3.57)\\
&\le \sum_q \sum_{s_1,\cdots,s_{\ell-1}} \prod_{i=1}^{\ell} C_i (\frac{\lambda}{\lambda'})^{d_{\mathcal{M}}(u,s_1)+d_{\mathcal{M}}(s_1,s_2)+\cdots+d_{\mathcal{M}}(s_{\ell-1},q)} \\
&\le (\tilde{a})^\ell \prod_{i=1}^{\ell} C_i.
\end{aligned}$$

Hence, we obtain that

$$\left\| (\prod_{i=1}^{\ell} A_i)_{u,q} \right\| \le C'(\lambda')^{d_M(u,q)}.$$

□

Step 4: Establishing exponential decay properties of matrix M^{-1}

In this step, we use the property developed for general exponential-decay matrices on M and derive the perturbation bound in the Theorem 3.3.4.

Lemma 3.D.4. *For $\ell \ge 1$, period index $i, j \ge 1$, J^ℓ has the following properties:*

- $(J^\ell)_{i,j} = 0$ *if* $\ell < |i-j|$ *or* $\ell - |i-j|$ *is odd.*
- $(J^\ell)_{i,j}$ *is a summation of terms* $\prod_{k=1}^{\ell} A_{j_k,i_k} D^{-1}_{i_k,i_k}$ *and the number of such terms is bounded by* $\binom{\ell}{(\ell-|i-j|)/2}$.

Note for integers $m, k \ge 1$, we define $\binom{m}{k/2} = 0$ if k is odd.

Proof of 3.D.4. Since J is a tri-diagonal banded matrix, $J^\ell_{i,j} = 0$ for $\ell < |i-j|$. We prove the rest of properties of J by induction on ℓ.

When $\ell = 1$,

$$J_{i,i} = 0, \quad J_{i,i-1} = A_{i,i-1}D^{-1}_{i-1,i-1}, \quad J_{i,i+1} = A_{i,i+1}D^{-1}_{i+1,i+1}.$$

Lemma 3.D.4 holds for the base case. Suppose Lemma 3.D.4 holds for J^q for $q \leq \ell - 1$. Let $q = \ell$, then

$$J^{\ell}_{i,j} = \sum_k J^{\ell-1}_{i,k} J_{k,j} = J^{\ell-1}_{i,j-1} A_{j-1,j} D^{-1}_{j,j} + J^{\ell-1}_{i,j+1} A_{j+1,j} D^{-1}_{j,j}.$$

By induction hypothesis, $J^{\ell-1}_{i,j}$ is a summation of terms $\prod_{k=1}^{\ell-1} A_{j_k,i_k} D^{-1}_{i_k,i_k}$. Moreover, the number of such terms is bounded by $\binom{\ell-1}{(\ell-1-|i-j-1|)/2} + \binom{\ell-1}{(\ell-1-|i-j+1|)/2}$. Next we will show $\binom{\ell-1}{(\ell-1-|i-j-1|)/2} + \binom{\ell-1}{(\ell-1-|i-j+1|)/2} = \binom{\ell}{(\ell-|i-j|)/2}$ case by case.

Case 1: $\ell - |i-j|$ is odd.

If $\ell - |i-j|$ is odd, then $\ell - 1 - |i-j-1|$ and $\ell - 1 - |i-j+1|$ are both odd. Under this case,

$$\binom{\ell-1}{(\ell-1-|i-j-1|)/2} + \binom{\ell-1}{(\ell-1-|i-j+1|)/2} = 0,$$

which is equal to $\binom{\ell}{(\ell-|i-j|)/2}$.

Case 2: $\ell - |i-j|$ is even and $i = j$. Under this case, we have

$$\binom{\ell-1}{(\ell-1-|i-j-1|)/2} + \binom{\ell-1}{(\ell-1-|i-j+1|)/2} = \binom{\ell-1}{\ell/2-1} + \binom{\ell-1}{\ell/2-1}.$$

Since ℓ is even, $\binom{\ell-1}{\ell/2-1} + \binom{\ell-1}{\ell/2-1} = \binom{\ell}{\ell/2} = \binom{\ell}{(\ell-|i-j|)/2}$.

Case 3: $\ell - |i-j|$ is even and $i \neq j$.

If $\ell - |i-j|$ is even, then $\ell - 1 - |i-j-1|$ and $\ell - 1 - |i-j+1|$ are both even. We denote $(\ell - |i-j|)/2$ as k_0. By triangle inequality, $(\ell - 1 - |i-j-1|)/2$ and $(\ell - 1 - |i-j+1|)/2$ are in $\{k_0 - 1, k_0\}$. Since $i \neq j$,

$$\binom{\ell-1}{(\ell-1-|i-j-1|)/2} + \binom{\ell-1}{(\ell-1-|i-j+1|)/2} = \binom{\ell-1}{k_0-1} + \binom{\ell-1}{k_0},$$

which sums to $\binom{\ell}{k_0}$ by Pascal's triangle. □

Next we present the proof of Lemma 3.C.4.

Proof of Lemma 3.C.4. By Lemma 3.D.4, $(J^{\ell})_{i,j}$ equals to the summation of terms $\prod_{k=1}^{\ell} A_{j_k,i_k} D^{-1}_{i_k,i_k}$ and the number of such terms is bounded by $\binom{\ell}{(\ell-|i-j|)/2}$.

Define $B_k \coloneqq A_{j_k,i_k} D^{-1}_{i_k,i_k}$. Recall A_{j_k,i_k} is a diagonal matrix and D_{i_k,i_k} is a graph-induced banded matrix.

$$\begin{aligned}\left\|(B_k)_{u,q}\right\| = \left\|(A_{j_k,i_k} D^{-1}_{i_k,i_k})_{u,q}\right\| = \left\|(A_{j_k,i_k})_{u,u}(D^{-1}_{i_k,i_k})_{u,q}\right\| &\le \ell_T \left\|(D^{-1}_{i_k,i_k})_{u,v}\right\| \\ &\le \frac{2\ell_T}{\mu}\gamma_S^{d_{\mathcal{G}}(u,v)}.\end{aligned}$$

where the last inequality is by using Lemma 3.D.1 on D_{i_k,i_k}.

Under the condition $b < \infty$, we can use Lemma 3.C.3 to obtain the following bound,

$$\left\|(\prod_{k=1}^{\ell} A_{j_k,i_k} D^{-1}_{i_k})_{u,v}\right\| \le (b\frac{2\ell_T}{\mu})^{\ell}(\gamma'_S)^{d_{\mathcal{G}}(u,v)}.$$

Since the number of such terms is bounded by $\binom{\ell}{(\ell-|i-j|)/2}$, we have

$$\left\|((J^{\ell})_{i,j})_{u,q}\right\| \le \binom{\ell}{(\ell-|i-j|)/2}(b\frac{2\ell_T}{\mu})^{\ell}(\gamma'_S)^{d_{\mathcal{G}}(u,v)}.$$

□

Proof of 3.C.5. Given $1 \le \tau, \tau' \le k-1$ and $v_0 \in N_v^{r-1}$, since $M^{-1} = D^{-1}\sum_{\ell\ge 0}(-J)^{\ell}$, we have

$$\left\|\left((M)^{-1}_{\tau',\tau}K^{(\tau)}_{\tau}y\right)_{v_0}\right\| = \left\|\left(D^{-1}_{\tau',\tau'}\sum_{\ell\ge 0}(-J)^{\ell}_{\tau',\tau}K^{(\tau)}_{\tau}y\right)_{v_0}\right\|. \tag{3.58}$$

With slight abuse of notation, we use K to denote $K^{(\tau)}_{\tau}$, and Q^{-1} to denote $D^{-1}_{\tau',\tau'}$ in this proof from now. We can rewrite the right hand side of (3.58) using the new notation as follows:

$$\begin{aligned}\left\|\left(Q^{-1}\sum_{\ell\ge 0}(-J)^{\ell}_{\tau',\tau}Ky\right)_{v_0}\right\| &\le \sum_{\ell\ge 0}\left\|\left(Q^{-1}(-J)^{\ell}_{\tau',\tau}Ky\right)_{v_0}\right\| \\ &= \sum_{\ell\ge 0}\left\|\sum_{q\in N_v^{r-1}}\left(Q^{-1}(-J)^{\ell}_{\tau',\tau}\right)_{v_0,q}(Ky)_q\right\| \\ &\le \sum_{\ell\ge 0}\sum_{q\in N_v^{r-1}}\left\|\left(Q^{-1}(-J)^{\ell}_{\tau',\tau}\right)_{v_0,q}\right\|\left\|(Ky)_q\right\|.\end{aligned} \tag{3.59}$$

For a given $q \in N_v^{r-1}$ and $y \in \mathbb{R}^{|\partial N_v^r|d}$,

$$\left\|(Ky)_q\right\| = \left\|\sum_{u\in\partial N_v^r}K_{q,u}y_u\right\| = \left\|\sum_{u\in\partial N_v^r\cap N_q^1}K_{q,u}y_u\right\|.$$

where the last equality is since spacial interaction costs are only among neighboring nodes.

For a given $u \in \partial N_v^r$, since the spacial interaction cost for each edge is ℓ_S smooth,

$$\left\|K_{q,u} y_u\right\| \leq \left\|K_{q,u}\right\| \|y_u\| \leq \ell_S \|y_u\|,$$

which gives

$$\left\|(Ky)_q\right\| \leq \sum_{u \in \partial N_v^r \cap N_q^1} \ell_S \|y_u\|.$$

Therefore,

$$\left\|\left(Q^{-1} \sum_{\ell \geq 0} (-J)_{\tau',\tau}^{\ell} K y\right)_{v_0}\right\| \leq \ell_S \sum_{\ell \geq 0} \sum_{q \in N_v^{r-1}} \left\|\left(Q^{-1} (-J)_{\tau',\tau}^{\ell}\right)_{v_0,q}\right\| \sum_{u \in \partial N_v^r \cap N_q^1} \|y_u\|. \tag{3.60}$$

By Lemma 3.C.4, $(-J)_{\tau',\tau}^{\ell}$ satisfies the following exponential decay properties: for any $u, q \in N_v^{r-1}$,

$$\left\|((J^{\ell})_{\tau',\tau})_{u,q}\right\| \leq \binom{\ell}{(\ell - |\tau' - \tau|)/2} (\tilde{a} \frac{2\ell_T}{\mu})^{\ell} (\gamma_S')^{d_{\mathcal{G}}(u,q)},$$

where we choose $\delta = b_1 \cdot \gamma_S$, $\gamma_S' = (1 + b_1)\gamma_S$ and $\tilde{a} = \sum_{\gamma \geq 0} (\frac{1}{1+b_1})^{\gamma} h(\gamma)$.

Moreover, Q^{-1} (which denotes $D_{\tau',\tau'}^{-1}$) is the inverse of a graph-induced banded matrix. Q^{-1} satisfies: for any $u, q \in N_v^{r-1}$,

$$\left\|(Q^{-1})_{u,q}\right\| \leq \frac{2}{\mu} \gamma_S^{d_{\mathcal{G}}(u,q)} < \frac{2}{\mu} (\gamma_S')^{d_{\mathcal{G}}(u,q)},$$

where the first inequality is again by using Lemma 3.D.1 on $D_{\tau',\tau'}$.

Applying Lemma 3.C.3 on Q^{-1} and $\left\|((J^{\ell})_{\tau',\tau})\right\|$, we have for any $u, q \in N_v^{r-1}$, and $\ell \geq 1$,

$$\left\|\left(Q^{-1} (-J)_{\tau',\tau}^{\ell}\right)_{u,q}\right\| \leq a^2 \frac{2}{\mu} \binom{\ell}{(\ell - |\tau' - \tau|)/2} (\tilde{a} \frac{2\ell_T}{\mu})^{\ell} (\lambda')^{d_{\mathcal{G}}(u,q)},$$

where $\lambda' := \gamma_S' + b_2 \cdot \gamma_S < 1$ and $a := \sum_{\gamma \geq 0} (\frac{1+b_1}{1+b_1+b_2})^{\gamma} h(\gamma)$. Note that $J^0 := I$, it is straightforward to verify that the above inequality holds when $\ell = 0$.

With the exponential decay properties of $Q^{-1}(-J)^{\ell}_{\tau',\tau}$, we have

$$
\begin{aligned}
&\left\|\left(Q^{-1}\sum_{\ell\geq 0}(-J)^{\ell}_{\tau',\tau}Ky\right)_{v_0}\right\| \\
&\leq \ell_S a^2\frac{2}{\mu}\sum_{\ell\geq 0}\binom{\ell}{(\ell-|\tau'-\tau|)/2}(\tilde{a}\frac{2\ell_T}{\mu})^{\ell}\sum_{q\in N_v^{r-1}}(\lambda')^{d_{\mathcal{G}}(v_0,q)}\sum_{u\in\partial N_v^r\cup N_q^1}\|y_u\| \\
&\leq \ell_S a^2\frac{2}{\mu}\sum_{\ell\geq|\tau'-\tau|}\binom{\ell}{(\ell-|\tau'-\tau|)/2}(\tilde{a}\frac{2\ell_T}{\mu})^{\ell}\sum_{u\in\partial N_v^r}\Delta(\lambda')^{d_{\mathcal{G}}(v_0,u)-1}\|y_u\| \\
&\leq \Delta\ell_S a^2\frac{2}{\mu}\sum_{\ell\geq|\tau'-\tau|}(\frac{4\tilde{a}\ell_T}{\mu})^{\ell}\sum_{u\in\partial N_v^r}(\lambda')^{d_{\mathcal{G}}(v_0,u)-1}\|y_u\| \\
&\leq \frac{2\Delta\ell_S a^2}{\mu-4\tilde{a}\ell_T}(\frac{4\tilde{a}\ell_T}{\mu})^{|\tau'-\tau|}\sum_{u\in\partial N_v^r}(\lambda')^{d_{\mathcal{G}}(v_0,u)-1}\|y_u\| \\
&= \frac{2\Delta\ell_S a^2}{\lambda'(\mu-4\tilde{a}\ell_T)}(\frac{4\tilde{a}\ell_T}{\mu})^{|\tau'-\tau|}\sum_{u\in\partial N_v^r}(\lambda')^{d_{\mathcal{G}}(v_0,u)}\|y_u\|.
\end{aligned}
\tag{3.61}
$$

The third inequality uses $\binom{\ell}{(\ell-|\tau'-\tau|)/2}\leq 2^{\ell}$, which can be proved using the following version of Stirling's approximation: For all $n\geq 1$, e denotes the natural number,

$$
\sqrt{2\pi n}(n/e)^n e^{1/(12n+1)} < n! < \sqrt{2\pi n}(n/e)^n e^{1/(12n)}.
$$

Similarly, consider $\left\|((M^{-1})_{\tau',i})R_i^{(i)}e)_{v_0}\right\|$ for $i\in\{1,k-1\}$. With slight abuse of notation, in this proof, we use R to denote $R_i^{(i)}$ and use the notation Q^{-1} to denote $D^{-1}_{\tau',\tau'}$. Following the same steps as before, we have

$$
\left\|\left((M^{-1})_{\tau',i})R_i^{(i)}e\right)_{v_0}\right\| \leq \sum_{\ell\geq 0}\sum_{q\in N_v^r}\left\|\left(Q^{-1}(-J)^{\ell}_{\tau',i}\right)_{v_0,q}\right\|\left\|(Re)_q\right\|. \tag{3.62}
$$

Since temporal interactions occurs for the same node under consecutive time steps, R is a diagonal block matrix. Hence,

$$
\left\|(Re)_q\right\| = \left\|R_{q,q}e_q\right\| \leq \ell_T\left\|e_q\right\|.
$$

Moreover, using the exponential decay properties of $Q^{-1}(-J)^{\ell}_{\tau',i}$, we have for $u,q\in N_v^{r-1}$,

$$
\left\|\left(Q^{-1}(-J)^{\ell}_{\tau',i}\right)_{u,q}\right\| \leq a^2\frac{2}{\mu}\binom{\ell}{(\ell-|\tau'-i|)/2}(\tilde{a}\frac{2\ell_T}{\mu})^{\ell}(\lambda')^{d_{\mathcal{G}}(u,q)}.
$$

Therefore,

$$\begin{aligned}
&\left\|((M^{-1})_{\tau',i})R_i^{(i)}e)_{v_0}\right\| \\
&\le \sum_{\ell\ge 0}\sum_{q\in N_v^r} a^2\frac{2}{\mu}\binom{\ell}{(\ell-|\tau'-i|)/2}(\tilde a\frac{2\ell_T}{\mu})^{\ell}(\lambda')^{d_{\mathcal G}(v_0,q)}\ell_T\left\|e_q\right\| \\
&\le \sum_{\ell\ge |\tau'-i|}\sum_{q\in N_v^r} a^2\frac{2}{\mu}\binom{\ell}{(\ell-|\tau'-i|)/2}(\tilde a\frac{2\ell_T}{\mu})^{\ell}(\lambda')^{d_{\mathcal G}(v_0,q)}\ell_T\left\|e_q\right\| \\
&\le \frac{2\ell_T a^2}{\mu}\sum_{\ell\ge|\tau'-i|}(\frac{4\tilde a\ell_T}{\mu})^{\ell}\sum_{q\in N_v^r}(\lambda')^{d_{\mathcal G}(v_0,q)}\left\|e_q\right\| \\
&\le \frac{2\ell_T a^2}{\mu-4\tilde a\ell_T}(\frac{4\tilde a\ell_T}{\mu})^{|\tau'-i|}\sum_{q\in N_v^r}(\lambda')^{d_{\mathcal G}(v_0,q)}\left\|e_q\right\| \\
&= \frac{a^2\mu}{2\tilde a(\mu-4\tilde a\ell_T)}(\frac{4\tilde a\ell_T}{\mu})^{|\tau'-i|+1}\sum_{q\in N_v^r}(\lambda')^{d_{\mathcal G}(v_0,q)}\left\|e_q\right\|.
\end{aligned} \tag{3.63}$$

□

Given time index $1 \le \tau' \le k-1$, node $v_0 \in N_v^{r-1}$, and perturbation vector $e = (e_0, e_1, \cdots, e_k)$,

$$\begin{aligned}
&\left\|(\frac{d}{d\theta}\psi(\zeta+\theta e))_{\tau',v_0}\right\| \\
&\le \left\|\left(M^{-1}_{\tau',1}R_1^{(1)}e_0\right)_{v_0}\right\| + \left\|\left(M^{-1}_{\tau',k-1}R_{k-1}^{(k-1)}e_k\right)_{v_0}\right\| + \sum_{\tau=1}^{k-1}\left\|\left(M^{-1}_{\tau',\tau}K_\tau^{(\tau)}e_\tau\right)_{v_0}\right\| \\
&\le \frac{a^2\mu}{2\tilde a(\mu-4\tilde a\ell_T)}\left[\rho_T^{\tau'}\sum_{q\in N_v^r}\rho_S^{d_{\mathcal G}(v_0,q)}\left\|(e_0)_q\right\| + \rho_T^{k-\tau'}\sum_{q\in N_v^r}\rho_S^{d_{\mathcal G}(v_0,q)}\left\|(e_k)_q\right\|\right] \\
&+ \sum_{\tau=1}^{k-1}\frac{2\Delta\ell_S a^2}{\lambda'(\mu-4\tilde a\ell_T)}\rho_T^{|\tau'-\tau|}\sum_{u\in\partial N_v^r}(\rho_S)^{d_{\mathcal G}(v_0,u)}\|(e_\tau)_u\|
\end{aligned}$$

where $\rho_T = \frac{4\tilde a\ell_T}{\mu}$ and $\rho_S = \lambda' = (1+b_1+b_2)\gamma_S$. We let

$$C = \max\{\frac{a^2}{2\tilde a(1-4\tilde a\ell_T/\mu)}, \frac{2a^2\Delta\ell_S/\mu}{\gamma_S(1+b_1+b_2)(1-4\tilde a\ell_T/\mu)}\}.$$

Under the condition $\mu \ge \max\{8\tilde a\ell_T, \Delta\ell_S(b_1+b_2)/4\}$, $\rho_T < 1$ and $\rho_S < 1$.

Then,

$$\left\|(\frac{d}{d\theta}\psi(\zeta+\theta e))_{\tau',v_0}\right\|$$

$$\leq C\Bigg[\rho_T^{\tau'} \sum_{q \in N_v^r} \rho_S^{d_{\mathcal{G}}(v_0,q)} \left\|(e_0)_q\right\| + \rho_T^{k-\tau'} \sum_{q \in N_v^r} \rho_S^{d_{\mathcal{G}}(v_0,q)} \left\|(e_k)_q\right\|$$

$$+ \sum_{\tau=1}^{k-1} \rho_T^{|\tau'-\tau|} \sum_{u \in \partial N_v^r} (\rho_S)^{d_{\mathcal{G}}(v_0,u)} \|(e_\tau)_u\|\Bigg].$$

Finally, let $\zeta = \{y_{t-1}^u, z_\tau^u | (\tau, u) \in \partial N_{(v,t)}^{(k,r)}\}$ and $e = \{(y_{t-1}^u)' - y_{t-1}^u, (z_\tau^u)' - z_\tau^u\}$. By integration,

$$\left\|\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\}, \{z_\tau^u\}\right)_{(t_0,v_0)} - (\psi_{p,(t,v)}^{(k,r)}\left(\{(y_{t-1}^u)'\}, \{(z_\tau^u)'\}\right)_{(t_0,v_0)}\right\|$$

$$\leq \int_0^1 \left\|(\frac{d}{d\theta}\psi(\zeta + \theta e))_{t_0,v_0}\right\| d\theta,$$

which is bounded by

$$C \sum_{u \in N_v^r} \rho_T^{t_0-(t-1)} \rho_S^{d_{\mathcal{G}}(v_0,u)} \left\|y_{t-1}^u - (y_{t-1}^u)'\right\| + C \sum_{(u,\tau) \in \partial N_{(t,v)}^{(k,r)}} \rho_T^{|t_0-\tau|} \rho_S^{d_{\mathcal{G}}(v_0,u)} \left\|z_\tau^u - (z_\tau^u)'\right\|.$$

Adding Constraints to Perturbation Bounds

Recall that we have shown Theorem 3.3.3 and Theorem 3.3.4 under the assumption that the individual decisions are unconstrained to simplify the analysis. In this section, we present a general way to relax this assumption by incorporating logarithm barrier functions, which also applies for Theorem 3.C.6.

Recall that in Assumption 3.3.1, we assume that D_t^v is convex with a non-empty interior, and can be expressed as

$$D_t^v := \{x_t^v \in \mathbb{R}^n \mid (g_t^v)_i(x_t^v) \leq 0, \forall 1 \leq i \leq m_t^v\},$$

where the i th constraint $(g_t^v)_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function in C^2. For any time-vertex pair (τ, v), we can approximate the individual constraints

$$(g_\tau^v)_i(x_\tau^v) \leq 0, \forall 1 \leq i \leq m_\tau^v,$$

by adding the *logarithmic barrier function* $-\lambda \sum_{i=1}^{m_\tau^v} \ln\left(-(g_\tau^v)_i(x_\tau^v)\right)$ to the original node cost function f_τ^v. Here, parameter μ is a positive real number that controls how "good" the barrier function approximates the indicator function

$$\mathbf{I}_{D_\tau^v}(x_\tau^v) = \begin{cases} 0 & \text{if } (g_\tau^v)_i(x_\tau^v) \leq 0, \forall 1 \leq i \leq m_\tau^v, \\ +\infty & \text{otherwise.} \end{cases}$$

The approximation improves as parameter μ becomes closer to 0. Thus, the new node cost function will be

$$B_\tau^v(x_\tau^v;\mu) \coloneqq f_\tau^v(x_\tau^v) - \lambda \sum_{i=1}^{m_\tau^v} \ln\left(-(g_\tau^v)_i(x_\tau^v)\right).$$

As an extension of the original notation, we use $\psi_{p,(t,v)}^{(k,r)}(\{y_{t-1}^u\},\{z_\tau^u\};\lambda)$ to denote the optimal solution of the following optimization problem defined on a Networked OCO problem instance p:

$$\begin{aligned}\arg\min_{\{x_\tau^u\}} &\sum_{\tau=t}^{t+k-1}\left(\sum_{u\in N_v^r} B_\tau^u(x_\tau^u;\lambda) + \sum_{u\in N_v^r} c_\tau^u(x_\tau^u, x_{\tau-1}^u) + \sum_{(u,q)\in\mathcal{E}(N_v^r)} g_t^{(u,q)}(x_t^u, x_t^q)\right)\\ &\text{s.t. } x_{t-1}^u = y_{t-1}^u, \forall u\in N_v^r,\\ &\quad\ x_\tau^u = z_\tau^u, \forall(\tau,u)\in\partial N_{(t,v)}^{(k,r)}.\end{aligned}$$

Compared with $\psi_{p,(t,v)}^{(k,r)}(\{y_{t-1}^u\},\{z_\tau^u\})$ defined in Section 3.3, the constraints $x_\tau^u\in D_\tau^u$ are removed and the node costs $f_\tau^u(x_\tau^u)$ are replaced with $B_\tau^u(x_\tau;\lambda)$.

A key observation we need to point out is that the perturbation bounds we have shown do not depend on the smoothness constant ℓ_f of node cost functions. That means the perturbation bound

$$\begin{aligned}&\left\|\psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\},\{z_\tau^u\};\lambda\right)_{(t,v)} - \psi_{p,(t,v)}^{(k,r)}\left(\{y_{t-1}^u\},\{(z_\tau^u)'\};\lambda\right)_{(t,v)}\right\|\\ \le{}& C_1 \sum_{(u,\tau)\in\partial N_{(t,v)}^{(k,r)}} \rho_T^{|t-\tau|}\rho_S^{d_\mathcal{G}(v,u)}\left\|z_\tau^u - (z_\tau^u)'\right\|\end{aligned}$$

holds for arbitrary λ, where C_1,ρ_S,ρ_T are specified in Theorem 3.3.3 or Theorem 3.3.4 and are independent of parameter λ. Theorem 3.10 in Forsgren, Gill, and Wright (2002) guarantees that the solutions $\psi_{p,(t,v)}^{(k,r)}(\{y_{t-1}^u\},\{z_\tau^u\};\lambda_k)$ converge to $\psi_{p,(t,v)}^{(k,r)}(\{y_{t-1}^u\},\{z_\tau^u\})$ for any positive sequence $\{\lambda_k\}_{k=1}^\infty$ that tends to zero. Thus the above perturbation bound also holds for $\psi_{p,(t,v)}^{(k,r)}(\{y_{t-1}^u\},\{z_\tau^u\})$ which includes the constraints on individual decisions.

Note that the argument we present in this section also works for Theorems 3.C.6 and 3.3.9.

Proof of Theorem 3.C.8

We first derive an upper bound on the distance between x_t and x_t^*.

Note that for any period t, we have

$$\left\|x_t - \tilde{\psi}_t\left(x_{t-1}\right)_t\right\| \le e_t. \tag{3.64}$$

Thus we see that

$$
\begin{aligned}
\left\|x_t - x_t^*\right\| &= \left\|x_t - \tilde{\psi}_1(x_0)_t\right\| \\
&\le \left\|x_t - \tilde{\psi}_t(x_{t-1})_t\right\| + \sum_{i=1}^{t-1} \left\|\tilde{\psi}_{t-i+1}(x_{t-i})_t - \tilde{\psi}_{t-i}(x_{t-i-1})_t\right\| \\
&\le \left\|x_t - \tilde{\psi}_t(x_{t-1})_t\right\| + \sum_{i=1}^{t-1} C_G \rho_G^i \left\|x_{t-i} - \tilde{\psi}_{t-i}(x_{t-i-1})_{t-i}\right\| \qquad (3.65a) \\
&\le \sum_{i=0}^{t-1} C_0 \rho_G^i \left\|x_{t-i} - \tilde{\psi}_{t-i}(x_{t-i-1})_{t-i}\right\| \qquad (3.65b) \\
&\le \sum_{i=1}^{t} C_0 \rho_G^{t-i} e_i, \qquad (3.65c)
\end{aligned}
$$

where in (3.65a), we used Theorem 3.C.6 and the fact that $\tilde{\psi}_{t-i}(x_{t-i-1})_t$ can be written as

$$\tilde{\psi}_{t-i}(x_{t-i-1})_t = \tilde{\psi}_{t-i+1}\left(\tilde{\psi}_{t-i}(x_{t-i-1})_{t-i}\right)_t .$$

We also used $C_0 \coloneqq \max\{1, C_G\}$ in (3.65b) and (3.64) in (3.65c).

By (3.65) and the Cauchy-Schwarz Inequality, we see that

$$\left\|x_t - x_t^*\right\|^2 \le C_0^2 \left(\sum_{i=1}^{t} \rho_G^{t-i} e_i\right)^2 \le C_0^2 \left(\sum_{i=1}^{t} \rho_G^{t-i}\right) \cdot \left(\sum_{i=1}^{t} \rho_G^{t-i} e_i^2\right) \le \frac{C_0^2}{1-\rho_G} \cdot \left(\sum_{i=1}^{t} \rho_G^{t-i} e_i^2\right).$$

Summing up over t gives that

$$\sum_{t=1}^{H} \left\|x_t - x_t^*\right\|^2 \le \frac{C_0^2}{(1-\rho_G)^2} \cdot \sum_{t=1}^{H} e_t^2.$$

Proof of Lemma 3.C.7

In this section, we show Lemma 3.C.7 holds with following specific constants:

$$
\begin{aligned}
e_t^2 &\coloneqq \left\|x_t - x_{t|t-1}^*\right\|^2 \\
&\le 4C_1^2 C_0^2 \left(\frac{h(r)^2 \rho_G^2}{(1-\rho_T)(1-\rho_G^2 \rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k}\right) \left\|x_{t-1} - x_{t-1}^*\right\|^2 \\
&\quad + \frac{8C_1^2}{\mu} \left(\frac{h(r)^2}{1-\rho_T} \cdot \rho_S^{2r} \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} f_\tau(x_\tau^*) + C_3(r)^2 \cdot \rho_T^{2(k-1)} f_{t+k-1}(x_{t+k-1}^*)\right).
\end{aligned}
\tag{3.66}
$$

Note that, by the principle of optimality, we have

$$x_t^v = \psi_{p,(t,v)}^{(k,r)}\left(\{x_{t-1}^u\}, \{\theta_\tau^u\}\right)_{(t,v)},$$

$$(x^v_{t|t-1})^* = \psi^{(k,r)}_{p,(t,v)}\left(\{x^u_{t-1}\}, \{(x^u_{\tau|t-1})^*\}\right)_{(t,v)}.$$

Recall that we define the quantity $C_3(r) \coloneqq \sum_{\gamma=0}^{r} h(\gamma)\cdot\rho_S^\gamma$ to simplify the notation. Since the exponentially decaying local perturbation bound holds in Definition 3.3.4, we see that

$$\begin{aligned}\left\|x^v_t - (x^v_{t|t-1})^*\right\| \leq{}& C_1\rho_S^r \sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N^r_v}\left\|(x^u_{\tau|t-1})^* - \theta^u_\tau\right\| \\ &+ C_1\rho_T^{k-1}\sum_{u\in N^r_v}\rho_S^{d_{\mathcal{G}}(u,v)}\left\|(x^u_{t+k-1|t-1})^* - \theta^u_{t+k-1}\right\|, \qquad (3.67)\end{aligned}$$

which implies that

$$\begin{aligned}&\left\|x^v_t - (x^v_{t|t-1})^*\right\|^2 \\ &\leq 2C_1^2\rho_S^{2r}\left(\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N^r_v}\left\|(x^u_{\tau|t-1})^* - \theta^u_\tau\right\|\right)^2 \\ &\quad + 2C_1^2\rho_T^{2(k-1)}\left(\sum_{u\in N^r_v}\rho_S^{d_{\mathcal{G}}(u,v)}\left\|(x^u_{t+k-1|t-1})^* - \theta^u_{t+k-1}\right\|\right)^2 \qquad (3.68a) \\ &\leq 2C_1^2\rho_S^{2r}\left(\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N^r_v}1\right)\left(\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N^r_v}\left\|(x^u_{\tau|t-1})^* - \theta^u_\tau\right\|^2\right) \\ &\quad + 2C_1^2\rho_T^{2(k-1)}\left(\sum_{u\in N^r_v}\rho_S^{d_{\mathcal{G}}(u,v)}\right)\left(\sum_{u\in N^r_v}\rho_S^{d_{\mathcal{G}}(u,v)}\left\|(x^u_{t+k-1|t-1})^* - \theta^u_{t+k-1}\right\|^2\right) \qquad (3.68b) \\ &\leq \frac{2C_1^2h(r)}{1-\rho_T}\cdot\rho_S^{2r}\left(\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N^r_v}\left\|(x^u_{\tau|t-1})^* - \theta^u_\tau\right\|^2\right) \\ &\quad + 2C_1^2C_3(r)\cdot\rho_T^{2(k-1)}\left(\sum_{u\in N^r_v}\rho_S^{d_{\mathcal{G}}(u,v)}\left\|(x^u_{t+k-1|t-1})^* - \theta^u_{t+k-1}\right\|^2\right), \qquad (3.68c)\end{aligned}$$

where we used the AM-GM Inequality in (3.68a); we used the Cauchy-Schwarz Inequality in (3.68b); we used the definitions of functions $h(r)$ and $C_3(r)$ in (3.68c).

Summing up (3.68) over all $v\in\mathcal{V}$ and reorganizing terms gives

$$\begin{aligned}&\sum_{v\in\mathcal{V}}\left\|x^v_t - (x^v_{t|t-1})^*\right\|^2 \\ &\leq \frac{2C_1^2h(r)}{1-\rho_T}\cdot\rho_S^{2r}\sum_{v\in\mathcal{V}}\left(\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N^r_v}\left\|(x^u_{\tau|t-1})^* - \theta^u_\tau\right\|^2\right)\end{aligned}$$

$$
\begin{aligned}
&+ 2C_1^2 C_3(r) \cdot \rho_T^{2(k-1)} \sum_{v \in \mathcal{V}} \left(\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| (x_{t+k-1|t-1}^u)^* - \theta_{t+k-1}^u \right\|^2 \right) \\
&\le \frac{2C_1^2 h(r)^2}{1-\rho_T} \cdot \rho_S^{2r} \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \left\| x_{\tau|t-1}^* - \theta_\tau \right\|^2 \\
&\quad + 2C_1^2 C_3(r)^2 \cdot \rho_T^{2(k-1)} \left\| x_{t+k-1|t-1}^* - \theta_{t+k-1} \right\|^2, \qquad (3.69)
\end{aligned}
$$

where we used the facts that

$$
\sum_{v \in \mathcal{V}} \sum_{u \in \partial N_v^r} \left\| (x_{\tau|t-1}^u)^* - \theta_\tau^u \right\|^2 \le h(r) \sum_{v \in \mathcal{V}} \left\| (x_{\tau|t-1}^v)^* - \theta_\tau^v \right\|^2 = h(r) \cdot \left\| x_{\tau|t-1}^* - \theta_\tau \right\|^2,
$$

and

$$
\begin{aligned}
\sum_{v \in \mathcal{V}} \sum_{u \in \partial N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| (x_{t+k-1|t-1}^u)^* - \theta_{t+k-1}^u \right\|^2 &\le C_3(r) \sum_{v \in \mathcal{V}} \left\| (x_{t+k-1|t-1}^v)^* - \theta_{t+k-1}^v \right\|^2 \\
&= C_3(r) \cdot \left\| x_{t+k-1|t-1}^* - \theta_{t+k-1} \right\|^2.
\end{aligned}
$$

We also note that by the principle of optimality, the following equations hold for all $\tau \ge t$:

$$
\begin{aligned}
x_{\tau|t-1}^* &= \tilde{\psi}_t \left(x_{t-1} \right)_\tau, \\
x_\tau^* &= \tilde{\psi}_t \left(x_{t-1}^* \right)_\tau.
\end{aligned}
$$

Recall that $C_0 \coloneqq \max\{1, C_G\}$. By Theorem 3.C.6, we see that

$$
\left\| x_{\tau|t-1}^* - x_\tau^* \right\| \le C_0 \rho_G^{\tau-t+1} \left\| x_{t-1} - x_{t-1}^* \right\|, \qquad (3.70)
$$

which implies

$$
\begin{aligned}
\left\| x_{\tau|t-1}^* - \theta_\tau \right\|^2 &\le 2 \left\| x_{\tau|t-1}^* - x_\tau^* \right\|^2 + 2 \left\| x_\tau^* - \theta_\tau \right\|^2 \qquad (3.71a) \\
&\le 2C_0^2 \rho_G^{2(\tau-t+1)} \left\| x_{t-1} - x_{t-1}^* \right\|^2 + 2 \left\| x_\tau^* - \theta_\tau \right\|^2, \qquad (3.71b)
\end{aligned}
$$

where we used the triangle inequality and the AM-GM inequality in (3.71a); we used (3.70) in (3.71b).

Substituting (3.71) into (3.69) gives

$$
\begin{aligned}
&\sum_{v \in \mathcal{V}} \left\| x_t^v - (x_{t|t-1}^v)^* \right\|^2 \\
&\le 4C_1^2 C_0^2 \left(\frac{h(r)^2 \rho_G^2}{(1-\rho_T)(1-\rho_G^2 \rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k} \right) \left\| x_{t-1} - x_{t-1}^* \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 4C_1^2 \left(\frac{h(r)^2}{1-\rho_T} \cdot \rho_S^{2r} \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \left\| x_\tau^* - \theta_\tau \right\|^2 + C_3(r)^2 \cdot \rho_T^{2(k-1)} \left\| x_{t+k-1}^* - \theta_{t+k-1} \right\|^2 \right) \\
\leq{}& 4C_1^2 C_0^2 \left(\frac{h(r)^2 \rho_G^2}{(1-\rho_T)(1-\rho_G^2 \rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k} \right) \left\| x_{t-1} - x_{t-1}^* \right\|^2 \\
&+ \frac{8C_1^2}{\mu} \left(\frac{h(r)^2}{1-\rho_T} \cdot \rho_S^{2r} \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} f_\tau(x_\tau^*) + C_3(r)^2 \cdot \rho_T^{2(k-1)} f_{t+k-1}(x_{t+k-1}^*) \right),
\end{aligned}
\tag{3.72}
$$

where we used the fact that the node cost function f_τ^v is non-negative and μ-strongly convex for all τ, v, thus

$$
f_\tau(x_\tau^*) \geq \sum_{v \in \mathcal{V}} f_\tau^v((x_\tau^v)^*) \geq \frac{\mu}{2} \sum_{v \in \mathcal{V}} \left\| (x_\tau^v)^* - \theta_\tau^v \right\|^2 = \frac{\mu}{2} \left\| x_\tau^* - \theta_\tau \right\|^2 .
$$

Note that $\sum_{v \in \mathcal{V}} \left\| x_t^v - (x_{t|t-1}^v)^* \right\|^2 = \left\| x_t - x_{t|t-1}^* \right\|^2 = e_t^2$. Thus we have finished the proof of (3.66).

Proof of Lemma 3.C.9

In this section, we show Lemma 3.C.9 holds with following specific constants:

$$
\begin{aligned}
e_t^2 :={}& \left\| x_t - x_{t|t-1}^* \right\|^2 \\
\leq{}& 12C_1^2 C_0^2 \left(\frac{h(r)^2 \rho_G^2}{(1-\rho_T)(1-\rho_G^2 \rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k} \right) \left\| x_{t-1} - x_{t-1}^* \right\|^2 \\
&+ \frac{24C_1^2}{\mu} \left(\frac{h(r)^2}{1-\rho_T} \cdot \rho_S^{2r} \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} f_\tau(x_\tau^*) + C_3(r)^2 \cdot \rho_T^{2(k-1)} f_{t+k-1}(x_{t+k-1}^*) \right) \\
&+ \left(\frac{9C_2^2(1+\Delta^2) C_3(r)^2}{1-\rho_T} + \frac{12C_1^2 C_3(r)^2 \ell_w}{\mu} + \frac{12C_1^2 h(r)^2 \ell_w}{\mu(1-\rho_T)} \cdot \rho_S^{2r} \right) \cdot \\
&\mathsf{PredictionError}_{p,(t,k)}.
\end{aligned}
\tag{3.73}
$$

Note that, by the principle of optimality, we have

$$
\begin{aligned}
x_t^v &= \psi_{p,(t,v)}^{(k,r)} \left(\{x_{t-1}^u\}, \{\theta_{\tau|t}^u\}; \xi_{(t,v)}^{(k,r)} \right)_{(t,v)}, \\
(x_{t|t-1}^v)^* &= \psi_{p,(t,v)}^{(k,r)} \left(\{x_{t-1}^u\}, \{(x_{\tau|t-1}^u)^*\}; (\xi_{(t,v)}^{(k,r)})^* \right)_{(t,v)},
\end{aligned}
$$

where we recall that $\theta_{\tau|t}^u := \arg\min_{y \in D_\tau^u} f_\tau^u(y; \mu_{\tau|t}^u)$ in Algorithm 3 and $\theta_\tau^u := \arg\min_{y \in D_\tau^u} f_\tau^u(y)$.

Since the exponentially decaying local perturbation bound holds in Definition 3.3.7, we see that

$$\begin{aligned}
&\left\|x_t^v - (x_{t|t-1}^v)^*\right\| \\
&\leq C_1 \rho_S^r \underbrace{\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|(x_{\tau|t-1}^u)^* - \theta_{\tau|t}^u\right\|}_{\text{Term 1}} \\
&+ C_1 \rho_T^{k-1} \underbrace{\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\|(x_{t+k-1|t-1}^u)^* - \theta_{t+k-1|t}^u\right\|}_{\text{Term 2}} \\
&+ C_2 \cdot \mathsf{dist}_p\left(\tilde{\xi}_{(t,v)}^{(k,r)}, \left(\xi_{(t,v)}^{(k,r)}\right)^*\right). \qquad (3.74)
\end{aligned}$$

Note that the square of Term 1 in (3.74) can be upper bounded by

$$\begin{aligned}
&(\text{Term } 1)^2 \\
&\leq \left(\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} 1\right)\left(\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|(x_{\tau|t-1}^u)^* - \theta_{\tau|t}^u\right\|^2\right) \qquad (3.75a)\\
&\leq \frac{h(r)}{1-\rho_T}\left(\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|(x_{\tau|t-1}^u)^* - \theta_{\tau|t}^u\right\|^2\right) \qquad (3.75b)\\
&\leq \frac{2h(r)}{1-\rho_T}\left(\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|(x_{\tau|t-1}^u)^* - \theta_\tau^u\right\|^2 + \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|\theta_{\tau|t}^u - \theta_\tau^u\right\|^2\right) \qquad (3.75c)\\
&\leq \frac{2h(r)}{1-\rho_T}\left(\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|(x_{\tau|t-1}^u)^* - \theta_\tau^u\right\|^2 + \frac{2\ell_w}{\mu}\sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \sum_{u \in \partial N_v^r} \left\|\omega_{\tau|t}^u - (\omega_\tau^u)^*\right\|^2\right), \qquad (3.75d)
\end{aligned}$$

where we used the Cauchy-Schwarz Inequality in (3.75a); we used the definition of $h(r)$ in (3.75b); we used the triangle inequality and the AM-GM inequality in (3.75c); we used the special case of generalized exponentially decaying local perturbation bound when the graph is a single node in (3.75d). We also see that

$$\begin{aligned}
&(\text{Term } 2)^2 \\
&\leq \left(\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)}\right)\left(\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\|(x_{t+k-1|t-1}^u)^* - \theta_{t+k-1|t}^u\right\|^2\right) \qquad (3.76a)\\
&\leq C_3(r) \cdot \left(\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\|(x_{t+k-1|t-1}^u)^* - \theta_{t+k-1|t}^u\right\|^2\right) \qquad (3.76b)
\end{aligned}$$

$$\le 2C_3(r) \cdot \Bigg(\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| (x^u_{t+k-1|t-1})^* - \theta^u_{t+k-1} \right\|^2 + \sum_{u\in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| \theta^u_{t+k-1|t} - \theta^u_{t+k-1}\right\|^2 \Bigg) \tag{3.76c}$$

$$\le 2C_3(r) \cdot \Bigg(\sum_{u \in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| (x^u_{t+k-1|t-1})^* - \theta^u_{t+k-1} \right\|^2 + \frac{2\ell_w}{\mu} \sum_{u\in N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| \omega^u_{t+k-1|t} - (\omega^u_{t+k-1})^*\right\|^2 \Bigg), \tag{3.76d}$$

where we used Cauchy-Schwarz Inequality in (3.76a); we used the definition of $C_3(r)$ in (3.76b); we used the triangle inequality and the AM-GM inequality in (3.76c); we used the special case of generalized exponentially decaying local perturbation bound when the graph is a single node in (3.76d). Note that the square of $\mathsf{dist}\left(\tilde{\xi}^{(k,r)}_{(t,v)}, \left(\xi^{(k,r)}_{(t,v)}\right)^*\right)$ can also be bounded by

$$\begin{aligned}
&\mathsf{dist}\left(\tilde{\xi}^{(k,r)}_{(t,v)}, \left(\xi^{(k,r)}_{(t,v)}\right)^*\right)^2 \\
&\le 3\Bigg(\sum_{(u,\tau)\in N^{(k-1,r-1)}_{(t,v)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)}\Bigg)\Bigg(\sum_{(u,\tau)\in N^{(k-1,r-1)}_{(t,v)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)} \left\|\omega^u_{\tau|t} - (\omega^u_\tau)^*\right\|^2\Bigg) \\
&\quad + 3\Bigg(\sum_{(u,\tau)\in N^{(k,r)}_{(t,v)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)}\Bigg)\Bigg(\sum_{(u,\tau)\in N^{(k,r)}_{(t,v)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)} \left\|\alpha^u_{\tau|t} - (\alpha^u_\tau)^*\right\|^2\Bigg) \\
&\quad + 3\Bigg(\sum_{\tau=t}^{t+k}\sum_{e\in\mathcal{E}(N_v^r)} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,e)}\Bigg)\Bigg(\sum_{\tau=t}^{t+k}\sum_{e\in\mathcal{E}(N_v^r)} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,e)} \left\|\beta^e_{\tau|t} - (\beta^e_\tau)^*\right\|^2\Bigg)
\end{aligned} \tag{3.77a}$$

$$\begin{aligned}
&\le \frac{3C_3(r)}{1-\rho_T} \cdot \Bigg(\sum_{(u,\tau)\in N^{(k-1,r-1)}_{(t,v)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)} \left\|\omega^u_{\tau|t} - (\omega^u_\tau)^*\right\|^2 \\
&\quad + \sum_{(u,\tau)\in N^{(k,r)}_{(t,v)}} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)} \left\|\alpha^u_{\tau|t} - (\alpha^u_\tau)^*\right\|^2\Bigg) \\
&\quad + \frac{3\Delta C_3(r)}{1-\rho_T} \cdot \sum_{\tau=t}^{t+k}\sum_{e\in\mathcal{E}(N_v^r)} \rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,e)} \left\|\beta^e_{\tau|t} - (\beta^e_\tau)^*\right\|^2,
\end{aligned} \tag{3.77b}$$

where we used Cauchy-Schwarz Inequality in (3.77a); we used the definition of $C_4(r)$ in (3.77b).

Substituting (3.75), (3.76), and (3.77) into (3.74) gives that

$$\begin{aligned}
&\left\|x_t^v - (x_{t|t-1}^v)^*\right\|^2 \\
&\le 3C_1^2\rho_S^{2r}(\text{Term 1})^2 + 3C_1^2\rho_T^{2(k-1)}(\text{Term 2})^2 + 3C_2^2\mathsf{dist}_p\left(\tilde{\xi}_{(t,v)}^{(k,r)}, \left(\xi_{(t,v)}^{(k,r)}\right)^*\right)^2 \\
&\le \frac{6C_1^2 h(r)\cdot\rho_S^{2r}}{1-\rho_T}\left(\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N_v^r}\left\|(x_{\tau|t-1}^u)^* - \theta_\tau^u\right\|^2\right. \\
&\quad \left. + \frac{2\ell_w}{\mu}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\sum_{u\in\partial N_v^r}\left\|\omega_{\tau|t}^u - (\omega_\tau^u)^*\right\|^2\right) \\
&\quad + 6C_1^2C_3(r)\cdot\rho_T^{2(k-1)}\left(\sum_{u\in N_v^r}\rho_S^{d_{\mathcal{G}}(u,v)}\left\|(x_{t+k-1|t-1}^u)^* - \theta_{t+k-1}^u\right\|^2\right. \\
&\quad \left. + \frac{2\ell_w}{\mu}\sum_{u\in N_v^r}\rho_S^{d_{\mathcal{G}}(u,v)}\left\|\omega_{t+k-1|t}^u - (\omega_{t+k-1}^u)^*\right\|^2\right) \\
&\quad + \frac{9C_2^2C_3(r)}{1-\rho_T}\cdot\left(\sum_{(u,\tau)\in N_{(t,v)}^{(k-1,r-1)}}\rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)}\left\|\mu_{\tau|t}^u - (\mu_\tau^u)^*\right\|^2\right. \\
&\quad \left. + \sum_{(u,\tau)\in N_{(t,v)}^{(k,r)}}\rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,u)}\left\|\alpha_{\tau|t}^u - (\alpha_\tau^u)^*\right\|^2\right) \\
&\quad + \frac{9C_2^2\Delta C_3(r)}{1-\rho_T}\cdot\sum_{\tau=t}^{t+k}\sum_{e\in\mathcal{E}(N_v^r)}\rho_T^{|t-\tau|}\rho_S^{d_{\mathcal{G}}(v,e)}\left\|\beta_{\tau|t}^e - (\beta_\tau^e)^*\right\|^2. \qquad (3.78)
\end{aligned}$$

Summing up (3.78) for all $v \in \mathcal{V}$ and reorganizing terms gives

$$\begin{aligned}
&\sum_{v\in\mathcal{V}}\left\|x_t^v - (x_{t|t-1}^v)^*\right\|^2 \\
&\le \frac{6C_1^2h(r)^2}{1-\rho_T}\cdot\rho_S^{2r}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\left\|x_{\tau|t-1}^* - \theta_\tau\right\|^2 + 6C_1^2C_3(r)^2\cdot\rho_T^{2(k-1)}\left\|x_{t+k-1|t-1}^* - \theta_{t+k-1}\right\|^2 \\
&\quad + \frac{12C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\left\|\omega_{\tau|t} - \omega_\tau^*\right\|^2 \\
&\quad + \frac{12C_1^2C_3(r)^2\ell_w}{\mu}\cdot\rho_T^{2(k-1)}\left\|\omega_{t+k-1|t} - \omega_{t+k-1}^*\right\|^2 \\
&\quad + \frac{9C_5^2C_3(r)^2}{1-\rho_T}\cdot\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\left\|\omega_{\tau|t} - \omega_\tau^*\right\|^2 + \frac{9C_2^2C_3(r)^2}{1-\rho_T}\cdot\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\left\|\alpha_{\tau|t}^{\mathcal{V}} - (\alpha_\tau^{\mathcal{V}})^*\right\|^2 \\
&\quad + \frac{9C_2^2\Delta^2C_3(r)^2}{1-\rho_T}\sum_{\tau=t}^{t+k}\rho_T^{\tau-t}\left\|\beta_{\tau|t}^{\mathcal{E}} - (\beta_\tau^{\mathcal{E}})^*\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{6C_1^2 h(r)^2}{1-\rho_T} \cdot \rho_S^{2r} \sum_{\tau=t}^{t+k-1} \rho_T^{\tau-t} \left\| x_{\tau|t-1}^* - \theta_\tau \right\|^2 \\
&\quad + 6C_1^2 C_3(r)^2 \cdot \rho_T^{2(k-1)} \left\| x_{t+k-1|t-1}^* - \theta_{t+k-1} \right\|^2 \\
&\quad + \left(\frac{9C_2^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{12C_1^2 C_3(r)^2 \ell_w}{\mu} + \frac{12C_1^2 h(r)^2 \ell_w}{\mu(1-\rho_T)} \cdot \rho_S^{2r} \right) \\
&\quad \cdot \mathsf{PredictionError}_{p,(t,k)},
\end{aligned} \tag{3.79}
$$

where we used the facts that

$$
\sum_{v \in \mathcal{V}} \sum_{u \in \partial N_v^r} \left\| (x_{\tau|t-1}^u)^* - \theta_\tau^u \right\|^2 \leq h(r) \sum_{v \in \mathcal{V}} \left\| (x_{\tau|t-1}^v)^* - \theta_\tau^v \right\|^2 = h(r) \cdot \left\| x_{\tau|t-1}^* - \theta_\tau \right\|^2,
$$

and

$$
\begin{aligned}
\sum_{v \in \mathcal{V}} \sum_{u \in \partial N_v^r} \rho_S^{d_{\mathcal{G}}(u,v)} \left\| (x_{t+k-1|t-1}^u)^* - \theta_{t+k-1}^u \right\|^2 &\leq C_3(r) \sum_{v \in \mathcal{V}} \left\| (x_{t+k-1|t-1}^v)^* - \theta_{t+k-1}^v \right\|^2 \\
&= C_3(r) \cdot \left\| x_{t+k-1|t-1}^* - \theta_{t+k-1} \right\|^2.
\end{aligned}
$$

We also note that by the principle of optimality, the following equations hold for all $\tau \geq t$:

$$
x_{\tau|t-1}^* = \tilde{\psi}_t\left(x_{t-1}\right)_\tau, \; x_\tau^* = \tilde{\psi}_t\left(x_{t-1}^*\right)_\tau.
$$

Recall that $C_0 \coloneqq \max\{1, C_G\}$. By Theorem 3.C.6, we see that

$$
\left\| x_{\tau|t-1}^* - x_\tau^* \right\| \leq C_0 \rho_G^{\tau-t+1} \left\| x_{t-1} - x_{t-1}^* \right\|, \tag{3.80}
$$

which implies

$$
\begin{aligned}
\left\| x_{\tau|t-1}^* - \theta_\tau \right\|^2 &\leq 2\left\| x_{\tau|t-1}^* - x_\tau^* \right\|^2 + 2\left\| x_\tau^* - \theta_\tau \right\|^2 &\text{(3.81a)} \\
&\leq 2C_0^2 \rho_G^{2(\tau-t+1)} \left\| x_{t-1} - x_{t-1}^* \right\|^2 + 2\left\| x_\tau^* - \theta_\tau \right\|^2, &\text{(3.81b)}
\end{aligned}
$$

where we used the triangle inequality and the AM-GM inequality in (3.81a); we used (3.80) in (3.81b).

Substituting (3.81) into (3.79) gives

$$
\begin{aligned}
&\sum_{v \in \mathcal{V}} \left\| x_t^v - (x_{t|t-1}^v)^* \right\|^2 \\
&\leq 12C_1^2 C_0^2 \left(\frac{h(r)^2 \rho_G^2}{(1-\rho_T)(1-\rho_G^2 \rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k} \right) \left\| x_{t-1} - x_{t-1}^* \right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 12C_1^2\left(\frac{h(r)^2}{1-\rho_T}\cdot\rho_S^{2r}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}\left\|x_\tau^* - \theta_\tau\right\|^2 + C_3(r)^2\cdot\rho_T^{2(k-1)}\left\|x_{t+k-1}^* - \theta_{t+k-1}\right\|^2\right)\\
&+\left(\frac{9C_2^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{12C_1^2C_3(r)^2\ell_w}{\mu} + \frac{12C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\right)\\
&\cdot \mathsf{PredictionError}_{p,(t,k)}\\
\le{}& 12C_1^2C_0^2\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\cdot\rho_G^{2k}\right)\left\|x_{t-1} - x_{t-1}^*\right\|^2\\
&+\frac{24C_1^2}{\mu}\left(\frac{h(r)^2}{1-\rho_T}\cdot\rho_S^{2r}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}f_\tau(x_\tau^*) + C_3(r)^2\cdot\rho_T^{2(k-1)}f_{t+k-1}(x_{t+k-1}^*)\right)\\
&+\left(\frac{9C_2^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{12C_1^2C_3(r)^2\ell_w}{\mu} + \frac{12C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\right)\\
&\cdot \mathsf{PredictionError}_{p,(t,k)}, \qquad (3.82)
\end{aligned}
$$

where we used the fact that the node cost function f_τ^v is non-negative and μ-strongly convex for all τ, v, thus

$$
f_\tau(x_\tau^*) \ge \sum_{v\in\mathcal{V}} f_\tau^v((x_\tau^v)^*) \ge \frac{\mu}{2}\sum_{v\in\mathcal{V}}\left\|(x_\tau^v)^* - \theta_\tau^v\right\|^2 = \frac{\mu}{2}\left\|x_\tau^* - \theta_\tau\right\|^2.
$$

Note that $\sum_{v\in\mathcal{V}}\left\|x_t^v - (x_{t|t-1}^v)^*\right\|^2 = \left\|x_t - x_{t|t-1}^*\right\|^2 = e_t^2$. Thus we have finished the proof of (3.74).

Proof of Theorem 3.3.5

In this section, we show Theorem 3.3.5 holds with the following specific constants:

$$
\begin{aligned}
1 &+ \left(1 + \frac{32C_0^2C_1^2(\ell_f + \Delta\ell_S + 2\ell_T)\cdot h(r)^2}{\mu(1-\rho_G)^2(1-\rho_T)^2}\right)\cdot\rho_S^r\\
&+\left(1 + \frac{32C_0^2C_1^2(\ell_f + \Delta\ell_S + 2\ell_T)C_3(r)^2}{\mu(1-\rho_G)^2}\right)\rho_T^{k-1} \qquad (3.83)
\end{aligned}
$$

under the assumption that

$$
\frac{4C_1^2C_0^4}{(1-\rho_G)^2}\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\cdot\rho_G^{2k}\right)\le\frac{1}{2}. \qquad (3.84)
$$

Recall that C_0 is defined in Theorem 3.C.8. Note that Theorems 3.3.3 and 3.C.6 hold for $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \Delta, h)$. One can check that $C_0, C_1, (1-\rho_G)^{-1}$, and $(1-\rho_T)^{-1}$ are bounded by polynomials of ℓ_f/μ, ℓ_T/μ, and $(\Delta\ell_S)/\mu$.

In the proof, we need to use Lemma F.2 in Lin, Hu, Shi, et al. (2021) to bound LPC's total cost by a weighted sum of the offline optimal cost and the sum of squared distances between their trajectories. For completeness, we present Lemma F.2 in Lin, Hu, Shi, et al. (2021) below:

Lemma 3.D.5. *For a fixed dimension $m \in \mathbb{Z}_+$, assume a function $h : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is convex, ℓ-smooth and continuously differentiable. For all $x, y \in \mathbb{R}^m$, for all $\eta > 0$, we have*

$$h(x) \leq (1+\eta)h(y) + \frac{\ell}{2}\left(1 + \frac{1}{\eta}\right)\|x - y\|^2.$$

Now we come back to the proof of Theorem 3.3.5. We first bound the sum of squared distances between LPC's trajectory and the offline optimal trajectory:

$$\begin{aligned}
&\sum_{t=1}^{H} \|x_t - x_t^*\|^2 \\
&\leq \frac{C_0^2}{(1-\rho_G)^2}\sum_{t=1}^{H} e_t^2 && (3.85a)\\
&\leq \frac{4C_1^2C_0^4}{(1-\rho_G)^2}\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\cdot\rho_G^{2k}\right)\sum_{t=1}^{H}\|x_{t-1} - x_{t-1}^*\|^2 \\
&\quad + \frac{8C_0^2C_1^2}{\mu(1-\rho_G)^2}\sum_{t=1}^{H}\left(\frac{h(r)^2}{1-\rho_T}\cdot\rho_S^{2r}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}f_\tau(x_\tau^*)\right. \\
&\quad \left. + C_3(r)^2\cdot\rho_T^{2(k-1)}f_{t+k-1}(x_{t+k-1}^*)\right), && (3.85b)
\end{aligned}$$

where we used Theorem 3.C.8 in (3.85a); we used Lemma 3.C.7 with the specific constants given in Section 3.D in (3.85b).

Recall that in (3.84), we assume r and k are sufficient large so that the coefficient of the first term in (3.85) satisfies

$$\frac{4C_1^2C_0^4}{(1-\rho_G)^2}\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\cdot\rho_G^{2k}\right) \leq \frac{1}{2}.$$

Substituting this bound into (3.85) gives that

$$\sum_{t=1}^{H}\|x_t - x_t^*\|^2 \leq \frac{16C_0^2C_1^2}{\mu(1-\rho_G)^2}\left(\frac{h(r)^2}{(1-\rho_T)^2}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\right)\cdot\sum_{t=1}^{H}f_t(x_t^*). \tag{3.86}$$

By Lemma 3.D.5, since f_t is $(\ell_f + \Delta\ell_S)$-smooth, convex, and non-negative on $\mathbb{R}^n$, and c_t is ℓ_T-smooth, convex, and non-negative on $\mathbb{R}^n \times \mathbb{R}^n$, we know that

$$\begin{aligned} f_t(x_t) &\le (1+\eta) f_t(x_t^*) + \frac{\ell_f + \Delta\ell_S}{2}\left(1 + \frac{1}{\eta}\right)\|x_t - x_t^*\|^2 \\ c_t(x_t, x_{t-1}) &\le (1+\eta) c_t(x_t^*, x_{t-1}^*) + \frac{\ell_T}{2}\left(1 + \frac{1}{\eta}\right)\left(\|x_t - x_t^*\|^2 + \|x_{t-1} - x_{t-1}^*\|^2\right) \end{aligned} \tag{3.87}$$

holds for any $\eta > 0$. Summing the above inequality over t gives

$$\begin{aligned} &\sum_{t=1}^{H}\left(f_t(x_t) + c_t(x_t, x_{t-1})\right) \\ &\le (1+\eta)\sum_{t=1}^{H}\left(f_t(x_t^*) + c_t(x_t^*, x_{t-1}^*)\right) + \frac{(\ell_f + \Delta\ell_S + 2\ell_T)}{2}\left(1 + \frac{1}{\eta}\right)\sum_{t=1}^{H}\|x_t - x_t^*\|^2 \\ &\le (1+\eta)\mathrm{cost}_p(\mathsf{OPT}) \\ &\quad + \left(1 + \frac{1}{\eta}\right)\frac{16C_0^2C_1^2(\ell_f + \Delta\ell_S + 2\ell_T)}{\mu(1-\rho_G)^2}\left(\frac{h(r)^2}{(1-\rho_T)^2}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\right) \\ &\quad \cdot\mathrm{cost}_p(\mathsf{OPT}), \end{aligned} \tag{3.88}$$

where we used (3.86) and $\sum_{t=1}^{H} f_t(x_t^*) \le \mathrm{cost}_p(\mathsf{OPT})$ in the last inequality. Setting $\eta = \rho_S^r + \rho_T^{k-1}$ in (3.88) finishes the proof of (3.83).

As a remark, we require the local cost function $\left(f_t^v, c_t^v, s_t^e\right)$ to be non-negative, convex, and smooth in the whole Euclidean spaces $(\mathbb{R}^n, \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n \times \mathbb{R}^n)$ in Assumption 3.3.1 because we want to apply Lemma 3.D.5 in (3.87).

Proof of Theorem 3.3.10

In this section, we show Theorem 3.3.10 holds with the following specific constants: Under the assumption that the following inequality holds

$$\frac{12C_1^2C_0^4}{(1-\rho_G)^2}\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\cdot\rho_G^{2k}\right) \le \frac{1}{2}, \tag{3.89}$$

the coefficient before $\mathrm{cost}_p(\mathsf{OPT})$ is given by

$$\left(1 + \frac{96C_0^2C_1^2(\ell_f + \Delta\ell_S + 2\ell_T)\cdot h(r)^2}{\mu(1-\rho_G)^2(1-\rho_T)^2}\right)\rho_S^r + \left(1 + \frac{96C_0^2C_1^2(\ell_f + \Delta\ell_S + 2\ell_T)C_3(r)^2}{\mu(1-\rho_G)^2}\right)\rho_T^{k-1}, \tag{3.90}$$

and the additive term is

$$\left(\frac{18C_5^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{24C_1^2C_3(r)^2\ell_w}{\mu} + \frac{24C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\right)\cdot\sum_{\tau=0}^{k-1}\rho_T^\tau\Gamma_\tau(p). \tag{3.91}$$

Recall that C_0 is defined in Theorem 3.C.8. Note that Theorems 3.C.6 and 3.3.9 hold for $\mathcal{P}(\mu, \ell_f, \ell_T, \ell_S, \ell_w, \Delta, h)$. One can check that $C_0, C_1, (1-\rho_G)^{-1}$, and $(1-\rho_T)^{-1}$ are bounded by polynomials of $\ell_f/\mu, \ell_T/\mu$, and $(\Delta \ell_S)/\mu$.

In the proof, we need to use Lemma F.2 in Lin, Hu, Shi, et al. (2021) to bound LPC's total cost by a weighted sum of the offline optimal cost and the sum of squared distances between their trajectories. For completeness, we present Lemma F.2 in Lin, Hu, Shi, et al. (2021) below:

Lemma 3.D.6. *For a fixed dimension $m \in \mathbb{Z}_+$, assume a function $h : \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is convex, ℓ-smooth and continuously differentiable. For all $x, y \in \mathbb{R}^m$, for all $\eta > 0$, we have*

$$h(x) \leq (1+\eta)h(y) + \frac{\ell}{2}\left(1 + \frac{1}{\eta}\right)\|x - y\|^2.$$

Now we come back to the proof of Theorem 3.3.10. We first bound the sum of squared distances between LPC's trajectory and the offline optimal trajectory:

$$\begin{aligned}
&\sum_{t=1}^{H} \|x_t - x_t^*\|^2 \\
&\leq \frac{C_0^2}{(1-\rho_G)^2}\sum_{t=1}^{H} e_t^2 \qquad\qquad (3.92a)\\
&\leq \frac{12C_1^2C_0^4}{(1-\rho_G)^2}\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k}\right)\sum_{t=1}^{H}\|x_{t-1} - x_{t-1}^*\|^2 \\
&\quad + \frac{24C_0^2C_1^2}{\mu(1-\rho_G)^2}\sum_{t=1}^{H}\left(\frac{h(r)^2}{1-\rho_T}\cdot\rho_S^{2r}\sum_{\tau=t}^{t+k-1}\rho_T^{\tau-t}f_\tau(x_\tau^*) + C_3(r)^2\cdot\rho_T^{2(k-1)}f_{t+k-1}(x_{t+k-1}^*)\right) \\
&\quad + \left(\frac{9C_5^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{12C_1^2C_3(r)^2\ell_w}{\mu} + \frac{12C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\right)\cdot\sum_{\tau=0}^{k-1}\rho_T^\tau\Gamma_\tau(p),
\end{aligned}$$
$$(3.92b)$$

where we used Theorem 3.C.8 in (3.92a); we used Lemma 3.C.7 with the specific constants given in Section 3.D in (3.92b).

Recall that in (3.89), we assume r and k are sufficient large so that the coefficient of the first term in (3.92) satisfies

$$\frac{12C_1^2C_0^4}{(1-\rho_G)^2}\left(\frac{h(r)^2\rho_G^2}{(1-\rho_T)(1-\rho_G^2\rho_T)} \cdot \rho_S^{2r} + C_3(r)^2 \cdot \rho_T^{2(k-1)} \cdot \rho_G^{2k}\right) \leq \frac{1}{2}.$$

Substituting this bound into (3.92) gives that

$$\sum_{t=1}^{H}\|x_t - x_t^*\|^2$$

$$\leq \frac{48C_0^2C_1^2}{\mu(1-\rho_G)^2}\left(\frac{h(r)^2}{(1-\rho_T)^2}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\right)\cdot\sum_{t=1}^{H} f_t(x_t^*)$$
$$+\left(\frac{18C_5^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{24C_1^2C_3(r)^2\ell_w}{\mu} + \frac{24C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\right)\cdot\sum_{\tau=0}^{k-1}\rho_T^{\tau}\Gamma_\tau(p). \tag{3.93}$$

By Lemma 3.D.6, since f_t is $(\ell_f + \Delta\ell_S)$-smooth, convex, and non-negative on $\mathbb{R}^n$, and c_t is ℓ_T-smooth, convex, and non-negative on $\mathbb{R}^n \times \mathbb{R}^n$, we know that

$$\begin{aligned} f_t(x_t) &\leq (1+\eta)f_t(x_t^*) + \frac{\ell_f + \Delta\ell_S}{2}\left(1+\frac{1}{\eta}\right)\left\|x_t - x_t^*\right\|^2 \\ c_t(x_t, x_{t-1}) &\leq (1+\eta)c_t(x_t^*, x_{t-1}^*) + \frac{\ell_T}{2}\left(1+\frac{1}{\eta}\right)\left(\left\|x_t - x_t^*\right\|^2 + \left\|x_{t-1} - x_{t-1}^*\right\|^2\right) \end{aligned} \tag{3.94}$$

holds for any $\eta > 0$. Summing the above inequality over t gives

$$\begin{aligned} &\sum_{t=1}^{H}\left(f_t(x_t) + c_t(x_t, x_{t-1})\right) \\ &\leq (1+\eta)\sum_{t=1}^{H}\left(f_t(x_t^*) + c_t(x_t^*, x_{t-1}^*)\right) + \frac{(\ell_f + \Delta\ell_S + 2\ell_T)}{2}\left(1+\frac{1}{\eta}\right)\sum_{t=1}^{H}\left\|x_t - x_t^*\right\|^2 \\ &\leq (1+\eta)\mathrm{cost}_p(\mathsf{OPT}) \\ &\quad + \left(1+\frac{1}{\eta}\right)\frac{16C_0^2C_1^2(\ell_f + \Delta\ell_S + 2\ell_T)}{\mu(1-\rho_G)^2}\left(\frac{h(r)^2}{(1-\rho_T)^2}\cdot\rho_S^{2r} + C_3(r)^2\cdot\rho_T^{2(k-1)}\right) \\ &\quad \cdot \mathrm{cost}(\mathsf{OPT}) \\ &\quad + \left(\frac{18C_5^2(1+\Delta^2)C_3(r)^2}{1-\rho_T} + \frac{24C_1^2C_3(r)^2\ell_w}{\mu} + \frac{24C_1^2h(r)^2\ell_w}{\mu(1-\rho_T)}\cdot\rho_S^{2r}\right)\cdot\sum_{\tau=0}^{k-1}\rho_T^{\tau}\Upsilon_\tau, \end{aligned} \tag{3.95}$$

where we used (3.93) and $\sum_{t=1}^{H} f_t(x_t^*) \leq \mathrm{cost}_p(\mathsf{OPT})$ in the last inequality. Setting $\eta = \rho_S^r + \rho_T^{k-1}$ in (3.95) finishes the proof of (3.90).

As a remark, we require the local cost function $\left(f_t^v, c_t^v, s_t^e\right)$ to be non-negative, convex, and smooth in the whole Euclidean spaces $(\mathbb{R}^n, \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n \times \mathbb{R}^n)$ in Assumption 3.3.1 because we want to apply Lemma 3.D.6 in (3.94).

Proof of Corollary 3.3.1

Under the assumptions, by Theorem 3.3.3 in Section 3.D, we see that the exponentially decaying local perturbation bound (Definition 3.3.4) holds with

$$\rho_T = \sqrt{1 - 2\left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1}} \leq \frac{1}{2},\ \rho_S = \sqrt{1 - 2\left(\sqrt{1 + (\Delta\ell_S/\mu)} + 1\right)^{-1}} \leq \frac{1}{4},$$

and $C_1 = 2\sqrt{(\Delta\ell_S/\mu)(\ell_T/\mu)} \leq 0.702$. We also see that

$$\rho_G = 1 - 2\left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1} \leq \frac{1}{4}, \text{ and}$$
$$C_3(r) = \sum_{\gamma=0}^{r} h(\gamma) \cdot \rho_S^\gamma \leq C \sum_{\gamma=0}^{r} 2^{-\frac{3}{2}\gamma} \leq \frac{4C}{4 - \sqrt{2}}.$$

Substituting these bounds into (3.83) and (3.84) in the proof of Theorem 3.3.5 finishes the proof of Corollary 3.3.1, where the numerical constants are $C_1' = 782, C_2' = 936$.

Proof of Corollary 3.3.2

We apply Theorem 3.3.4 and set $b_1 = 2\Delta - 1$ and $b_2 = 4\Delta^2 - 2\Delta$. Since we have $h(\gamma) \leq \Delta^\gamma$, we see that $a \leq 2$ and $\tilde{a} \leq 2$. We also see that

$$\max\{8\tilde{a}\ell_T/\mu, \Delta\ell_S(b_1 + b_2)/(4\mu)\} \leq 1.$$

Thus, by Theorem 3.3.4, we see that the exponentially decaying local perturbation bound holds with

$$\rho_T \leq \frac{1}{2},\ \rho_S \leq \Delta^{-4},\ C_1 \leq 2.$$

We also see that

$$\rho_G = 1 - 2\left(\sqrt{1 + (2\ell_T/\mu)} + 1\right)^{-1} \leq \frac{1}{32}, \text{ and}$$
$$C_3(r) = \sum_{\gamma=0}^{r} h(\gamma) \cdot \rho_S^\gamma \leq \sum_{\gamma=0}^{r} \Delta^{-3\gamma} \leq 2.$$

Substituting these bounds into (3.83) and (3.84) in the proof of Theorem 3.3.5 finishes the proof of Corollary 3.3.1, where the numerical constants are $C_1' = 546, C_2' = 1092$.

Proof of Theorem 3.3.6

In this section we prove a lower bound on the competitive ratio of any online algorithm. Our proof focuses on temporal and spatial lower bounds separately first, and then combines them.

Step 1: Temporal Lower Bounds We first show that the competitive ratio of any online algorithm with k steps of future predictions is lower bounded by $1 + \Omega(\lambda_T^k)$. To show this, we consider the special case when there are no spatial interaction costs (i.e., $s_t^e \equiv 0$ for all t and e). In this case, since all agents are independent with each other, it suffices to assume there is only one agent in the network $\mathcal{G}$. Thus we will drop the agent index in the following analysis. To further simplify the problem, we assume dimension $n = 1$, $c_t(x_t, x_{t-1}) = \frac{\ell_T}{2}(x_t - x_{t-1})^2$, and the feasible set is $D_t \equiv D = [0, 1]$ for all t. Let R denote the diameter of D, i.e., $R = \sup_{x,y \in D} |x - y| = 1$.

By Theorem 2 in Li, Qu, and Li (2021) and Case 1 in its proof, we know that for any online algorithm ALG with k-period future predictions and $L_T \in (2R, RH)$, there exists a problem instance with quadratic functions $f_1, f_2, \ldots, f_H$ that have the form $f_t(x_t) = \frac{\mu}{2}(x_t - \theta_t)^2, \theta_t \in D$ such that

$$\text{cost}_p(\mathsf{ALG}) - \text{cost}_p(\mathsf{OPT}) \geq \frac{\mu^3(1 - \sqrt{\lambda_T})^2}{96(\mu + 1)^2} \cdot \lambda_T^k \cdot R \cdot L_H, \tag{3.96}$$

where $L_H \geq \sum_{t=1}^H |\theta_t - \theta_{t-1}|$. Note that

$$R \cdot L_T \geq \sum_{t=1}^H |v_t - v_{t-1}|^2 = \frac{2}{\ell_T} \cdot \sum_{t=1}^H (f_t(v_t) + c_t(v_t, v_{t-1})) \geq \frac{2}{\ell_T} \cdot \text{cost}_p(\mathsf{OPT}).$$

Substituting this into (3.96) gives

$$\text{cost}_p(\mathsf{ALG}) \geq \left(1 + \frac{\mu^3(1 - \sqrt{\lambda_T})^2}{48(\mu + 1)^2 \ell_T} \cdot \lambda_T^k\right) \cdot \text{cost}_p(\mathsf{OPT}). \tag{3.97}$$

Note that (3.96) implies $\text{cost}_p(\mathsf{ALG}) > 0$, hence the competitive ratio can be unbounded if $\text{cost}_p(\mathsf{OPT}) = 0$.

Step 2: Spatial Lower Bounds We next show that the competitive ratio of any online algorithm that can communicate within r-hop neighborhood according to the scheme defined in Section 3.3 is lower bounded by $1 + \Omega(\lambda_S^r)$. To show this, we will construct a special Networked OCO instance with random cost functions and show there exists a realization that achieves the lower bound by probabilistic methods.

Theorem 3.D.7. *Under the assumption that $\Delta \geq 3$, the competitive ratio of any decentralized online algorithm ALG with communication radius r is lower bounded by $1+\Omega(\lambda_S^r)$, where $\Omega(\cdot)$ notation hides factors that depend polynomially on $1/\mu$, ℓ_T, ℓ_S,* *and Δ. Depending on the value of $\delta\ell_S/\mu$, the decay factor λ_S is given by the following equations:*

$$\lambda_S = \begin{cases} \frac{(\Delta\ell_S/\mu)}{3+3(\Delta\ell_S/\mu)} & \text{if } \Delta\ell_S/\mu < 48, \\ \max\left(\frac{(\Delta\ell_S/\mu)}{3+3(\Delta\ell_S/\mu)}, \left(1 - 4\sqrt{3}\cdot(\Delta\ell_S/\mu)^{-\frac{1}{2}}\right)^2\right) & \text{otherwise.} \end{cases} \tag{3.98}$$

Proof of Theorem 3.D.7. In the proof, we assume the online game only lasts one period before it ends, i.e., $H = 1$. Note that when $H > 1$, the same counterexample can be constructed repeatedly by letting the temporal interaction costs $c_t^v \equiv 0$ for every node v and period t. To simplify the notation, we define $\ell \coloneqq \ell_S/\mu$ and $d \coloneqq [\Delta/2]$. Without the loss of generality, we assume $\mathcal{V} = \{1, 2, \cdots, n\}$ so that each node has a positive integer index.

We consider the case where the node cost function for each node i is $(x_i+w_i)^2$ and the spatial interaction cost between two neighboring nodes i and j is $\ell(x_i - x_j)^2$. Here, $x_i \in \mathbb{R}$ is the scalar action of node i, and parameter $w_i \in \mathbb{R}$ is a local information that corresponds to node i. The parameters $\{w_i\}_{i=1}^n$ are sampled i.i.d. from some distribution $\mathcal{D}$, which we will discuss later.

For a general graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of nodes, let L denote its graph Laplacian matrix. Recall that the graph Laplacian matrix $L \in \mathcal{V} \times \mathcal{V}$ is a symmetric $n \times n$ matrix and it is defined as

$$L_{i,j} = \begin{cases} deg(i) & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } (i, j) \in \mathcal{E}, \\ 0 & \text{otherwise,} \end{cases}$$

for nodes $i, j \in \mathcal{V}$. Here $deg(\cdot)$ denotes the degree of a node in graph $\mathcal{G}$. We know that L is a symmetric semi-definite positive semi-definite and has bandwidth 1 w.r.t. to $\mathcal{G}$. The centralized optimization problem can be expressed as

$$\begin{aligned} \text{cost}_p(\mathsf{OPT}) &= \min_{x\in\mathbb{R}^n} (x + w)^\top (x + w) + \ell \cdot x^\top Lx \\ &= \min_{x\in\mathbb{R}^n} \left\|(I + \ell \cdot L)^{\frac{1}{2}} x + (I + \ell \cdot L)^{-\frac{1}{2}} w\right\|^2 + w^\top (I - (I + \ell \cdot L)^{-1})w \\ &= w^\top (I - (I + \ell \cdot L)^{-1})w, \end{aligned}$$

where the minimum is attained at $x^* = (I + \ell \cdot L)^{-1} w$. Since each node i only has an observation radius of r, it can only observe the part of w that is within N_i^r. To simplify the notation, we define the mask operator $\phi_S : \mathbb{R}^n \to \mathbb{R}^n$ with respect to a subset of nodes $S \subseteq \mathcal{V}$ as

$$\phi_S(w)_i = \begin{cases} w_i & \text{if } i \in S, \\ 0 & \text{otherwise,} \end{cases}$$

for $i \in \mathcal{V}$. The local policy of node i (denote as π_i) is a mapping from $w_{N_i^r}$ to the local decision x_i.

Suppose the distribution $\mathcal{D}$ of each local parameters w_i is a mean-zero distribution with support on $\mathbb{R}$. For every node $i \in \mathcal{V}$, we see that

$$\begin{aligned} \mathbb{E}_w \left| x_i(w) - x_i^*(w) \right|^2 &= \min_{\pi_i} \mathbb{E}_w \left| \pi_i(w_{N_i^r}) - x_i^*(w) \right|^2 \\ &\geq \mathbb{E}_w \left| \mathbb{E}[x_i^*(w) \mid w_{N_i^r}] - x_i^*(w) \right|^2 \qquad (3.99a) \\ &= \mathbb{E}_w \left| \mathbb{E}[\left((I + \ell \cdot L)^{-1} w\right)_i \mid w_{N_i^r}] - \left((I + \ell \cdot L)^{-1} w\right)_i \right|^2 \\ &= \mathbb{E}_w \left| \left((I + \ell \cdot L)^{-1} \phi_{N_i^r}(w)\right)_i - \left((I + \ell \cdot L)^{-1} w\right)_i \right|^2 \qquad (3.99b) \\ &= \mathbb{E}_w \left| \left((I + \ell \cdot L)^{-1} \phi_{N_{-i}^r}(w)\right)_i \right|^2, \qquad (3.99c) \end{aligned}$$

where we use the fact that conditional expectations minimize the mean square prediction error in (3.99a); we use the requirement that the distribution of w is mean-zero in (3.99b).

To bound the variance term in (3.99c), we need the following lemma to lower bound the magnitude of every entry in the exponential decaying matrix $(I + \ell \cdot L)^{-1}$:

Lemma 3.D.8. *There exists a finite graph $\mathcal{G}$ with maximum degree $2d$ that satisfies the following conditions: For any two vertices i, j such that $d_{\mathcal{G}}(i, j) \geq 3$, the following inequality holds:*

$$\left((I + \ell \cdot L)^{-1}\right)_{ij} \geq \frac{d_{\mathcal{G}}(i, j)}{d^2(2d\ell + 1)} \cdot \left(\frac{d\ell}{2d\ell + 1}\right)^{d_{\mathcal{G}}(i,j)}.$$

If we make the additional assumption that $\ell > \frac{16}{d}$, we have that

$$\left((I + \ell \cdot L)^{-1}\right)_{ij} \geq \frac{1}{4\sqrt{\pi \cdot d_{\mathcal{G}}(i, j) \cdot \sqrt{d\ell}} \cdot d^2(2d\ell + 1)} \cdot \left(1 - 4(d\ell)^{-\frac{1}{2}}\right)^{d_{\mathcal{G}}(i,j)}.$$

We defer the proof of Lemma 3.D.8 to the end of this section. Note that Lemma 3.D.8 implies that there exists a graph $\mathcal{G}$ that satisfies $\left((I + \ell \cdot L)^{-1}\right)_{i,j} = \Omega\left(\lambda_S^r\right)$, where

$\Omega(\cdot)$ notation hides factors that depend polynomially on $1/\mu, \ell_T, \ell_S$, and Δ, and λ_S is as defined in (3.98). We assume the nodes are located in this graph $\mathcal{G}$ for the rest of the proof.

Using Lemma 3.D.8, we can derive the following lower bound of the variance term in (3.99b):

$$\begin{aligned}\mathbb{E}_w\left|\left((I+\ell\cdot L)^{-1}\phi_{N^r_{-i}}(w)\right)_i\right|^2 &= \mathbb{E}_w\left(\sum_{j\in N^r_{-i}}\left((I+\ell\cdot L)^{-1}\right)_{ij}w_j\right)^2\\ &= \sum_{j\in N^r_{-i}}\left((I+\ell\cdot L)^{-1}\right)^2_{ij}\mathbf{Var}\left[w_j\right]\\ &\geq \sum_{j\in\partial N^{r+1}_i}\left((I+\ell\cdot L)^{-1}\right)^2_{ij}\mathbf{Var}\left[w_j\right]\\ &\geq \Theta\left(\lambda_S^r\cdot\mathbf{Var}\left[w_i\right]\right). \end{aligned} \tag{3.100}$$

Substituting (3.100) into (3.99) and summing over all vertices i, we obtain that

$$\mathbb{E}_w\|x(w)-x^*(w)\|^2 \geq \sum_{i=1}^n \mathbb{E}_w\left|x_i(w)-x_i^*(w)\right|^2 \geq \Theta\left(n\cdot\lambda_S^r\cdot\mathbf{Var}\left[w_i\right]\right).$$

We also see that

$$\mathbb{E}_w\left[\mathrm{cost}_p(\mathsf{OPT})\right] = \mathbb{E}_w\left[w^\top(I-(I+\ell\cdot L)^{-1})w\right] = O(n\cdot\mathbf{Var}\left[w_i\right]). \tag{3.101}$$

Note that the global objective function $(x+w)^\top(x+w)+\ell\cdot x^\top Lx$ is 1-strongly convex, and $x^*(w)$ is minimizer of this function. Thus, we have that for any outcome of w,

$$\mathrm{cost}_p(\mathsf{ALG}) - \mathrm{cost}_p(\mathsf{OPT}) \geq \frac{1}{2}\|x(w)-x^*(w)\|^2.$$

Taking expectations on both sides w.r.t. w gives that

$$\mathbb{E}_w\mathrm{cost}_p(\mathsf{ALG}) - \mathbb{E}_w\mathrm{cost}_p(\mathsf{OPT}) \geq \frac{1}{2}\mathbb{E}_w\|x(w)-x^*(w)\|^2 \geq \Theta\left(n\cdot\lambda_S^r\cdot\mathbf{Var}\left[w_i\right]\right). \tag{3.102}$$

Dividing (3.102) by (3.101), we obtain that

$$\frac{\mathbb{E}_w\mathrm{cost}_p(\mathsf{ALG})}{\mathbb{E}_w\mathrm{cost}_p(\mathsf{OPT})} \geq 1+\Omega\left(\lambda_S^r\right).$$

Note that $\mathbb{P}_w\left[\mathrm{cost}_p(\mathsf{OPT})=0\right]=0$. Thus, there must exist an instance of w such that $\mathrm{cost}_p(\mathsf{OPT})>0$ and

$$\frac{\mathrm{cost}_p(\mathsf{ALG})}{\mathrm{cost}_p(\mathsf{OPT})} \geq 1+\Omega\left(\lambda_S^r\right).$$

□

Before we present the proof of Lemma 3.D.8, we first need to introduce two technical lemmas that will be used in the proof of Lemma 3.D.8. The first lemma (Lemma 3.D.9) provides a lower bound for binomial coefficient $\binom{(2+\epsilon)m}{m}$.

Lemma 3.D.9. *For any positive integer m and $\epsilon \in \mathbb{R}_{\geq 0}$ such that ϵm is an integer, the following inequality holds:*

$$\binom{(2+\epsilon)m}{m} > \frac{1}{\sqrt{2\pi}} m^{-\frac{1}{2}} \cdot \frac{(2+\epsilon)^{(2+\epsilon)m+\frac{1}{2}}}{(1+\epsilon)^{(1+\epsilon)m+\frac{1}{2}}} \cdot e^{-\frac{1}{6m}}.$$

Proof of Lemma 3.D.9. By Lemma 2.1 in Stanica, 2001, we know for any $n \in \mathbb{Z}_+$,

$$n! = \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+r(n)},$$

where $r(n)$ satisfies $\frac{1}{12n+1} < r(n) < \frac{1}{12n}$. Thus we see that

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n+1}} < n! < \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n+\frac{1}{12n}}, \forall n \in \mathbb{Z}_+.$$

Therefore, we can lower bound $\binom{(2+\epsilon)m}{m}$ by

$$\begin{aligned}
\binom{(2+\epsilon)m}{m} &= \frac{((2+\epsilon)m)!}{m! \cdot ((1+\epsilon)m)!} \\
&> \frac{\sqrt{2\pi}((2+\epsilon)m)^{(2+\epsilon)m+\frac{1}{2}} e^{-(2+\epsilon)m+\frac{1}{12(2+\epsilon)m+1}}}{\sqrt{2\pi} m^{m+\frac{1}{2}} e^{-m+\frac{1}{12m}} \cdot \sqrt{2\pi}((1+\epsilon)m)^{(1+\epsilon)m+\frac{1}{2}} e^{-(1+\epsilon)m+\frac{1}{12(1+\epsilon)m}}} \\
&= \frac{1}{\sqrt{2\pi}} m^{-\frac{1}{2}} \cdot \frac{(2+\epsilon)^{(2+\epsilon)m+\frac{1}{2}}}{(1+\epsilon)^{(1+\epsilon)m+\frac{1}{2}}} \cdot e^{\frac{1}{12(2+\epsilon)m+1} - \frac{1}{12m} - \frac{1}{12(1+\epsilon)m}} \\
&> \frac{1}{\sqrt{2\pi}} m^{-\frac{1}{2}} \cdot \frac{(2+\epsilon)^{(2+\epsilon)m+\frac{1}{2}}}{(1+\epsilon)^{(1+\epsilon)m+\frac{1}{2}}} \cdot e^{-\frac{1}{6m}}.
\end{aligned}$$

□

The second technical lemma (Lemma 3.D.10) will be used to simplify the decay factor in the proof of Lemma 3.D.8.

Lemma 3.D.10. *For all $\epsilon \in [0, \sqrt{2})$, the following inequality holds*

$$\frac{2+\epsilon}{2 \cdot (1+\epsilon)^{\frac{1+\epsilon}{2+\epsilon}}} \geq 1 - \frac{\epsilon^2}{2}.$$

Proof of Lemma 3.D.10. By taking logarithm on both sides, we see the original inequality is equivalent to

$$\ln\left(1 + \frac{\epsilon}{2}\right) - \frac{1+\epsilon}{2+\epsilon} \ln(1+\epsilon) \geq \ln\left(1 - \frac{1}{2}\epsilon^2\right), \text{ thus}$$

$$\ln\left(1+\frac{\epsilon}{2}\right)-\frac{1+\epsilon}{2+\epsilon}\ln(1+\epsilon)-\ln\left(1-\frac{1}{2}\epsilon^2\right)\geq 0. \tag{3.103}$$

Note that the LHS can be lower bounded by

$$\begin{aligned}&\ln\left(1+\frac{\epsilon}{2}\right)-\frac{1+\epsilon}{2+\epsilon}\ln(1+\epsilon)-\ln\left(1-\frac{1}{2}\epsilon^2\right)\\ \geq\ &\ln\left(1+\frac{\epsilon}{2}\right)-\frac{1+\epsilon}{2}\ln(1+\epsilon)-\ln\left(1-\frac{1}{2}\epsilon^2\right)=: g(\epsilon).\end{aligned}$$

Function g satisfies that $g(0)=0$, and its derivative is

$$\begin{aligned}g'(\epsilon)&=\frac{1}{2+\epsilon}-\frac{1}{2}-\frac{1}{2}\ln(1+\epsilon)+\frac{\epsilon}{1-\frac{1}{2}\epsilon^2}\\ &\geq\frac{1}{2+\epsilon}-\frac{1}{2}-\frac{\epsilon}{2}+\epsilon\\ &=\frac{2-(2+\epsilon)(1-\epsilon)}{2(2+\epsilon)}\\ &=\frac{\epsilon+\epsilon^2}{2(2+\epsilon)}\geq 0.\end{aligned}$$

Thus, $g(\epsilon)\geq 0$ for all $\epsilon\in[0,\sqrt{2})$. Hence (3.103) holds for all $\epsilon\in[0,\sqrt{2})$. □

Now we are ready to present the proof of Lemma 3.D.8.

Proof of Lemma 3.D.8. Consider the graph $\mathcal{G}$ constructed as Figure 3.11: Let N be a positive integer that is sufficiently large. N blocks form a ring, where each block contains d nodes. Every pair of blocks are connected by a complete bipartite graph. The graph Laplacian of $\mathcal{G}$ can be decomposed as $L=2dI-M$, where M is the adjacency matrix of $\mathcal{G}$. We see that

$$\begin{aligned}(I+\ell\cdot L)^{-1}&=((2d\ell+1)I-\ell\cdot M)^{-1}\\ &=\frac{1}{2d\ell+1}\left(I-\frac{\ell}{2d\ell+1}M\right)^{-1}\\ &=\frac{1}{2d\ell+1}\sum_{t=0}^{\infty}\frac{\ell^t}{(2d\ell+1)^t}M^t.\end{aligned}$$

Fix two nodes i and j and denote $\kappa := d_{\mathcal{G}}(i,j)$ and assume $\kappa\geq 3$. Without the loss of generality, we can assume j is on the clockwise direction of i. We see that

$$\left((I+\ell\cdot L)^{-1}\right)_{ij}=\frac{1}{2d\ell+1}\sum_{t=0}^{\infty}\frac{\ell^t}{(2d\ell+1)^t}(M^t)_{ij}$$

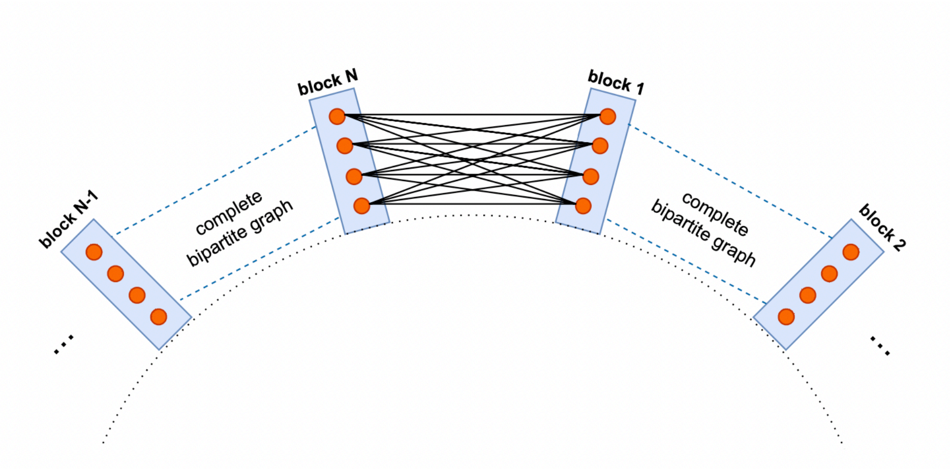

Figure 3.11: Graph structure of $\mathcal{G}$ to obtain the lower bound: N blocks form a ring. Each block contains d nodes.

$$= \frac{\ell^\kappa}{(2d\ell+1)^{\kappa+1}} \sum_{m=0}^{\infty} \frac{\ell^{2m}}{(2d\ell+1)^{2m}} (M^{\kappa+2m})_{ij}. \tag{3.104}$$

Note that $(M^{\kappa+2m})_{ij}$ denotes the number of paths from i to j with length $\kappa+2m$ in graph $\mathcal{G}$. Note that the shortest paths from i to j have length κ. To pick a path with length $(\kappa+2m)$ from i to j, we can first pick a path on the level of blocks: The number of possible block-level paths is lower bounded by $\binom{\kappa+2m}{m}$ because we can choose m in $(\kappa+2m)$ steps to go in the counter-clockwise direction. After a block-level path is fixed, we can choose which specific nodes in the blocks we want to land at, and there are $d^{\kappa+2m-2}$ choices. Thus we see that

$$(M^{\kappa+2m})_{ij} \geq \binom{\kappa+2m}{m} d^{\kappa+2m-2}.$$

Substituting this into (3.104) gives

$$\left((I+\ell\cdot L)^{-1}\right)_{ij} \geq \frac{\ell^\kappa d^{\kappa-2}}{(2d\ell+1)^{\kappa+1}} \sum_{m=0}^{\infty} \frac{\ell^{2m} d^{2m}}{(2d\ell+1)^{2m}} \binom{\kappa+2m}{m}. \tag{3.105}$$

Let $m=0$ will give that $\left((I+\ell\cdot L)^{-1}\right)_{ij} \geq \frac{\kappa}{d^2(2d\ell+1)} \cdot \left(\frac{d\ell}{2d\ell+1}\right)^\kappa$, which shows the first claim of Lemma 3.D.8. Now we proceed to show the second claim of Lemma 3.D.8.

By Lemma 3.D.9, we know that when $\kappa = \epsilon m$, we have that the following inequality holds:

$$\begin{aligned}
\binom{\kappa + 2m}{m} &= \binom{(2+\epsilon)m}{m} \\
&> \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{6m}} m^{-\frac{1}{2}} \frac{(2+\epsilon)^{(2+\epsilon)m}}{(1+\epsilon)^{(1+\epsilon)m}} \cdot \left(\frac{2+\epsilon}{1+\epsilon}\right)^{\frac{1}{2}} \\
&\geq \frac{1}{2\sqrt{2\pi}} m^{-\frac{1}{2}} \left(\frac{2+\epsilon}{(1+\epsilon)^{\frac{1+\epsilon}{2+\epsilon}}}\right)^{(2+\epsilon)m} .
\end{aligned}$$

For any $m > \kappa$, the inequality we just showed can help us bound a term in the summation of (3.105) below:

$$\begin{aligned}
&\frac{\ell^{\kappa} d^{\kappa-2}}{(2d\ell+1)^{\kappa+1}} \cdot \frac{\ell^{2m} d^{2m}}{(2d\ell+1)^{2m}} \binom{\kappa+2m}{m} \\
&\geq \frac{1}{d^2(2d\ell+1)} \cdot \frac{1}{2^{\kappa+2m}} \binom{\kappa+2m}{m} \cdot \left(1 - \frac{1}{2d\ell+1}\right)^{\kappa+2m} \\
&\geq \frac{1}{2\sqrt{2\pi} \cdot \sqrt{\frac{\kappa}{\epsilon}} \cdot d^2(2d\ell+1)} \cdot \left(\frac{2+\epsilon}{2 \cdot (1+\epsilon)^{\frac{1+\epsilon}{2+\epsilon}}}\right)^{(2+\epsilon)\kappa/\epsilon} \cdot \left(1 - \frac{1}{2d\ell+1}\right)^{(1+\frac{2}{\epsilon})\kappa} \\
&\geq \frac{1}{2\sqrt{2\pi} \cdot \sqrt{\frac{\kappa}{\epsilon}} \cdot d^2(2d\ell+1)} \cdot \left(\left(1 - \frac{\epsilon^2}{2}\right)^{\frac{1}{\epsilon}} \cdot \left(1 - \frac{1}{2d\ell+1}\right)^{\frac{1}{\epsilon}}\right)^{(2+\epsilon)\kappa} ,
\end{aligned}$$

where the last line follows from Lemma 3.D.10.

Thus, we obtain that the following inequality holds for arbitrary $\epsilon \in (0,1)$:

$$\left((I+L)^{-1}\right)_{ij} \geq \frac{1}{2\sqrt{2\pi} \cdot \sqrt{\frac{\kappa}{\epsilon}} \cdot d^2(2d\ell+1)} \cdot \left(\left(1 - \frac{\epsilon^2}{2}\right)^{\frac{2}{\epsilon}+1} \cdot \left(1 - \frac{1}{2d\ell+1}\right)^{\frac{2}{\epsilon}+1}\right)^{\kappa} . \tag{3.106}$$

By setting ϵ such that $1/\epsilon = \left\lceil 2(d\ell)^{\frac{1}{2}} \right\rceil$ in (3.106), we obtain that:

$$\begin{aligned}
&\left((I + \ell \cdot L)^{-1}\right)_{ij} \\
&\geq \frac{1}{2\sqrt{2\pi} \cdot \sqrt{2\kappa \cdot \sqrt{d\ell}} \cdot d^2(2d\ell+1)} \cdot \left(\left(1 - \frac{1}{2d\ell}\right)^{4\sqrt{d\ell}+1} \cdot \left(1 - \frac{1}{2d\ell+1}\right)^{4\sqrt{d\ell}+1}\right)^{\kappa} \\
&\geq \frac{1}{4\sqrt{\pi \cdot \kappa \cdot \sqrt{d\ell}} \cdot d^2(2d\ell+1)} \cdot \left(\left(1 - \frac{4\sqrt{d\ell}+1}{2d\ell}\right) \cdot \left(1 - \frac{4\sqrt{d\ell}+1}{2d\ell+1}\right)\right)^{\kappa} \\
&\geq \frac{1}{4\sqrt{\pi \cdot \kappa \cdot \sqrt{d\ell}} \cdot d^2(2d\ell+1)} \cdot \left(1 - \frac{4}{\sqrt{d\ell}}\right)^{\kappa} .
\end{aligned} \tag{3.107}$$

□

Step 3: Combine Temporal and Spatial Lower Bounds Combining the results of Steps 1 and 2 together, we know that the competitive ratio of any decentralized online algorithm is lower bounded by

$$\max\{1 + \frac{\mu^3(1-\sqrt{\lambda_T})^2}{48(\mu+1)^2\ell_T} \cdot \lambda_T^k, 1 + \Omega(\lambda_S^r)\} = 1 + \Omega(\lambda^k) + \Omega(\lambda_S^r).$$

Proof of Corollary 3.3.7

In this appendix we prove a resource augmentation bound for LPC. To simplify the notation, we define the shorthand $a_T := \ell_T/\mu$ and $a_S := \ell_S/\mu$. a_T and a_S are positive real numbers. We first show two lemmas about the relationships between the decay factors ρ_T and λ_T, and ρ_S and λ_S.

Lemma 3.D.11. *Under the assumptions of Theorem 3.3.3, we have $\rho_T^4 \le \lambda_T \le \rho_T^2$.*

Proof of Lemma 3.D.11. Recall that ρ_T is given by

$$\rho_T = \sqrt{1 - \frac{2}{\sqrt{1+2a_T}+1}}$$

in Theorem 3.3.3. Thus we see that

$$\rho_T^4 = \left(1 - \frac{2}{\sqrt{1+2a_T}+1}\right)^2 \le \left(1 - \frac{2}{\sqrt{1+4a_T}+1}\right)^2 = \lambda_T.$$

On the other hand, we have that

$$\begin{aligned}\lambda_T - \rho_T^2 &= \left(1 - \frac{2}{\sqrt{1+4a_T}+1}\right)^2 - 1 + \frac{2}{\sqrt{1+2a_T}+1}\\ &= \frac{4\sqrt{(1+2a_T)}\left(\sqrt{1+2a_T} - \sqrt{1+4a_T}\right)}{\left(\sqrt{1+2a_T}+1\right)\left(\sqrt{1+4a_T}+1\right)^2} \le 0.\end{aligned}$$

□

Lemma 3.D.12. *Under the assumptions of Theorem 3.3.3, we have $\rho_S^{8\Delta \log \Delta} \le \lambda_S$.*

Proof of Lemma 3.D.12. Recall that ρ_T is given by

$$\rho_S = \sqrt{1 - \frac{2}{\sqrt{1+\Delta a_S}+1}}$$

in Theorem 3.3.3. We consider the following 3 cases separately.

Case 1: $a_S \geq 16\Delta - 8$.

We first show that the following inequality holds for any $x \in [0, 1/(2\Delta)]$:

$$(1-x)^{2\Delta} \leq 1 - \Delta x. \tag{3.108}$$

To see this, define function $g(x) = (1-x)^{2\Delta} + \Delta x - 1$. Note that g is a convex function with $g(0) = 0$ and

$$g\left(\frac{1}{2\Delta}\right) = \left(1 - \frac{1}{2\Delta}\right)^{2\Delta} - \frac{1}{2} \leq e^{-1} - \frac{1}{2} < 0.$$

Thus, we see that $g(x) \leq 0$ holds for all $x \in [0, 1/(2\Delta)]$. Hence (3.108) holds.

Note that under the assumption $a_S \geq 16\Delta - 8$, we have

$$0 \leq \frac{2}{\sqrt{1+\Delta a_S}+1} \leq \frac{1}{2\Delta}.$$

Thus substituting $x = \frac{2}{\sqrt{1+\Delta a_S}+1} \leq \frac{1}{2\Delta}$ gives

$$\rho_S^{4\Delta} = \left(1 - \frac{2}{\sqrt{1+\Delta a_S}+1}\right)^{2\Delta} \leq 1 - \frac{2\Delta}{\sqrt{1+\Delta a_S}+1} \leq 1 - \frac{2}{\sqrt{1+4a_S}+1} = \lambda_S.$$

Case 2: $a_S \leq \frac{\Delta^2}{(\Delta^2-1)^2}$. Recall that $\Delta \geq 2$. Thus, in this case, we have $a_S < 1$ and

$$\sqrt{\lambda_S} = 1 - \frac{2}{\sqrt{1+4a_S}+1} \leq \frac{1}{\Delta^2}.$$

Note that

$$\rho_S^2 = \frac{(\sqrt{1+\Delta a_S}-1)^2}{\Delta a_S} \leq \frac{\Delta a_S}{4} = \Delta \cdot \frac{(\sqrt{1+2a_S+a_S^2}-1)^2}{4a_S} \leq \Delta \cdot \frac{(\sqrt{1+4a_S}-1)^2}{4a_S}$$
$$= \Delta\sqrt{\lambda_S}.$$

Thus we see that

$$\rho_S^4 \leq (\Delta^2 \cdot \lambda_S) \cdot \lambda_S \leq \lambda_S.$$

Case 3: $\frac{\Delta^2}{(\Delta^2-1)^2} < a_S < 16\Delta - 8$.

In this case, we have

$$\lambda_S = \left(1 - \frac{2}{\sqrt{1+4a_S}+1}\right)^2 \geq \frac{1}{\Delta^4},$$
$$\rho_S = \left(1 - \frac{2}{\sqrt{1+\Delta a_S}+1}\right)^{\frac{1}{2}} \leq \sqrt{1 - \frac{1}{2\Delta}}.$$

Since $(1 - 1/(2\Delta))^{2\Delta} < e^{-1}$, we see that $\rho_S^{8\Delta\log(\Delta)} \leq \lambda_S$. □

Now we come back to the proof of Corollary 3.3.7. By Theorem 3.3.5 and Theorem 3.3.6, we know that the optimal competitive ratio is lower bounded by

$$c(k^*, r^*) \geq 1 + C_\lambda \left(\lambda_T^{k^*} + \lambda_S^{r^*}\right)$$

and LPC's competitive ratio is upper bounded by

$$\mathrm{CR}_{\mathcal{P}(v)}(\mathrm{LPC}) \coloneqq 1 + C_\rho \left(C_3(r)^2 \cdot \rho_T^k + h(r)^2 \cdot \rho_S^r\right),$$

where C_λ and C_ρ are some positive constants. To achieve $c_{LPC}(k, r) \leq c(k^*, r^*)$, it suffices to guarantee that

$$C_\rho \cdot C_3(r)^2 \cdot \rho_T^k \leq C_\lambda \lambda_T^{k^*} \text{ and } C_\rho \cdot h(r)^2 \cdot \rho_S^r \leq C_\lambda \lambda_S^{r^*}.$$

Note that $C_3(r)$ can be upper bounded by some constant and $h(r)^2 \leq poly(r) \cdot \rho_S^{-\frac{r}{2}}$ under our assumptions. Applying Lemma 3.D.11 and Lemma 3.D.12 finishes the proof.

Proof of Corollary 3.3.8

By Corollary 3.3.2 and Theorem 3.3.6, we see the competitive ratio of LPC with prediction horizon k and observation radius r will be less than or equal to $c(k^*, r^*)$ if the following two inequalities holds:

$$\left(2 + C_4' \cdot \frac{\ell_f + \Delta \ell_S + 2\ell_T}{\mu}\right) \cdot \left(\frac{8\ell_T}{\mu}\right)^k \leq \frac{\mu^3(1 - \sqrt{\lambda_T})^2}{96(\mu+1)^2 \ell_T} \cdot \lambda_T^{k^*}, \tag{3.109}$$

$$\left(1 + C_3' \cdot \frac{\ell_f + \Delta \ell_S + 2\ell_T}{\mu}\right) \cdot \left(\frac{\Delta^3 \ell_S}{\mu}\right)^{\frac{r}{2}} \leq \frac{\mu^3(1 - \sqrt{\lambda_S})^2}{96(\mu+1)^2 \ell_S} \cdot \lambda_S^{r^*+1}, \tag{3.110}$$

where

$$\lambda_T = \left(1 - 2\left(\sqrt{1 + (4\ell_T/\mu)} + 1\right)^{-1}\right)^2, \text{ and } \lambda_S = \frac{(\Delta \ell_S/\mu)}{3 + 3(\Delta \ell_S/\mu)}.$$

Note that we have

$$\frac{1}{2}\lambda_T^{\frac{1}{2}} = \frac{2\ell_T/\mu}{(\sqrt{1 + 4\ell_T/\mu} + 1)^2} \geq 8(\sqrt{5} - 2)^2 \cdot \frac{\ell_T}{\mu} \geq \left(\frac{8\ell_T}{\mu}\right)^{1 + c_1 \log(\mu/(8\ell_T))}$$

for some positive constant $c_1 < \infty$. We also have that

$$\frac{1}{2}\lambda_S \geq \frac{(\Delta \ell_S/\mu)}{8} \geq \left(\frac{\Delta^3 \ell_S}{\mu}\right)^{1 + c_2 \log\left(\mu/(\Delta^3 \ell_S)\right)}$$

holds for some positive constant $c_2 = c_2(\Delta) < \infty$. Therefore, we see that there exists $C = C(\Delta) < \infty$ such that

$$\left(\frac{8\ell_T}{\mu}\right)^{(2+C\log(\mu/(8\ell_T)))k^*} \leq \left(\frac{1}{2}\right)^{k^*} \cdot \lambda_T^{k^*}, \text{ and } \left(\frac{\Delta^3\ell_S}{\mu}\right)^{\frac{1}{2}\cdot(2+C\log\left(\mu/(\Delta^3\ell_S)\right))r^*} \leq \left(\frac{1}{2}\right)^{r^*} \cdot \lambda_S^{r^*}.$$

Therefore, let $k = (2 + C\log(\mu/(8\ell_T)))k^*$ and $r = (2 + C\log\left(\mu/(\Delta^3\ell_S)\right))r^*$. We know that (3.109) and (3.110) hold when k^* and r^* are sufficiently large.

3.E Proofs for Adaptive Video Streaming

We first introduce the notation used to define the performance metrics and the variant of `SODA` studied in our theoretical analysis. To make the formulation of the video streaming problem closer to a classic control problem, we define the "control action" u_t as the inverse of the bitrate (i.e., $u_t = \frac{1}{r_t}$). Recall that we set $v(r) = \frac{1}{r}$ in our theoretical analysis. Thus, we can write down a general form of the optimization problem solved by `SODA` and use $\psi_t^{t+p}\left((\sigma_{t-1}, \nu_{t-1}); \hat{\omega}_{t:t+p}; F\right)$ to denote its optimal solution:

$$\arg\min_{x_{t:t+p}, u_{t+1:t+p}} \sum_{\tau=t}^{t+p} \hat{\omega}_\tau u_\tau^2 + \beta \sum_{\tau=t}^{t+p} b(x_\tau) + \gamma \sum_{\tau=t}^{t+p+1} |u_\tau - u_{\tau-1}|^2 + F(x_{t+p}, u_{t+p+1}) \tag{3.111a}$$

$$\text{s.t. } x_\tau = x_{\tau-1} + \hat{\omega}_\tau u_\tau - 1, \text{ for } \tau = t, \ldots, t+p, \tag{3.111b}$$

$$0 \leq x_\tau \leq x_{\max}, \frac{1}{r_{\max}} \leq u_\tau \leq \frac{1}{r_{\min}}, \text{ for } \tau = t, \ldots, t+p, \tag{3.111c}$$

$$x_{t-1} = \sigma_{t-1}, u_{t-1} = \nu_{t-1}. \tag{3.111d}$$

Here, $\psi_t^{t+p}\left((\sigma_{t-1}, \nu_{t-1}); \hat{\omega}_{t:t+p}; F\right)$ is defined to be a vector that contains the states $x_{t:t+p}$ and control actions $u_{t+1:t+p}$ in the optimal solution. The initial condition $(\sigma_{t-1}, \nu_{t-1})$, bandwidth sequence $\hat{\omega}_{t:t+p}$, and terminal cost function F are the parameters of the optimization problem. For the terminal costs, we consider two types of functions: (1) The zero function $F = \mathbf{0}$, i.e., $F(x, u) = 0$ for all x, u; (2) The indicator function $F = \mathbb{I}_{\sigma,\nu}$, which is defined as

$$F(x, u) = \mathbb{I}_{\sigma,\nu}(x, u) = \begin{cases} 0 & \text{if } x = \sigma, u = \nu, \\ +\infty & \text{otherwise.} \end{cases}$$

The first type of terminal cost will be used to define the performance metrics (competitive ratio and dynamic regret), and the second type will be used in the algorithm design. Since we will use the indicator terminal cost frequently, we introduce the shorthand $\tilde{\psi}_t^{t+p}\left((\sigma_{t-1}, \nu_{t-1}); \hat{\omega}_{t:t+p}; (\sigma_{t+p}, \nu_{t+p+1})\right)$, which denotes $\tilde{\psi}_t^{t+p}\left((\sigma_{t-1}, \nu_{t-1}); \hat{\omega}_{t:t+p}; \mathbb{I}_{\sigma_{t+p}, \nu_{t+p+1}}\right)$. We use $\iota_t^{t+p}\left((\sigma_{t-1}, \nu_{t-1}); \hat{\omega}_{t:t+p}; F\right)$ to denote the optimal objective value of the optimization problem (3.111).

Proof of Theorem 3.4.1

In this section, we establish the critical exponentially decaying perturbation bounds (Definition 3.4.1). Instead of just focusing on the video streaming application itself, we establish the perturbation bound for a more general SOCO with memory framework.

Specifically, we consider the following finite-time optimal control problem with memory H.

$$\psi(y, z; \mu, w, \delta) = \underset{x_{-H+1:p+H-1}}{\arg\min} \sum_{t=0}^{p} f_t(x_t; \mu_t) + \sum_{t=0}^{p+H-1} c_t(x_{t:t-H+1}; w_t) \tag{3.112a}$$

$$\text{s.t. } x_t \in [0, x_{\max}] \subseteq \mathbb{R}, \forall 0 \le t \le p, \tag{3.112b}$$

$$x_t - x_{t-1} \ge -\delta_t, \forall 0 \le t \le p+1, \tag{3.112c}$$

$$x_{-H+1:-1} = y, x_{p+1:p+H-1} = z, \tag{3.112d}$$

where $y, z \in [0, x_{\max}]^{H-1}$, $\mu \in [0, x_{\max}]^{p+1}, w \in \mathcal{W}^{p+H}, \delta \in \Delta^{p+2}$. Here, the objective function (3.112a) contains the hitting costs $f_t(x_t; \mu_t)$ (parameterized by μ_t) and the switching costs $c_t(x_{t:t-H+1}; w_t)$ (parameterized by w_t). For the constraints, (3.112b) imposes a box constraint on each decision variable x_t; (3.112c) imposes a constraint on how much x_t can decrease at each time step; and (3.112d) specifies the boundary conditions of the optimization problem.

In the special case of video streaming, the decision is on the buffer level x_t. Given the buffer levels, the inverse of the bitrate $u_t \coloneqq 1/r_t$ is uniquely decided by the equation

$$u_t = (x_t - x_{t-1} + 1)/\omega_t,$$

where ω_t denotes the bandwidth. The memory length $H = 3$. For the hitting cost, we have $\mu_t \equiv \bar{x}$, and

$$f_t(x; \mu_t) = \beta b(x) = \begin{cases} \beta(x - \bar{x})^2, & \text{if } x \le \bar{x}, \\ \epsilon\beta(x - \bar{x})^2, & \text{otherwise.} \end{cases}$$

For the switching cost, we have $w_t = (\omega_t, \omega_{t-1})$ and

$$c_t(x_{t:t-2}; w_t)$$

$$= \omega_t u_t^2 + \gamma(u_t - u_{t-1})^2$$
$$= \frac{(x_t - x_{t-1} + 1)^2}{\omega_t} + \gamma \frac{(\omega_{t-1}x_t + \omega_t x_{t-2} - (\omega_t + \omega_{t-1})x_{t-1} + (\omega_{t-1} - \omega_t))^2}{\omega_t^2 \omega_{t-1}^2}.$$

The first constraint $x_t \in [0, x_{\max}]$ of (3.112) matches the buffer constraint of the video streaming problem exactly.

The second constraint $x_t - x_{t-1} \geq -\delta_t$ corresponds to the constraint that $u_t \geq \frac{1}{r_{\max}}$ in (3.111). Thus, when applying (3.112) to video streaming, we have $\delta_t = 1 - \frac{\omega_t}{r_{\max}}$. By Assumption 3.4.1, we have $\delta_t \geq \delta > 0$.

Given the relationship between SOCO with memory problem and adaptive video streaming problem, we only need to establish the exponentially decaying perturbation bound for the more general SOCO with memory problem. To show this perturbation bound, we need the following assumption about the objective function and constraints:

Assumption 3.5.1. *We need the following assumption on the optimization problem* (3.112) *for the exponentially decaying perturbation property to hold:*

1. $f_t(\cdot;\mu_t) : \mathbb{R} \to \mathbb{R}$ *is strongly convex for all* t *and* $\mu_t \in [0, x_{max}]$. *We further assume there exists two* m_f*-strongly convex and* ℓ_f*-smooth functions* $f_t^{(0)}(\cdot;\mu_t), f_t^{(1)}(\cdot;\mu_t) : \mathbb{R} \to \mathbb{R}$ *in* C^2 *such that* $f_t(x_t;\mu_t) = f_t^{(0)}(x_t;\mu_t)$ *for* $x_t \in [0, \mu_t]$ *and* $f_t(x_t) = f_t^{(1)}(x_t;\mu_t)$ *for* $x_t \in [\mu_t, x_{max}]$. *We also assume that for* $j = 1, 2$, $f_t^{(j)}$ *satisfies that for all* $x_t, \mu_t \in [0, x_{max}]$,
$$\left\|\nabla_{x_t} f_t^{(j)}(x_t;\mu_t)\right\| + \left\|\nabla_{\mu_t} f_t^{(j)}(x_t;\mu_t)\right\| \leq L_f, \text{ and } \left\|\nabla_{\mu_t}\nabla_{x_t} f_t^{(j)}(x_t;\mu_t)\right\| \leq \ell_\mu.$$

2. $c_t(\cdot; w_t) : \mathbb{R}^H \to \mathbb{R}$ *is convex and* ℓ_c*-smooth for all* t *and* $w_t \in \mathcal{W} \subset \mathbb{R}^q$. $c_t(\cdot; w_t)$ *is in* C^2 *on* $[0, x_{max}]^H$. *We also assume that for all* $w_t \in \mathcal{W}$ *and feasible* $x_{t:t-H+1}$, *we have*
$$\left\|\nabla_{x_{t:t-H+1}} c_t(x_{t:t-H+1}; w_t)\right\| + \left\|\nabla_{w_t} c_t(x_{t:t-H+1}; w_t)\right\| \leq L_c, \text{ and}$$
$$\left\|\nabla_{w_t}\nabla_{x_{t:t-H+1}} c_t(x_{t:t-H+1}; w_t)\right\| \leq \ell_w.$$

3. *We have* $\delta_t \in \Delta$ *holds for all* t, *where* Δ *is a closed interval on* $\mathbb{R}$ *and is bounded below by some positive constant* δ. *Denote* $d := \lceil x_{max}/\delta \rceil$.

In the special case of the video streaming problem, Assumption 3.5.1 is satisfied with the parameters $m_f = \epsilon\beta$, $\ell_f = \ell_\mu = \beta$, $\ell_c = \frac{2(\omega_{\min}+3)}{\omega_{\min}^2}$, $\ell_w = \frac{4x_{\max}(\omega_{\min}+8\gamma)}{\omega_{\min}^3}$. In addition, both L_f and L_c are bounded.

We state the exponentially decaying perturbation bound for the SOCO with memory problem formally in Theorem 3.5.1 and defer its proof to Appendix 3.4.

Theorem 3.5.1. *Under Assumption 3.5.1, if $p \geq d$, the inequality*

$$\begin{aligned}
&\|\psi(y, z; \mu, w, \delta)_t - \psi(y', z'; \mu', w', \delta')_t\| \\
&\leq C\left(\rho^t \|y - y'\| + \rho^{p-t}\|z - z'\|\right) \\
&\quad + C\left(\sum_{\tau=0}^{p} \rho^{|t-\tau|}\left|\mu_\tau - \mu'_\tau\right| + \sum_{\tau=0}^{p+H-1} \rho^{|t-\tau|}\left\|w_\tau - w'_\tau\right\| + \sum_{\tau=0}^{p+1} \rho^{|t-\tau|}\left\|\delta_\tau - \delta'_\tau\right\|\right)
\end{aligned} \tag{3.113}$$

holds for all $t \in [0, p]$ and $y, z \in [\underline{x}, \overline{x}]^{H-1}$. Here,

$$\rho = \left(1 - \frac{2}{1 + \sqrt{1 + (\underline{\ell}/m_f)}}\right)^{\frac{1}{H(H+d)}}, C = \frac{2\overline{\ell}}{m_f \rho^{(H-2)(H+d)}},$$

where $\underline{\ell} \coloneqq \max\{H\ell_c, \ell_w\}$ and $\bar{\ell} \coloneqq \max\{H\ell_f, \ell_\mu, \underline{\ell}\}$.

In the special case of the video streaming, we see that

$$\underline{\ell} = \max\{3\ell_c, \ell_w\} = \frac{\max\{6\omega_{\min}(\omega_{\min} + 3), 4x_{\max}(\omega_{\min} + 8\gamma)\}}{\omega_{\min}^3}.$$

Therefore, we have

$$\rho = \left(1 - \frac{2}{1 + \sqrt{1 + \frac{\max\{6\omega_{\min}(\omega_{\min}+3), 4x_{\max}(\omega_{\min}+8\gamma)\}}{\omega_{\min}^3 \epsilon\beta}}}\right)^{\frac{1}{3(3+\lceil x_{\max}/\delta\rceil)}}.$$

The coefficient C is bounded by

$$C \leq \frac{3\beta\omega_{\min}^3 + \max\{6\omega_{\min}(\omega_{\min} + 3), 4x_{\max}(\omega_{\min} + 8\gamma)\}}{\omega_{\min}^3 \rho^{3+\lceil x_{\max}/\delta\rceil}}.$$

Discussion about different distortion costs. Note that Assumption 3.5.1 still holds if we replace the distortion cost function $v(r) = \frac{1}{r}$ by $v(r) = \log(r_{\max}/r)$. This is because the new switching cost

$$\begin{aligned}
c'_t(x_{t:t-2}; w_t) &= \omega_t u_t \log(r_{\max} u_t) + \gamma(u_t - u_{t-1})^2 \\
&= (x_t - x_{t-1} + 1) \log\left(\frac{r_{\max}(x_t - x_{t-1} + 1)}{\omega_t}\right)
\end{aligned}$$

$$+\gamma\frac{(\omega_{t-1}x_t+\omega_t x_{t-2}-(\omega_t+\omega_{t-1})x_{t-1}+(\omega_{t-1}-\omega_t))^2}{\omega_t^2\omega_{t-1}^2}$$

also satisfies Assumption 3.5.1 for any $w_t=(\omega_t,\omega_{t-1})\in[\omega_{\min},\omega_{\max}]^2$ and feasible $x_{t:t-2}$.

Proof of Theorem 3.5.1

To show Theorem 3.5.1, we first need to define *indicators of active constraints*, denoted as $\xi\in\{0,1\}^{4p+5}$. Specifically, given the unique optimal solution $x_{0:p}=\psi(y,z;\mu,w,\delta)$ under a tuple of parameters $(y,z;\mu,w,\delta)$, we consider whether the following equality conditions hold:

$$\begin{aligned}
\xi_{1,t}&=\mathbf{1}\{x_t=0\},\forall 0\le t\le p;\\
\xi_{2,t}&=\mathbf{1}\{x_t=x_{\max}\},\forall 0\le t\le p;\\
\xi_{3,t}&=\mathbf{1}\{x_t=\mu_t\},\forall 0\le t\le p;\\
\xi_{4,t}&=\mathbf{1}\{x_t-x_{t-1}=-\delta_t\},\forall 0\le t\le p+1.
\end{aligned}$$

And we define *indicators of the sides* (denoted as $\sigma\in\{0,1\}^{p+1}$) as the following:

$$\sigma_t=\mathbf{1}\{x_t\in[\mu_t,x_{\max}]\},\forall 0\le t\le p.$$

To simplify the notation, we let $\theta:=(\mu,w,\delta)\in\Theta:=[0,x_{\max}]^{p+1}\times\mathcal{W}^{p+H}\times\Delta^{p+2}$. While $\psi(y,z;\theta)$ can decide a unique pair of (ξ,σ), we can also define a new equality-constrained optimization problem using $(y,z;\theta)$ and (ξ,σ):

Definition 3.5.1. *Define the equality-constrained optimization problem*

$$\hat{\psi}(y,z;\theta;\xi,\sigma)=\operatorname*{arg\,min}_{x_{-H+1:p+H-1}}\sum_{t=0}^{p}f_t^{(\sigma_t)}(x_t;\mu_t)+\sum_{t=0}^{p+H-1}c_t(x_{t:t-H+1};w_t)\tag{3.114a}$$

$$\textit{s.t. }x_t=\begin{cases}0, & \textit{if }\xi_{1,t}=1\\ x_{max}, & \textit{if }\xi_{2,t}=1\\ \mu_t, & \textit{if }\xi_{3,t}=1\end{cases},\forall 0\le t\le p,\tag{3.114b}$$

$$x_t-x_{t-1}=-\delta_t,\ \textit{if }\xi_{4,t}=1,\forall 0\le t\le p+1,\tag{3.114c}$$

$$x_{-H+1:-1}=y,x_{p+1:p+H-1}=z.\tag{3.114d}$$

Note that it is possible that the optimization problem $\hat{\psi}(y,z;\theta;\xi,\sigma)$ for some parameters and constraint configurations. We use $\hat{\iota}(y,z;\theta;\xi,\sigma)$ to denote the optimal value of this optimization problem. The following lemma states that the optimal solution of (3.112) will not change if we remove all inactive inequality constraints and leave active constraints as equality constraints.

Lemma 3.5.2. *Suppose Assumption 3.5.1 holds and $p \geq d$. For $y, z \in [0, x_{max}]^{H-1}$ and $\theta \in \Theta$, let ξ, σ be the corresponding indicators of active constraints/sides. Then, we have*

$$\psi(y, z; \theta) = \hat{\psi}(y, z; \theta; \xi, \sigma) \text{ and } \iota(y, z; \theta) = \hat{\iota}(y, z; \theta; \xi, \sigma).$$

Proof of Lemma 3.5.2. Note that

$$\iota\left(y, z; \theta\right) \geq \hat{\iota}\left(y, z; \theta; \xi, \sigma\right)$$

because the optimization problem on the RHS has less constraints. If the inequality holds with equality, we must have $\psi(y, z; \theta) = \hat{\psi}(y, z; \theta; \xi, \sigma)$ since the optimal solution for the LHS is feasible for the RHS by the assumption on active constraints, and the optimization problem on the RHS has a unique solution. Otherwise, we must have

$$\psi\left(y, z; \theta\right) \neq \hat{\psi}\left(y, z; \theta; \xi, \sigma\right), \text{ and } \iota\left(y, z; \theta\right) > \hat{\iota}\left(y, z; \theta; \xi, \sigma\right).$$

Consider the convex combination $\zeta(\eta)$ for $\eta \in [0, 1]$ defined as

$$\zeta(\eta) = (1 - \eta)\psi\left(y, z; \theta\right) + \eta\hat{\psi}\left(y, z; \theta; \xi, \sigma\right).$$

Note that $\zeta(\eta)$ satisfies all the active constraints and sides as specified by (ξ, σ) because they are active for all $\eta \in [0, 1]$. Since the constraints of (3.112) that are not in (ξ, σ) are inactive at $\eta = 0$, there must exist $\eta > 0$ such that $\zeta(\eta)$ is also feasible for (3.112). $\zeta(\eta)$ achieves a strictly smaller objective than $\zeta(0) = \psi\left(y, z; \theta\right)$, which leads to a contradiction. □

Lemma 3.5.2 establishes that given any feasible tuple of $(y, z; \theta)$, one can find at least one pair of (ξ, σ) such that $\psi(y, z; \theta) = \hat{\psi}(y, z; \theta; \xi, \sigma)$, while there can be other (ξ', σ') that satisfies $\psi(y, z; \theta) = \hat{\psi}(y, z; \theta; \xi', \sigma')$.

Lemma 3.5.3. *Suppose Assumption 3.5.1 holds and $p \geq d$. If both $\hat{\psi}(y, z; \theta; \xi, \sigma)$ and $\hat{\psi}(y', z'; \theta'; \xi, \sigma)$ exist for $y, z, y', z' \in [0, x_{max}]^{H-1}$ and (ξ, σ), then we have*

$$\begin{aligned}&\left\|\hat{\psi}(y, z; \theta; \xi, \sigma)_t - \hat{\psi}(y', z'; \theta'; \xi, \sigma)_t\right\| \\ \leq\ & C\left(\rho^t\|y - y'\| + \rho^{p-t}\|z - z'\|\right)\end{aligned}$$

$$+C\left(\sum_{\tau=0}^{p}\rho^{|t-\tau|}\left|\mu_\tau-\mu'_\tau\right|+\sum_{\tau=0}^{p+H-1}\rho^{|t-\tau|}\left\|w_\tau-w'_\tau\right\|+\sum_{\tau=0}^{p+1}\rho^{|t-\tau|}\left|\delta_\tau-\delta'_\tau\right|\right), \tag{3.115}$$

where

$$\rho=\left(1-\frac{2}{1+\sqrt{1+(\underline{\ell}/m_f)}}\right)^{\frac{1}{H(H+d)}}, C=\frac{2\bar{\ell}}{m_f\rho^{(H-2)(H+d)}}.$$

Here, $\underline{\ell}\coloneqq\max\{H\ell_c,\ell_w\}$ *and* $\bar{\ell}\coloneqq\max\{H\ell_f,\ell_\mu,\underline{\ell}\}$.

Proof of Lemma 3.5.3. We do a variable change to eliminate all constraints in the equality-constrained optimization problem. After the elimination, we get an unconstrained optimization problem with the free variables $x_{t_0},x_{t_1},\ldots,x_{t_q}$ where the indices satisfy $0\le t_0<t_1<\ldots<t_q\le p$. To simplify the notation, we let $t_{-1}=-1$ and $t_{q+1}=p+1$. For τ that satisfies $t_i<\tau<t_{i+1}$, we have either $x_\tau=x_{t_i}-\sum_{\gamma=t_i+1}^{\tau}\delta_\gamma$ or x_τ is some constant. Without loss of generality, we can assume $t_{i+1}\le t_i+d+H$, because otherwise we can find $\tau\in(t_i,t_{i+1}-H]$ such that $x_{\tau:\tau+H-1}$ are constants, which means the free variables after $x_{t_{i+1}}$ will not change, regardless of how we perturb y, and the free variables before x_{t_i} will not change, regardless of how we perturb z. Thus, we can decompose the perturbation to the left side and the right side and derive them separately.

After the change of variable, the objective becomes a function $\hat{h}$ of $x_{t_0},x_{t_1},\ldots,x_{t_q}$. To simplify the notation, we let $\hat{x}_\tau\coloneqq x_{t_\tau}$, where $\tau=0,\ldots,q$. We can decompose $\hat{h}$ as

$$\hat{h}(\hat{x}_{0:q};\zeta)=\hat{h}_a(\hat{x}_{0:q};\mu)+\hat{h}_b(\hat{x}_{0:q};\zeta),$$

where $\zeta=(y,z,\theta)$, $\hat{h}_a$ is the sum of the original hitting costs minus $\frac{m_f}{2}\left\|\hat{x}_{0:q}\right\|^2$, and $\hat{h}_b$ is the sum of the original switching costs plus $\frac{m_f}{2}\left\|\hat{x}_{0:q}\right\|^2$. By Assumption 3.5.1, we see that

$$\nabla^2_{\hat{x}_{0:q}}\hat{h}_a(\hat{x}_{0:q};\mu)\succeq 0,(m_f+H\ell_c)I\succeq\nabla^2_{\hat{x}_{0:q}}\hat{h}_b(\hat{x}_{0:q};\zeta)\succeq m_fI. \tag{3.116}$$

We also note that $\nabla^2_{\hat{x}_{0:q}}\hat{h}_a(\hat{x}_{0:q};\mu)$ is a diagonal matrix and $\nabla^2_{\hat{x}_{0:q}}\hat{h}_b(\hat{x}_{0:q};\zeta)$ is a $2H$-banded matrix.

We can follow a similar procedure as Theorem 3.1 in Lin, Hu, Shi, et al., 2021 to show

$$\left\|\hat{\psi}(y,z;\theta;\xi,\sigma)_{t_\tau}-\hat{\psi}(y',z';\theta';\xi,\sigma)_{t_\tau}\right\|$$

$$\begin{aligned}
&\leq C_0 \left(\rho_0^{\tau} \|y - y'\| + \rho_0^{q-\tau} \|z - z'\| \right) \\
&\quad + C_0 \left(\sum_{i=0}^{p} \rho_0^{|\phi(i)-\tau|} \left|\mu_i - \mu_i'\right| + \sum_{i=0}^{p+H-1} \rho_0^{|\phi(i)-\tau|} \left\|w_i - w_i'\right\| + \sum_{i=0}^{p+1} \rho_0^{|\phi(i)-\tau|} \left\|\delta_i - \delta_i'\right\| \right),
\end{aligned} \tag{3.117}$$

where $\phi(i)$ denotes the integer j that satisfies $t_j \leq i < t_{j+1}$ and

$$\rho_0 = \left(1 - \frac{2}{\sqrt{1 + (\underline{\ell}/m_f)}}\right)^{\frac{1}{H}}, C_0 = \frac{2\bar{\ell}}{m_f \rho_0^{H-2}}.$$

Here, $\underline{\ell} \coloneqq \max\{H\ell_c, \ell_w\}$ and $\bar{\ell} \coloneqq \max\{H\ell_f, \ell_\mu, \underline{\ell}\}$. For completeness, we give the detailed proof below: Let e be a vector such that both ζ and $\zeta + e$ are in $\mathcal{Y} \times \mathcal{Z} \times \Theta$. Consider the function

$$\overline{\psi}(\zeta + \eta e) \coloneqq \hat{\psi}(\zeta + \eta e; \xi, \sigma)_{t_{0:q}},$$

which is implicitly determined by the equation

$$\nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) = 0.$$

By the implicit function theorem we know that the function $\overline{\psi}$ is differentiable. Taking the derivative with respect to θ gives that

$$\begin{aligned}
&\nabla_{\hat{x}_{0:q}}^2 \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) \frac{d}{d\eta} \overline{\psi}(\zeta + \eta e) \\
= &-\nabla_y \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) e_y - \nabla_z \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) e_z \\
&- \sum_{t=0}^{p} \nabla_{\mu_t} \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) e_{\mu_t} - \sum_{t=0}^{p+H-1} \nabla_{w_t} \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) e_{w_t} \\
&- \sum_{t=0}^{p} \nabla_{\delta_t} \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e) e_{\delta_t}.
\end{aligned}$$

To simplify the notation, we define

$$\begin{aligned}
M &\coloneqq \nabla_{\hat{x}_{0:q}}^2 \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \text{which is a } (q+1) \times (q+1) \text{ matrix}, \\
R^{(y)} &\coloneqq -\nabla_y \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \text{which is a } (q+1) \times (H-1) \text{ matrix}, \\
R^{(z)} &\coloneqq -\nabla_z \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \text{which is a } (q+1) \times (H-1) \text{ matrix}, \\
R^{(\mu_t)} &\coloneqq -\nabla_{\mu_t} \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \text{which is a } (q+1) \times 1 \text{ matrix}, \\
R^{(w_t)} &\coloneqq -\nabla_{w_t} \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \text{which is a } (q+1) \times d \text{ matrix}, \\
R^{(\delta_t)} &\coloneqq -\nabla_{\delta_t} \nabla_{\hat{x}_{0:q}} \hat{h}(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \text{which is a } (q+1) \times 1 \text{ matrix}.
\end{aligned}$$

Hence we can write

$$\frac{d}{d\theta}\overline{\psi}(\zeta + \eta e) = M^{-1}\left(R^{(y)}e_y + R^{(z)}e_z + \sum_{t=0}^{p} R^{(\mu_t)}e_{\mu_t} + \sum_{t=0}^{p+H-1} R^{(w_t)}e_{w_t} + \sum_{t=0}^{p} R^{(\delta_t)}e_{\delta_t}\right).$$

Recall that $R^{(y)}, R^{(z)}$ are $(q+1) \times (H-1)$ matrices. For $R^{(y)}$, only the first $H-1$ rows are non-zero. For $R^{(z)}$, only the last $H-1$ rows are non-zero. Hence we see that

$$\begin{aligned}
&\frac{d}{d\eta}\overline{\psi}(\zeta + \eta e)_\tau \\
&= (M^{-1})_{\tau,0:H-2}R^{(y)}_{0:H-2,:}e_y + (M^{-1})_{\tau,q-H+2:q}R^{(z)}_{q-H+2:q,:}e_z \\
&\quad + \sum_{j=0}^{q}\sum_{i=t_j}^{t_{j+1}-1}(M^{-1})_{\tau,j}R^{(\mu_i)}_{j,:}e_{\mu_i} + \sum_{j=0}^{q+1}\sum_{i=t_j}^{t_{j+1}-1}(M^{-1})_{\tau,j-H+1:j+H-1}R^{(w_i)}_{j-H+1:j+H-1,:}e_{w_i} \\
&\quad + \sum_{j=0}^{q}\sum_{i=t_j}^{t_{j+1}-1}(M^{-1})_{\tau,j}R^{(\delta_i)}_{j,:}e_{\delta_i}. \qquad (3.118)
\end{aligned}$$

Recall that $\bar{\ell} \coloneqq \max\{H\ell_c, H\ell_f, \ell_\mu, \ell_w\}$. We know that the norms of

$$R^{(y)}_{0:H-2,:}, R^{(z)}_{q-H+2:q,:}, R^{(\mu_i)}_{j,:}, R^{(w_i)}_{j-H+1:j+H-1,:}, \text{ and } R^{(\delta_i)}_{j,:}$$

are all upper bounded by $\bar{\ell}$. Taking norm on both sides of (3.118) gives

$$\begin{aligned}
&\left\|\frac{d}{d\theta}\overline{\psi}(\zeta + \eta e)_\tau\right\| \\
&\leq \bar{\ell}\left\|(M^{-1})_{\tau,0:H-2}\right\|\left\|e_y\right\| + \bar{\ell}\left\|(M^{-1})_{\tau,q-H+2:q}\right\|\|e_z\| \\
&\quad + \bar{\ell}\sum_{j=0}^{q}\sum_{i=t_j}^{t_{j+1}-1}\left\|(M^{-1})_{\tau,j}\right\|\left\|e_{\mu_i}\right\| + \bar{\ell}\sum_{j=0}^{q+1}\sum_{i=t_j}^{t_{j+1}-1}\left\|(M^{-1})_{\tau,j-H+1:j+H-1}\right\|\left\|e_{w_i}\right\| \\
&\quad + \bar{\ell}\sum_{j=0}^{q}\sum_{i=t_j}^{t_{j+1}-1}\left\|(M^{-1})_{\tau,j}\right\|\left\|e_{\delta_i}\right\|. \qquad (3.119)
\end{aligned}$$

Note that M can be decomposed as $M = M_a + M_b$, where

$$\begin{aligned}
M_a &\coloneqq \nabla^2_{\hat{x}_{0:q}}\hat{h}_a(\overline{\psi}(\zeta + \eta e), \zeta + \eta e), \\
M_b &\coloneqq \nabla^2_{\hat{x}_{0:q}}\hat{h}_b(\overline{\psi}(\zeta + \eta e), \zeta + \eta e).
\end{aligned}$$

Since M_a is a diagonal $(q+1) \times (q+1)$ matrix and satisfies $M_a \succeq 0$, and M_b is $2H$-banded and satisfies $(m_f + \underline{\ell})I \succeq M_b \succeq m_f I$, we obtain the following with Lemma B.1 in Lin, Hu, Shi, et al., 2021:

$$\left\|(M^{-1})_{\tau,0:H-2}\right\| \leq \frac{2}{m_f}\rho_0^{\tau-(H-2)}, \left\|(M^{-1})_{\tau,q-H+2:q}\right\| \leq \frac{2}{m_f}\rho_0^{q-\tau-(H-2)}$$

$$\left\|(M^{-1})_{\tau,j}\right\| \le \frac{2}{m_f}\rho_0^{|\tau-j|}, \left\|(M^{-1})_{\tau,j-H+1:j+H-1}\right\| \le \frac{2}{m_f}\rho_0^{|\tau-j|-(H-1)},$$

where $\rho_0 := (\sqrt{cond(M_b)}-1)/(\sqrt{cond(M_b)}+1) = 1-2\cdot\left(\sqrt{1+(\underline{\ell}/\mu)}+1\right)^{-1}$. Substituting this into (3.119), we see that

$$\begin{aligned}&\left\|\frac{d}{d\theta}\overline{\psi}(\zeta+\theta e)_\tau\right\|\\ \le\ & C_0\Bigg(\rho_0^\tau\left\|e_y\right\| + \rho_0^{q-\tau}\|e_z\| + \sum_{i=0}^{p}\rho_0^{|\phi(i)-\tau|}\left\|e_{\mu_i}\right\| + \sum_{i=0}^{p+H-1}\rho_0^{|\phi(i)-\tau|}\left\|e_{w_i}\right\|\\ &+\sum_{i=0}^{p}\rho_0^{|\phi(i)-\tau|}\left\|e_{\delta_i}\right\|\Bigg).\end{aligned}$$

Hence we obtain

$$\begin{aligned}&\left\|\overline{\psi}(\zeta)_\tau - \overline{\psi}(\zeta+e)_\tau\right\|\\ =\ &\left\|\int_0^1 \frac{d}{d\eta}\overline{\psi}(\zeta+\eta e)_\tau d\eta\right\|\\ \le\ &\int_0^1 \left\|\frac{d}{d\eta}\overline{\psi}(\zeta+\eta e)_\tau\right\| d\eta\\ \le\ & C_0\Bigg(\rho_0^\tau\left\|e_y\right\| + \rho_0^{q-\tau}\|e_z\| + \sum_{i=0}^{p}\rho_0^{|\phi(i)-\tau|}\left\|e_{\mu_i}\right\| + \sum_{i=0}^{p+H-1}\rho_0^{|\phi(i)-\tau|}\left\|e_{w_i}\right\|\\ &+\sum_{i=0}^{p}\rho_0^{|\phi(i)-\tau|}\left\|e_{\delta_i}\right\|\Bigg).\end{aligned}$$

This finishes the proof of (3.117). Recall that we have $t_i < t_{i+1} \le t_i + d + H$. Therefore, (3.117) implies (3.115). □

In the next lemma, we show a continuity property of the "equality-constrained labeling" method.

Lemma 3.5.4. *Suppose Assumption 3.5.1 holds and $p \ge d$. For a pair of (ξ,σ), if any tuple in the sequence $\{(y_q,z_q;\theta_q)\}_{q=1}^\infty$ satisfies $\psi(y_q,z_q;\theta_q) = \hat{\psi}(y_q,z_q;\theta_q;\xi,\sigma)$ and $\lim_{q\to\infty}(y_q,z_q,\theta_q) = (y,z,\theta)$, then we have*

$$\psi(y,z;\theta) = \hat{\psi}(y,z;\theta;\xi,\sigma).$$

Proof of Lemma 3.5.4. Note that the perturbation bound in Lemma 3.5.3 also establishes the continuity of the function $\hat{\psi}(\cdot,\cdot;\cdot;\xi,\sigma)$. Therefore, we see that

$$\lim_{q\to\infty}\psi(y_q,z_q;\theta_q) = \lim_{q\to\infty}\hat{\psi}(y_q,z_q;\theta_q;\xi,\sigma) = \hat{\psi}(y,z;\theta;\xi,\sigma).$$

Since the constraint set of (3.112) is closed, we know $\hat{\psi}(y, z; \theta; \xi, \sigma)$ is a feasible solution of (3.112).

For the sake of contradiction, we assume $\psi(y, z; \theta) \neq \hat{\psi}(y, z; \theta; \xi, \sigma)$. In this case, since $\hat{\psi}(y, z; \theta; \xi, \sigma)$ is feasible for (3.112), we must have

$$\iota(y, z; \theta) < \hat{\iota}(y, z; \theta; \xi, \sigma).$$

Define the optimality gap as $\Lambda \coloneqq \hat{\iota}(y, z; \theta; \xi, \sigma) - \iota(y, z; \theta)$.

Since $\lim_{q \to \infty}(y_q, z_q; \theta_q) = (y, z; \theta)$, for an arbitrary small positive real number ϵ, we can find a positive integer q such that

$$\left\|y_q - y\right\| + \left\|z_q - z\right\| + dist(\theta, \theta_q) < \epsilon,$$

where $dist(\theta, \theta') = \sum_{i=0}^{p} \left|\mu_i - \mu_i'\right| + \sum_{i=0}^{p+H-1} \left\|w_i - w_i'\right\| + \sum_{i=0}^{p+1} \left|\delta_i - \delta_i'\right|$. Based on $x_{-H+1:p+H-1} \coloneqq \psi(y, z; \theta)$, we construct a feasible solution $x'_{-H+1:p+H-1} \eqqcolon x'$ for the optimization problem (3.112) with parameters $(y_q, z_q; \theta_q)$ as following: Let $x'_{0:p} = x_{0:p}, x_{-H+1:-1} = y, x_{p+1:p+H-1} = z$. For $t = 0, 1, \ldots$, if $x'_t - x'_{t-1} < -\delta_t^{(q)}$, we increase x'_t such that $x'_t = x'_{t-1} - \delta_t^{(q)}$. Then, for $t = p, p-1, \ldots$, if $x'_{t+1} - x'_t < -\delta_{t+1}^{(q)}$, we decrease x'_t such that $x'_t = x'_{t+1} + \delta_{t+1}^{(q)}$. Note that this procedure can guarantee that x' is a feasible solution for (3.112), and their distance are upper bounded by

$$\|\psi(y, z; \theta) - x'\| \leq (2d + 1)\epsilon. \tag{3.120}$$

Since the objective function of (3.112) is Lipschitz in (x, y, z, θ), by (3.120), we know there exists some positive constant c_0 such that

$$\iota(y_q, z_q; \theta_q) - \iota(y, z; \theta) \leq c_0 \left(\|x' - \psi(y, z; \theta)\| + \epsilon\right) \leq (2d + 2)c_0\epsilon. \tag{3.121}$$

On the other hand, by Lemma 3.5.3, we see that

$$\left\|\hat{\psi}(y_q, z_q; \theta_q; \xi, \sigma) - \hat{\psi}(y, z; \theta; \xi, \sigma)\right\| \leq \left(\frac{C}{1 - \rho} + 1\right)\epsilon. \tag{3.122}$$

Since the objective function of (3.112) is smooth in (x, y, z, θ), by (3.122), we see that

$$\left|\hat{\iota}(y_q, z_q; \theta_q; \xi, \sigma) - \hat{\iota}(y, z; \theta; \xi, \sigma)\right| \leq c_0 \left(\frac{C}{1 - \rho} + 2\right)\epsilon. \tag{3.123}$$

Therefore, we see that

$$\begin{aligned} & \hat{\iota}(y_q, z_q; \theta_q; \xi, \sigma) - \iota(y_q, z_q; \theta_q) && \text{(3.124a)} \\ \geq\ & -\left|\hat{\iota}(y_q, z_q; \theta_q; \xi, \sigma) - \hat{\iota}(y, z; \theta; \xi, \sigma)\right| + \left(\hat{\iota}(y, z; \theta; \xi, \sigma) - \iota(y, z; \theta)\right) \end{aligned}$$

$$
\begin{aligned}
&+ (\iota(y,z;\theta) - \iota(y_q,z_q;\theta_q)) \\
\geq\ & - c_0\left(\frac{C}{1-\rho} + 2\right)\epsilon + \Lambda - c_0(2d+2)\epsilon \qquad (3.124b) \\
=\ & \Lambda - c_0\left(\frac{C}{1-\rho} + 2d + 4\right)\epsilon,
\end{aligned}
$$

where we used (3.121) and (3.123) in (3.124b). Let $\epsilon \coloneqq \frac{1}{2}\Lambda c_0^{-1}\left(\frac{C}{1-\rho} + 2d + 4\right)^{-1}$ leads to a contradiction with the assumption that $\hat{\iota}(y_q,z_q;\theta_q;\xi,\sigma) = \iota(y_q,z_q;\theta_q)$. Therefore, we have shown that $\psi(y,z;\theta) = \hat{\psi}(y,z;\theta;\xi,\sigma)$. □

With the above technical lemmas, we are ready to finish the proof of Theorem 3.5.1.

Proof of Theorem 3.5.1. Consider the segment

$$((1-\eta)y + \eta y', (1-\eta)z + \eta z'; (1-\eta)\theta + \eta\theta'), \eta \in [0,1].$$

Note that since $(1-\eta)\psi(y,z;\theta) + \eta\psi(y',z';\theta')$ is a feasible solution for the optimization problem (3.112) parameterized by

$$((1-\eta)y + \eta y', (1-\eta)z + \eta z'; (1-\eta)\theta + \eta\theta'),$$

we know that the corresponding optimization problem is feasible. With some slight abuse of notation, we use $(\xi,\sigma)(\eta) \subseteq \Xi \times \Sigma$ to denote the set of indicators of active constraints and sides such that

$$
\begin{aligned}
&\psi((1-\eta)y + \eta y', (1-\eta)z + \eta z'; (1-\eta)\theta + \eta\theta') \\
=\ &\hat{\psi}((1-\eta)y + \eta y', (1-\eta)z + \eta z'; (1-\eta)\theta + \eta\theta'; \xi,\sigma), \forall(\xi,\sigma) \in (\xi,\sigma)(\eta).
\end{aligned}
$$

By Lemma 3.5.2, we know this set is not empty for any $\eta \in [0,1]$.

We can divide the interval $[0,1]$ into $0 = \eta_0 < \eta_1 < \ldots < \eta_q = 1$ for some positive integer $q \leq 2^{5p+6}$ such that there exists a sequence of *different* indicators of active constraints and sides $(\xi,\sigma)_{0:q-1}$ which satisfies

$$
\begin{aligned}
\psi((1-\eta_i)(y,z;\theta) + \eta_i(y',z';\theta')) &= \hat{\psi}((1-\eta_i)(y,z;\theta) + \eta_i(y',z';\theta'); (\xi,\sigma)_i), \\
\psi((1-\eta_{i+1})(y,z;\theta) + \eta_{i+1}(y',z';\theta')) &= \hat{\psi}((1-\eta_{i+1})(y,z;\theta) + \eta_{i+1}(y',z';\theta'); (\xi,\sigma)_i)
\end{aligned}
$$

for all $0 \leq i \leq q-1$. Note that this requires $(\xi,\sigma)(\eta_i)$ to contain both $(\xi,\sigma)_{i-1}$ and $(\xi,\sigma)_i$ for $i = 1,\ldots,q-1$. To construct the sequence $\eta_{0:q}$ and $(\xi,\sigma)_{0:q-1}$, we first have $\eta_0 = 0$ and let $(\xi,\sigma)_0$ be any pair $(\xi,\sigma) \in (\xi,\sigma)(\eta_0)$ such that

$$\sup\{\eta \in [0,1] \mid \psi((1-\eta)(y,z;\theta) + \eta(y',z';\theta'))$$

$$= \hat{\psi}\left((1-\eta)(y,z;\theta) + \eta(y',z';\theta'); \xi, \sigma\right)\} > 0,$$

and let η_1 be the supremum value above. Since $0 = \inf(0,1]$ and $(\xi,\sigma)(\eta) \subseteq \Xi \times \Sigma$ is nonempty for every $\eta \in (0,1]$, we know such $(\xi,\sigma)_0$ exists by Lemma 3.5.4. Suppose we have already constructed $\eta_{0:i}$, $(\xi,\sigma)_{0:i-1}$, and $\eta_i < 1$. Then we select $(\xi,\sigma)_i$ to be any pair (ξ,σ) such that

$$\begin{aligned}&\sup\{\eta \in [0,1] \mid \psi\left((1-\eta)(y,z;\theta) + \eta(y',z';\theta')\right)\\ &= \hat{\psi}\left((1-\eta)(y,z;\theta) + \eta(y',z';\theta'); \xi, \sigma\right)\} > \eta_i,\end{aligned}$$

and let η_{i+1} be the supremum value above. We can repeat this construction and stop when $\eta_{i+1} = 1$. By the construction, we know all pairs in the sequence $(\xi,\sigma)_{0:i-1}$ are distinct, thus the construction will terminate in finite time. Hence, we have a finite index q such that $\eta_q = 1$.

By Lemma 3.5.3, we know that

$$\begin{aligned}&\left\|\psi\left((1-\eta_i)(y,z;\theta) + \eta_i(y',z';\theta')\right)_t - \psi\left((1-\eta_{i+1})(y,z;\theta) + \eta_{i+1}(y',z';\theta')\right)_t\right\|\\ &\le (\eta_{i+1}-\eta_i)C\left(\rho^t\|y-y'\| + \rho^{p-t}\|z-z'\|\right) \qquad (3.125)\\ &\quad + (\eta_{i+1}-\eta_i)C\left(\sum_{\tau=0}^{p}\rho^{|t-\tau|}\left|\mu_\tau - \mu_\tau'\right| + \sum_{\tau=0}^{p+H-1}\rho^{|t-\tau|}\left\|w_\tau - w_\tau'\right\| + \sum_{\tau=0}^{p+1}\rho^{|t-\tau|}\left\|\delta_\tau - \delta_\tau'\right\|\right). \qquad (3.126)\end{aligned}$$

Summing (3.125) over $i = 0, 1, \ldots, q-1$ finishes the proof. □

Chapter 4

STOCHASTIC PREDICTIONS

The results in Chapter 3 provide performance guarantees for MPC-style policies under potentially adversarial ways to generate the predictions under constraints such as prediction error bounds. Although such worst-case guarantees are important for many safety-critical or risk-averse scenarios, it may be overly-conservative in other applications where we care about the performance in expectation. The adversarial prediction model can be insufficient to characterize the benefit of predictive control because it overlooks the stochastic dependence between predictions and unknown problem parameters. In addition, instead of asking what a standard predictive policy like MPC can achieve, we seek to answer a more fundamental question:

What is the maximum achievable cost improvement under the optimal policy to leverage predictions relative to the no-prediction scenario?

As we will show in this chapter, the optimal predictive policy can be expressed equivalently as MPC in specific problem settings, but this equivalence does not hold in general.

In this chapter, we introduce a stochastic prediction model and define *prediction power* as the maximum cost improvement in the above question. We show that prediction power is always non-negative and establish a lower bound under two sufficient conditions, characterizing the fundamental benefit of incorporating stochastic predictions. We instantiate this in two settings: (i) in linear quadratic regulator, we derive a closed-form prediction power expression and reveal a mismatch between prediction accuracy and control cost, and (ii) for non-quadratic costs, we show that even weakly dependent predictions yield significant performance gains.

This chapter is based on the following paper:

[Lin, Chen, et al., 2025] Lin, Yiheng, Zaiwei Chen, Christopher Yeh, and Adam Wierman. “Maximizing the value of stochastic prediction in control: Accuracy is not enough.” Under submission.

4.1 Problem Setting

We consider a finite-horizon discrete-time optimal control problem with time-varying dynamics and cost functions, where state transitions are subject to random disturbances:

$$\begin{aligned}&\text{Control dynamics: } X_{t+1} = f_t\left(X_t, U_t; W_t\right), \quad 0 \le t < T, \text{ with } X_0 = x_0;\\ &\text{Stage cost: } h_t(X_t, U_t), \quad 0 \le t < T, \quad \text{and terminal cost: } h_T(X_T).\end{aligned} \tag{4.1}$$

At each time step t, we let X_t denote the system state and U_t denote the control action chosen by an agent. The function $f_t : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^n$ defines how the next state X_{t+1} depends on the current state X_t, the control action U_t, and the random disturbance W_t. The agent incurs a stage cost $h_t(X_t, U_t)$ at each intermediate time step $t < T$ and a terminal cost $h_T(X_T)$ at the final time step T. At each time step t, the controller observes the past disturbance W_{t-1} and a (possibly random) prediction vector $V_t(\theta) \in \mathbb{R}^d$ before selecting a control action U_t, where θ is a parameter of the predictor generating the prediction. We formally define the concept of *predictions* and the parameter θ in the following.

Definition 4.1.1 (Predictions)**.** *At each time step t, the predictor with parameter $\theta \in \Theta$ generates a prediction $V_t(\theta)$, where Θ denotes the set of all possible predictor parameters. The predictions $\{V_{0:T-1}(\theta)\}_{\theta\in\Theta}$ and the disturbances $W_{0:T-1}$ live in the same probability space.*

Compared with previous works (Lin, Hu, Shi, et al., 2021; Li, Yang, Qu, Shi, et al., 2022) that assume predictions targeting specific disturbances, Definition 4.1.1 focuses on the stochastic relationship between predictions and system uncertainties, yielding a unified framework for comparing different forms of prediction based on their effectiveness for control—even if their precise nature is unknown. Because predictions and disturbances share the same probability space, we can compare prediction sequences $V_{0:T-1}(\theta)$ and $V_{0:T-1}(\theta')$, generated by different predictors with parameters θ and θ'.

Observe that the disturbances $W_{0:T-1}$ and predictions in Definition 4.1.1 do not depend on the current state or past trajectory, reflecting their exogenous nature. For example, consider the problem of quadcopter control in windy conditions (O'Connell et al., 2022). In this case, the wind disturbances are not influenced by the quadcopter's state or control inputs. Under this causal relationship, we define the *problem instance* as $\Xi = \left(W_{0:T-1}, \{V_{0:T-1}(\theta)\}_{\theta\in\Theta}\right)$, and make the following assumption.

Assumption 4.1.1. *The generation of the problem instance Ξ is oblivious*—i.e., *the process of sampling Ξ from the distribution of problem instances is not affected by the agent's states or actions.*

Let $\xi = \left(w_{0:T-1}, \{v_{0:T-1}(\theta)\}_{\theta\in\Theta}\right)$ denote a realization of the problem instance, including disturbances and all parameterized predictions. Under Assumption 4.1.1, Ξ is viewed as realized to ξ before control begins, although the agent observes each disturbance and prediction step by step. Similar assumptions about oblivious environments or predictions appear in online optimization (Hazan, 2016; Rutten et al., 2023), ensuring that future disturbances or predictions remain unchanged by past states or actions. Hence, for a fixed predictor parameter θ, we define a *predictive policy* as a mapping from the current state and past disturbances and predictions to a control action.

Definition 4.1.2 (Predictive policy)**.** *Consider a fixed predictor parameter θ. For each time step t, let $I_t(\theta) \coloneqq (W_{0:t-1}, V_{0:t}(\theta))$ denote the history of past disturbances and predictions, and let $\mathcal{F}_t(\theta) \coloneqq \sigma(I_t(\theta))$[1]. A* predictive policy *that applies to the predictor with parameter θ is a sequence of functions $\pi_{0:T-1}$, where π_t maps a state/history pair to a control action.*

Given a fixed predictive policy sequence $\pi = \pi_{0:T-1}$ for a predictor parameter θ, we evaluate its performance via the expected total cost over Ξ:

$$J^\pi(\theta) \coloneqq \mathbb{E}[\sum_{t=0}^{T-1} h_t(X_t, U_t) + h_T(X_T)],$$

where $X_0 = x_0$, $X_{t+1} = f_t(X_t, U_t; W_t)$, $U_t = \pi_t(X_t; I_t(\theta))$, for $t = 0, \ldots, T-1$. The optimal cost under θ is defined as $J^*(\theta) = \min_\pi J^\pi(\theta)$, where the minimum is over all predictive policies that use the predictor parameter θ.

Following prior works on the benefits of using predictions in online decision making (Yu et al., 2020), we define *prediction power* by comparing against a baseline that provides minimal information (*e.g.*, no prediction). Without loss of generality, let $\mathbf{0} \in \Theta$ be the baseline predictor parameter so that any $\theta \neq \mathbf{0}$ provides at least as much information as $\mathbf{0}$, *i.e.*, $\mathcal{F}_t(\theta) \supseteq \mathcal{F}_t(\mathbf{0})$. Based on this baseline, we define *prediction power* as the maximum possible cost improvement achieved by using predictions under θ relative to the baseline, formally stated in Definition 4.1.3.

[1] For any random variable Y, we use $\sigma(Y)$ to denote the σ-algebra it generates.

Definition 4.1.3 (Prediction power)**.** *For a predictor with parameter θ, its prediction power in the optimal control problem* (4.1) *is $P(\theta) \coloneqq J^*(\mathbf{0}) - J^*(\theta)$.*

Our definition of prediction power is based on the optimal control policy under a given predictor parameter and, therefore, is independent of any specific policy class. Many previous works have considered prediction-enabled improvement within a specific policy class (Li, Qu, and Li, 2018; Li, Chen, and Li, 2019; Chen, Agarwal, et al., 2015), where they focus on changes in $J^\pi(\theta)$ rather than $J^*(\theta)$. In other works, policies include parameters that can be tuned to perform optimally under a specific predictor; that is, $\min_{\pi \in \text{a policy class}} J^\pi(\theta)$. While these approaches are useful in specific application scenarios, our definition, based on the general optimal policy, is more universal because: (1) imposing policy class constraints may lead to performance loss, and (2) the extent of improvement can depend on policy design and parameterization, which shifts the focus away from valuing predictions themselves.

4.2 Sufficient Conditions for Characterization

Our main results characterize prediction power $P(\theta)$ to help determine whether and which predictions yield better performance. If a lower bound of $P(\theta)$ is greater than the cost of obtaining the predictor with parameter θ, it is beneficial to use the predictor, assuming that we can design or learn a near-optimal predictive policy.

Throughout this paper, let $\bar{\pi} = \bar{\pi}_{0:T-1}$ denote the optimal policy for the predictor with parameter $\mathbf{0}$ and $\pi^\theta = \pi^\theta_{0:T-1}$ denote the optimal policy for the predictor with parameter θ. In other words, $J^{\bar{\pi}}(\mathbf{0}) = J^*(\mathbf{0})$ and $J^{\pi^\theta}(\theta) = J^*(\theta)$. To compare the policies π^θ and $\bar{\pi}$, we introduce a function that we call the *instance-dependent Q function*, inspired by the Q function in the study of Markov decision processes (MDPs). For a given state-action pair (x, u) and problem instance ξ, the instance-dependent Q function for a policy π evaluates the remaining cost incurred by taking action u from state x and then following policy π for all future time steps. Using $\iota_\tau(\theta)$ to denote the realization of $I_\tau(\theta)$, for any $\tau = 0, \ldots, T-1$, the instance-dependent Q function is defined as

$$Q_t^{\pi^\theta}(x, u; \xi) = \sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T), \text{ where } x_t = x,\ u_t = u, \text{ and}$$
$$x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_\tau), \text{ for } t \le \tau < T;\ u_\tau = \pi_\tau^\theta(x_\tau; \iota_\tau(\theta)), \text{ for } t < \tau < T. \quad (4.2)$$

The disturbance w_τ and the history $\iota_\tau(\theta)$ in (4.2) are decided by the problem instance ξ, which is an input to $Q_t^{\pi^\theta}$. Similarly, we can define $Q_t^{\bar{\pi}}(x, u; \xi)$ by replacing θ

with $\mathbf{0}$ and π^θ with $\bar{\pi}$ in (4.2). Importantly, our instance-dependent Q function is different from the classical definition of the Q function for MDPs or reinforcement learning (RL), where it is the *expectation* of the cost to go. The instance-dependent Q function denotes the actual remaining cost, which is a $\sigma(\Xi)$-measurable *random variable*. The classic definition of the Q function can be recovered by taking the conditional expectation, *i.e.*, $\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$. It is worth noting that our instance-dependent Q function is about the *cost* instead of the *reward*, so lower values are better.

With this definition of the instance-dependent Q function, the optimal policies $\bar{\pi}$ and π^θ can be expressed as recursively minimizing the corresponding expected Q functions conditioned on the available history. Starting with $C_T^{\pi^\theta}(x; \xi) = h_T(x)$, for time step $t = T-1, \dots, 0$, we have

$$\begin{aligned}
&Q_t^{\pi^\theta}(x, u; \xi) \coloneqq h_t(x, u) + C_{t+1}^{\pi^\theta}(f_t(x, u; w_t); \xi), \text{ for } x \in \mathbb{R}^n,\ u \in \mathbb{R}^m, \text{ and } \xi;\\
&\pi_t^\theta(x; \iota_t(\theta)) \coloneqq \arg\min_{u \in \mathbb{R}^m} \mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right], \text{ for } x \in \mathbb{R}^n \text{ and } \iota_t(\theta);\\
&C_t^{\pi^\theta}(x; \xi) \coloneqq Q_t^{\pi^\theta}(x, \pi_t^\theta(x; \iota_t(\theta)); \xi), \text{ for } x \in \mathbb{R}^n \text{ and problem instance } \xi. \qquad (4.3)
\end{aligned}$$

Similar recursive relationships also defines the optimal policy $\bar{\pi}$ for the baseline predictions, and we only need to replace θ with $\mathbf{0}$ and π^θ with $\bar{\pi}$ in the above equations. The recursive equations in (4.3) can be viewed as a generalization of the classical Bellman optimality equation for general MDPs.

We are now ready to introduce our main result, which is a lower bound on the prediction power $P(\theta)$. Our result relies on two conditions about a growth property of the expected Q function under π^θ and the covariance of the optimal policy's action when conditioned on the σ-algebra $\mathcal{F}_t(\mathbf{0})$ of the baseline. We state these conditions formally and provide intuitive explanations.

Condition 4.2.1. *For a sequence of positive semi-definite matrices $M_{0:T-1}$, the following inequality holds for all time steps $0 \le t < T$: For any $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and history $\iota_t(\theta)$,*

$$\begin{aligned}
&\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) - C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]\\
&\ge (u - \pi_t^\theta(x; \iota_t(\theta)))^\top M_t (u - \pi_t^\theta(x; \iota_t(\theta))). \qquad (4.4)
\end{aligned}$$

Condition 4.2.1 states that conditioned on any history $\iota_t(\theta)$, the expected Q function of policy π^θ grows at least quadratically as the action u deviates from the optimal policy's action. Note that one can always pick M_t to be the all-zeros matrix to make Condition 4.2.1 hold, but the choice of M_t will affect the prediction power bound in Theorem 4.2.3. When $M_t \succ 0$, deviating from the action of policy π^θ causes a non-negligible loss. The loss is characterized by the difference between the resulting Q function value and the cost-to-go function value. When this condition does not hold with any non-zero matrix M_t, one can construct an extreme case when $Q_t^{\pi^\theta}$ is a constant by letting all cost functions $h_{0:T}$ be constants; in this case, the prediction power must be zero because every policy achieves the same total cost no matter what predictions they use.

Condition 4.2.2. *One of the following holds for the optimal policy π^θ:*

(a) For positive semi-definite matrices $\Sigma_{0:T-1}$, the following holds for all time steps $0 \le t < T$:

$$\mathbb{E}\left[\mathbf{Cov}\left[\pi_t^\theta(X; I_t(\theta)) \mid I_t(\mathbf{0})\right]\right] \succeq \Sigma_t, \textit{ for any } \mathcal{F}_t(\mathbf{0})\textit{-measurable } X. \quad (4.5)$$

(b) For nonnegative scalars $\sigma_{0:T-1}$, the following holds for all time steps $0 \le t < T$:

$$\mathbb{E}\left[\mathrm{Tr}\left\{\mathbf{Cov}\left[\pi_t^\theta(X; I_t(\theta)) \mid I_t(\mathbf{0})\right]\right\}\right] \ge \sigma_t, \textit{ for any } \mathcal{F}_t(\mathbf{0})\textit{-measurable } X. \quad (4.6)$$

Before discussing the details, we note that by setting $\sigma_t = \mathrm{Tr}(\Sigma_t)$, Condition 4.2.2 (a) implies (and is therefore stronger than) Condition 4.2.2 (b). Similar to Condition 4.2.1, one can always pick Σ_t to be all-zeros matrix to satisfy Condition 4.2.2 (a), but it will affect the prediction power bound in Theorem 4.2.3.

Condition 4.2.2 (a) states that conditioned on the history $I_t(\mathbf{0})$ from the baseline, the covariance matrix of policy π^θ's action from any $\mathcal{F}_t(\mathbf{0})$-measurable state is positive semi-definite in expectation. Recall that $\mathcal{F}_t(\mathbf{0}) = \sigma(I_t(\mathbf{0}))$. To understand this, suppose that the agent only has access to the baseline information. Then, the agent cannot predict the action that policy π^θ would take. This should usually hold because the action $\pi_t^\theta(X; I_t(\theta))$ is not $\mathcal{F}_t(\mathbf{0})$-measurable, and the lower bound in (4.5) implies the mean-square prediction error cannot improve below a certain threshold. When this condition does not hold with non-zero matrix Σ_t (or scalar σ_t), one can design a policy $\bar{\pi}'$ that always picks the same action as π^θ but only requires access to the baseline information $I_t(\mathbf{0})$, which implies $P(\theta) = 0$ because $J^*(\mathbf{0}) \le J^{\bar{\pi}'}(\mathbf{0}) = J^*(\theta)$. This can happen, for example, when all disturbances $W_{0:T-1}$ are deterministic.

Note it is possible that the optimal action at different states has a positive variance in different directions, but there is no non-trivial lower bound on the covariance matrix as required by Condition 4.2.2 (a). In this case, Condition 4.2.2 (b) provides a weaker alternative and would be useful when we can only establish a lower bound on the trace of the optimal action's covariance matrix.

Theorem 4.2.3. *If Conditions 4.2.1 and 4.2.2 (a) hold with matrices $M_{0:T-1}$ and $\Sigma_{0:T-1}$, then $P(\theta) \geq \sum_{t=0}^{T-1} \mathrm{Tr}\{M_t \Sigma_t\}$. Alternatively, if Conditions 4.2.1 and 4.2.2 (b) hold with matrices $M_{0:T-1}$ and scalars $\sigma_{0:T-1}$, then $P(\theta) \geq \sum_{t=0}^{T-1} \mu_{min}(M_t) \cdot \sigma_t$, where $\mu_{min}(\cdot)$ returns the smallest eigenvalue.*

We defer the proof of Theorem 4.2.3 to Section 4.A. There are two main takeaways of Theorem 4.2.3. First, recall that one can always pick M_t and Σ_t to be the all-zeros matrices to satisfy Conditions 4.2.1 and 4.2.2. In this case, Theorem 4.2.3 states that $P(\theta) \geq 0$, which means that having predictions, no matter how weak they are, does not hurt. Second, to characterize the improvement in having predictions, the two Conditions 4.2.1 and 4.2.2 can establish a lower bound for the prediction power that is strictly positive if $\mathrm{Tr}\{M_t \Sigma_t\} > 0$ or $\mu_{\min}(M_t)\sigma_t > 0$. We provide an example to help illustrate how Conditions 4.2.1 and 4.2.2 (a) can work together to ensure that the predictions can lead to a strict improvement on the control cost (see Figure 4.1 for an illustration).

Example 4.2.4. *Consider the following optimal control problem*

$$\text{Dynamics: } X_{t+1} = U_t + W_t, \text{ Stage cost: } h_t(x,u) = x^2, \text{ Terminal cost: } h_T(x) = x^2,$$

*where each disturbance W_t is sampled independently according to $\mathbb{P}(W_t = -1) = \mathbb{P}(W_t = 1) = \frac{1}{2}$. Suppose that the predictor with parameter θ can predict W_t exactly (*i.e.*, $V_t(\theta) = W_t$), while the baseline predictor is uninformative (*e.g.*, $V_t(\mathbf{0}) = 0$). The Q functions, cumulative cost, and optimal actions under each predictor are*

$$\begin{aligned} Q_t^{\pi^\theta}(x,u;\Xi) &= x^2 + (u + V_t(\theta))^2, & Q_t^{\bar{\pi}}(x,u;\Xi) &= x^2 + (u + W_t)^2 + (T - t - 1), \\ C_t^{\pi^\theta}(x;\Xi) &= x^2, & C_t^{\bar{\pi}}(x;\Xi) &= x^2 + (T - t), \\ \pi_t^\theta(x; I_t(\theta)) &= -V_t(\theta) = -W_t, & \bar{\pi}_t(x; I_t(\mathbf{0})) &= 0. \end{aligned}$$

The Q function $Q_t^{\pi^\theta}$ is strongly convex in u, with Condition 4.2.1 holding for any $M_t \in [0,1]$. Furthermore, the optimal action has positive variance, with Condition 4.2.2 (a) holding for any $\Sigma_t \in [0,1]$. Thus, by Theorem 4.2.3, the prediction power satisfies $P(\theta) \geq T$. Indeed, by comparing the cumulative cost functions, we see

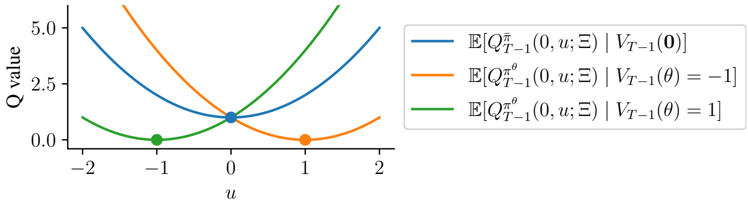

Figure 4.1: An illustration of why predictions are helpful, corresponding to Example 4.2.4. The expected Q functions with perfect predictions (green and orange lines) have lower minima than the expected Q function with uninformative predictions (blue line).

that the predictor with parameter θ incurs a lower cumulative cost by exactly T (as expected by Theorem 4.3.1).

Figure 4.1 illustrates the expected Q functions at time $t = T - 1$ and $x = 0$, which the policies $\pi_t^\theta(x; I_t(\theta))$ and $\bar{\pi}_t(x; I_t(\mathbf{0}))$ seek to minimize. The expected Q functions with perfect predictions have lower minima than the expected Q function with uninformative predictions.

Theorem 4.2.3 provides a useful tool to characterize the prediction power by reducing the problem of comparing two policies π^θ and $\bar{\pi}$ over the whole horizon to studying the properties of one policy π^θ at each time step. Our proof of Theorem 4.2.3 follows the same intuition as the widely-used performance difference lemma in RL (see Lemma 6.1 in Kakade and Langford (2002)), comparing the per-step "advantage" of π^θ along the trajectory of $\bar{\pi}$. When only the baseline information is available, the agent must pick a suboptimal action (4.5) and incur a loss (4.4) at each time step. The per-step losses accumulate to the total cost difference.

While Theorem 4.2.3 applies to the general dynamical system and cost functions in (4.1), the two conditions with their key coefficients M_t and Σ_t (or σ_t) still depend on the optimal Q function and the optimal policy that are implicitly defined through the recursive equations (4.3). To instantiate Theorem 4.2.3, we need to derive explicit expressions of M_t and Σ_t under more specific dynamics/costs. We study two cases in the rest of the paper. In Section 4.3, we first study the linear-quadratic regulator (LQR) problem to characterize the key factors for deciding the prediction power. In this setting, the optimal Q functions and the optimal policy have closed-form solutions. We obtain M_t and Σ_t that characterize the exact prediction power $P(\theta)$. Then, we study a time-varying linear system with general well-conditioned cost functions, where the optimal Q functions and the optimal policy do not have closed-form solutions. In this case, we can still verify that Conditions 4.2.1 and 4.2.2 (b) hold with nonzero M_t and σ_t, which yields a non-trivial lower bound of $P(\theta)$.

4.3 Applications: Linear Quadratic Regulator and Beyond

LTV Dynamics with Quadratic Costs. Consider a linear time-varying (LTV) dynamical system with quadratic costs:

$$\begin{aligned}&\text{Control dynamics: } X_{t+1} = A_t X_t + B_t U_t + W_t, \text{ for } 0 \le t < T;\\ &\text{stage cost: } X_t^\top Q_t X_t + U_t^\top R_t U_t, \text{ for } 0 \le t < T; \text{and terminal cost: } X_T^\top P_T X_T, \quad (4.7)\end{aligned}$$

where $Q_{0:T-1}, R_{0:T-1}$, and P_T are symmetric positive definite. The classic linear quadratic regulator (LQR) problem, along with its time-varying variant that we consider, has been used widely as a benchmark setting in the learning-for-control literature. It also serves as a good approximation of nonlinear systems near equilibrium points, making it amenable to standard analytical tools.

To apply Theorem 4.2.3, we first derive closed-form expressions for the optimal Q function, Q^{π^θ}, and the optimal policy, π^θ, which are used to verify Conditions 4.2.1 and 4.2.2 (a). We begin by defining key quantities that will be useful for stating the main results in this section.

Definition 4.3.1. *For $t = T-1, \ldots, 0$, we define the matrices H_t, P_t, and K_t recursively according to*

$$\begin{aligned}&H_t = B_t(R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top, \quad P_t = Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} H_t P_{t+1} A_t, \text{ and}\\ &K_t = (R_t + B_t^\top P_{t+1} B_t)^{-1}(B_t^\top P_{t+1} A_t). \quad (4.8)\end{aligned}$$

Moreover, we define the transition matrix Φ_{t_2,t_1} as $\Phi_{t_2,t_1} = I$ if $t_2 \le t_1$ and

$$\Phi_{t_2,t_1} = (A_{t_2-1} - B_{t_2-1}K_{t_2-1})(A_{t_2-2} - B_{t_2-2}K_{t_2-2}) \cdots (A_{t_1} - B_{t_1}K_{t_1}), \text{ if } t_2 > t_1. \quad (4.9)$$

The matrix K_t is the feedback gain matrix in the optimal policy, and P_t is the matrix that defines the quadratic term in the optimal cost-to-go function. To simplify notation, we define the shorthands $W_{\tau|t}^\theta \coloneqq \mathbb{E}\left[W_\tau \mid I_t(\theta)\right]$ and $w_{\tau|t}^\theta \coloneqq \mathbb{E}\left[W_\tau \mid I_t(\theta) = \iota_t(\theta)\right]$.

Proposition 4.3.1. *In the case of LTV dynamics with quadratic costs, the conditional expectation of the optimal Q function $\mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$ can be expressed as*

$$\left(u + K_t x - \bar{u}_t^\theta(\iota_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t) \left(u + K_t x - \bar{u}_t^\theta(\iota_t(\theta))\right) + \psi_t^{\pi^\theta}(x; \iota_t(\theta)),$$

where $\psi_t^{\pi^\theta}(x; \iota_t(\theta))$ is a function of the state x and the history $\iota_t(\theta)$ that does not depend on the control action u. Here,

$$\bar{u}_t^\theta(\iota_t(\theta)) \coloneqq -(R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} w_{\tau|t}^\theta.$$

And the optimal policy can be expressed as $\pi_t^\theta(x; \iota_t(\theta)) = -K_t x + \bar{u}_t^\theta(\iota_t(\theta))$.

We derive the closed-form expressions in Proposition 4.3.1 by induction following the backward recursive equations in (4.3); the full proof is deferred to Section 4.C. With these expressions, we can verify Conditions 4.2.1 and 4.2.2 (a) to obtain a closed-form expression of the prediction power.

Theorem 4.3.1. *In the case of LTV dynamics with quadratic costs, the prediction power of the predictor with parameter θ is $P(\theta) = \sum_{t=0}^{T-1} \mathrm{Tr}\{M_t \Sigma_t\}$, where $M_t \coloneqq R_t + B_t^\top P_{t+1} B_t$, and $\Sigma_t = \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\right]$.*

As a remark, the prediction power in Theorem 4.3.1 holds with *equality* due to the special structure of LQR. We provide a detailed discussion in the proof of Theorem 4.3.1 in Section 4.C.

While the optimal policy in Proposition 4.3.1 is restricted to the LQR case, we can interpret the optimal policy as planning according the conditional expectation following the idea of model predictive control (MPC) (Yu et al., 2020), which is easier to generalize. The agent needs to solve an optimization problem and re-plan at every time step. At time step t, the agent solves

$$\begin{aligned} \arg\min_{u_{t:T-1}} \quad & \mathbb{E}\left[\sum_{\tau=t}^{T-1} h_\tau(X_\tau, u_\tau) + h_T(X_T) \,\middle|\, I_t(\theta) = \iota_t(\theta)\right] \\ \text{s.t.} \quad & X_{\tau+1} = f_\tau(X_\tau, u_\tau; W_\tau), \text{ for } \tau \geq t, \text{ and } X_t = x. \end{aligned} \tag{4.10}$$

Then, the agent commits to the first entry $u_{t|t}$ of the optimal solution as $\pi_t^\theta(x; \iota_t(\theta))$. In the LQR setting, we can further simplify it to be *planning according to $w_{\tau|t}^\theta$* (see Section 4.C).

The MPC forms of the optimal policy in (4.10) extends the result in Yu et al. (2020), which shows that MPC is the optimal predictive policy under the accurate prediction model in time-variant LQR. When the predictions are inaccurate and the system is time-varying, MPC is still optimal if we solve the predictive optimal control problem in expectation (4.10).

Prediction Power ≠ Accuracy. As Proposition 4.3.1 suggests, one way to implement the optimal policy is to predict each of the future disturbances $W_{t:T-1}$ and generate the estimations $w^{\theta}_{(t:T-1)|t}$ in deciding the action at time step t. However, two controllers with the same estimation error (as measured by mean squared error (MSE)) can have very different control costs. Because of this reason, the control cost bounds depend on the estimation errors in previous works (Zhang, Li, and Li, 2021; Yu et al., 2022; Lin, Hu, Qu, et al., 2022) must be loose, so one cannot rely on them to infer or compare the values of different predictors.

To illustrate this point, we provide an example where the prediction power can change significantly when the prediction accuracy does not change.

Example 4.3.2. *Consider the time-invariant LQR setting,* i.e.*, assume $A_t = A, B_t = B, Q_t = Q, R_t = R$ for all t and $P_T = P$ is the solution to Discrete-time Riccati Equation (DARE) in* (4.7)*. Suppose the disturbance is sampled $W_t \overset{i.i.d.}{\sim} N(0, I)$ at every time step t. Let $\rho \in [0, \frac{\sqrt{2}}{2}]$ be a fixed coefficient. We construct a class of predictors from the disturbances $\{W_t\}$ by applying the affine transformation $V_t(\theta) := \rho\theta W_t + \epsilon_t(\rho, \theta)$ for $\theta \in \mathbb{R}^{2\times 2}$ that satisfies $\theta\theta^\top \preceq \frac{1}{2}I$, where the random noise $\epsilon_t(\rho, \theta)$ is independently sampled from a Gaussian distribution $N(0, I - \rho^2\theta\theta^\top)$.*

We can construct θ such that $V_t(\theta)$ and $V_t(I)$ achieve the same mean-square error (MSE) when predicting each individual entry of W_t, yet $P(I) > P(\theta)$. To construct θ, note that $(W_t, V_t(\theta))$ satisfies $\mathbb{E}[W_t \mid V_t(\theta)] = \rho\theta^\top V_t$ and $\mathbf{Cov}[W_t \mid V_t(\theta)] = I - \rho^2\theta^\top\theta$. Thus, we can change θ without affecting the MSE of predicting each individual entry as long as the diagonal entries of $\theta^\top\theta$ remain the same. However, by Theorem 4.3.1, we know the prediction power is equal to $\rho^2 T \cdot \mathrm{Tr}\{\theta^\top\theta PHP\}$, where $H = B(R + B^\top PB)^{-1}B^\top$. Thus, the off-diagonal entries of $\theta^\top\theta$ can also affect the value of$\mathrm{Tr}\{\theta^\top\theta PHP\}$*. We instantiate this example with a 2-D double-integrator dynamical system in Section 4.B: the predictors with parameters I and θ shares the same MSE but their prediction powers are significantly different.*

Example 4.3.2 shows how prediction power can vary even when the accuracy of predicting each entry of the disturbance W_t remains the same, where the construction leverages the covariance between the predictions for different entries of W_t. While the construction in Example 4.3.2 requires $n \geq 2$, we also provide an example with $n = 1$ and multiple steps of predictions in Section 4.B. From these examples, it is clear that one should not use the MSEs of predicting future disturbances to infer the prediction power. The intuition behind this mismatch is that MSE does not depend on matrices (A, B, Q, R), but the prediction power does. The mismatch also relates to the findings in the decision-focused learning literature, which we discuss in detail in Related Works.

Prediction Power Evaluation. In this section, we propose an algorithm (cf. Algorithm 4) to evaluate the prediction power efficiently given a set of historical problem instances $\{\xi_n\}_{n=1}^N$. We start with defining a quantity whose estimation error is closely related to the policy's performance:

$$\bar{u}_t^*(\Xi) \coloneqq -(R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} W_\tau. \tag{4.11}$$

We call $\bar{u}_t^*(\Xi)$ the surrogate-optimal action, because it is the optimal action that an agent should take with the oracle knowledge of all future disturbances at time t. In the prediction power given by Theorem 4.3.1, we can express $\bar{u}_t^\theta(I_t(\theta))$ as $\mathbb{E}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta)\right]$, which is the expectation of $\bar{u}_t^*(\Xi)$ condition on the the history at time step t.

Now we come back to the design of Algorithm 4. While iterating backward from time step $T - 1$ to 0, the algorithm first constructs a dataset of the surrogate optimal action $\bar{u}_t^*(\Xi)$ as the fitting target. Then, the algorithm estimates the covariance of $\bar{u}_t^*(\Xi)$ when conditioning on $I_t(\mathbf{0})$ and $I_t(\theta)$, respectively, using a subroutine (see Algorithm 5 in Section 4.C). The last step of Algorithm 4 gives the prediction power because $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\right]$ can be decomposed as $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\mathbf{0})\right]\right] - \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta)\right]\right]$, and we defer the proof to Section 4.C.

LTV Dynamics with General Costs Consider a system with linear time-varying dynamics and more general cost functions that depend on the states and control actions.

$$\begin{aligned}&\text{Control dynamics: } X_{t+1} = A_t X_t + B_t U_t + W_t, \text{ for } 0 \le t < T;\\ &\text{stage cost: } h_t^x(X_t) + h_t^u(U_t), \text{ for } 0 \le t < T; \text{and terminal cost: } h_T^x(X_T).\end{aligned} \tag{4.12}$$

The LTV system with quadratic cost functions studied in the previous section is a special case of (4.12). The generality of (4.12) leads to more challenging because the optimal Q function Q^{π^θ} and the optimal policy π^θ no longer have closed-form expressions like Proposition 4.3.1. While one may consider using the MPC policy in (4.10) to evaluate the prediction power, we can construct an example where this policy is suboptimal for non-quadratic costs (see Section 4.B for details).

Algorithm 4: Prediction Power Evaluation

```
Require: Dataset $D$ of problem instances $\{\xi_n\}_{n=1}^N$.
for $t = T-1, T-2, \ldots, 0$ do
    Compute $P_t, H_t, K_t$ and $\{\Phi_{t,t'}\}_{t' \geq t}$ according to (4.8) and (4.9).
    Compute $M_t = R_t + B_t^\top P_{t+1} B_t$.
    for $n = 1, 2, \ldots, N$ do
        Compute $\bar{u}_t^*(\xi_n)$ according to (4.11) in problem instance $\xi_n$.
    end
    Call Algorithm 5 to estimate $\Sigma_t^0 \coloneqq \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(0)\right]\right]$ using
      $\{(\bar{u}_t^*(\xi_n), \iota_t^n(0))\}_{n=1}^N$.
    Call Algorithm 5 to estimate $\Sigma_t^\theta \coloneqq \mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi) \mid I_t(\theta)\right]\right]$ using
      $\{(\bar{u}_t^*(\xi_n), \iota_t^n(\theta))\}_{n=1}^N$.
end
return $P(\theta) = \sum_{t=0}^{T-1} \mathrm{Tr}\{\Sigma_t^0 M_t\} - \sum_{t=0}^{T-1} \mathrm{Tr}\{\Sigma_t^\theta M_t\}$
```

We follow the recursive equations (4.3) to establish Conditions 4.2.1 and 4.2.2 (b). We make the following assumptions about the cost functions and dynamical matrices:

Assumption 4.3.1. *For every time step t, h_t^x is μ_x-strongly convex and ℓ_x-smooth; h_t^u is μ_u-strongly convex and ℓ_u-smooth; The dynamical matrices satisfy that $\mu_A I \preceq A_t^\top A_t \preceq \ell_A I$ and $\mu_B I \preceq B_t^\top B_t \preceq \ell_B I$. Further, we assume $\ell_A < 1$.*

The first two requirements about the well-conditioned cost functions in Assumption 4.3.1 are standard in the literature of online optimization and control (Lin, Hu, Shi, et al., 2021; Lin, Hu, Qu, et al., 2022). For the last requirement, we additionally require $\ell_A < 1$, which implies that the system is open-loop stable. Under Assumption 4.3.1, the expected cost-to-go function is a well-conditioned function, which is important for establishing Conditions 4.2.1 and 4.2.2 (b). We state this result formally in Lemma 4.3.3.

Lemma 4.3.3. *Under Assumption 4.3.1, Condition 4.2.1 holds with $M_t = \mu_u I$. Further, conditional expectation $\mathbb{E}[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)]$ as a function of x is μ_t-strongly convex and ℓ_t-smooth for any history $\iota_t(\theta)$, where μ_t and ℓ_t are defined as following: Let $\mu_T = \mu_x$ and $\ell_T = \ell_x$,*

$$\mu_t = \mu_x + \mu_A \cdot \frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}, \text{ and } \ell_t = \ell_x + \ell_A \cdot \ell_{t+1}, \text{ for time } t = T-1, \ldots, 0. \tag{4.13}$$

As a remark, μ_t is uniformly bounded below by μ_x and ℓ_t is uniformly bounded above by $\frac{\ell_x}{1-\ell_A}$. We present a proof sketch of Lemma 4.3.3 and defer the formal proof to Section 4.D.

Starting from time step T, we know the cost-to-go $C_T^{\pi^\theta}(x;\Xi)$ equals to the terminal cost $h_t^x(x)$. It satisfies the strong convexity/smoothness directly by Assumption 4.3.1. We repeat the following induction iterations: Given $\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x;\Xi) \mid I_{t+1}(\theta)\right]$ at time $t+1$,we define an auxiliary function that adds in the disturbance residual $W_t - W_{t|t}^\theta$ and condition on the history at time t:

$$\bar{C}_{t+1}^{\pi^\theta}(x;\iota_t(\theta)) \coloneqq \mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^\theta;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right]. \tag{4.14}$$

It can be expressed as $\mathbb{E}\left[\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^\theta;\Xi) \mid I_{t+1}(\theta)\right]\middle| I_t(\theta) = \iota_t(\theta)\right]$ by the towering rule. Thus, we know function $\bar{C}_{t+1}^{\pi^\theta}$ is strongly convex and smooth in x because these properties are preserved after taking the expectation. Then, we can obtain the expected cost-to-go function $\mathbb{E}\left[C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right] = h_t^x(x) + \min_u\left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + w_{t|t}^\theta;\iota_t(\theta))\right)$. We use an existing tool called *infimal convolution* to study the optimal value of the this optimization problem as a function of x. Specifically, define an operator $\square_B$:[2]

$$\left(f \square_B \omega\right)(x) \coloneqq \min_{u\in\mathbb{R}^m}\left\{f(u) + \omega(x - Bu)\right\} \text{ for } f:\mathbb{R}^m \to \mathbb{R} \text{ and } \omega:\mathbb{R}^n \to \mathbb{R}. \tag{4.15}$$

One can show that if f and ω are well-conditioned functions, then $(f\square_B\omega)$ is also well-conditioned (see Section 4.D for the formal statement and proof). We can use this result to show the expected cost-to-go function $\mathbb{E}\left[C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right] = h_t^x(x) + (h_t^u \square_{(-B_t)} \bar{C}_{t+1}^{\pi^\theta})(A_t x + w_{t|t}^\theta;\iota_t(\theta))$, is also well-conditioned in x at time step t, which completes the induction.

To establish the second condition about the covariance of the optimal policy's action, we make the following assumption about the joint distribution of the disturbances and the predictions:

Assumption 4.3.2. *The disturbances and predictions can be grouped as pairs* $\{(W_t, V_t(\theta))\}_{t=0}^{T-1}$*, where* $(W_t, V_t(\theta))$ *is joint Gaussian, and it is independent with* $(W_{t'}, V_{t'}(\theta))$ *when* $t \neq t'$*. Further, assume that the baseline is no prediction, i.e.,* $V_t(\mathbf{0}) = 0$*. And for* $\theta \in \Theta$*, there exists* $\lambda_t(\theta) \in \mathbb{R}_{\geq 0}$ *such that* $\mathbf{Cov}[W_t] - \mathbf{Cov}[W_t \mid V_t(\theta)] \succeq \lambda_t(\theta) I$*, for any* $0 \leq t < T$.

[2] If ω takes an additional parameter w, we denote $\left(f\square_B\omega\right)(x;w) \coloneqq \min_{u\in\mathbb{R}^m}\left\{f(u) + \omega(x - Bu;w)\right\}$

Note that $\lambda_t(\theta)$ should be positive as long as $V_t(\theta)$ has some weak correlation with W_t. Under Assumption 4.3.2, we can express the optimal policy as

$$\pi_t^\theta(x; I_t(\theta)) \coloneqq \arg\min_u \left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + W_{t|t}^\theta)\right). \tag{4.16}$$

While the original definition of $\bar{C}_{t+1}^{\pi^\theta}$ in (4.14) requires the history $\iota_t(\theta)$ as an input, it no longer depends on the history under Assumption 4.3.2. We defer the proof to Section 4.D.

We can express $\pi_t^\theta(x; I_t(\theta))$ as the solution to $(h_t^u \square_{(-B_t)} \bar{C}_{t+1}^{\pi^\theta})(A_t x + W_{t|t})$. For some distributions including Gaussian, the covariance in the input of an infimal convolution will be passed through to its optimal solution. Specifically, let $u_{(f \square_B \omega)}(x)$ denote the solution to the optimization problem (4.15). When ω and f are well-conditioned, we can derive a lower bound on the trace of the covariance $\mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f \square_B \omega)}(X)\right]\right\}$ that depends on the covariance of X. Due to space limit, we defer the formal statement of this result and its proof to Lemma 4.D.2 in Section 4.D. Using this property and the observation that $\pi_t^\theta(x; I_t(\theta))$ can be expressed as $u_{(h_t^u \square_{-B_t} \bar{C}_{t+1}^{\pi^\theta})}(A_t x + W_{t|t}^\theta)$, we can directly verify that Condition 4.2.2 (b) holds with

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[\pi_t^\theta(x; I_t(\theta)) \mid \mathcal{F}_t(0)\right]\right\} \geq \sigma_t \coloneqq \frac{n\lambda_t(\theta)\mu_{t+1}^2 \cdot \mu_B}{2(\ell_u + \ell_{t+1}\sqrt{\ell_B})^2}. \tag{4.17}$$

Since Lemma 4.3.3 and (4.17) imply that Conditions 4.2.1 and 4.2.2 (b) hold with $M_t = \mu_t I$ and σ_t, respectively, we obtain a lower bound on the prediction power by Theorem 4.2.3.

Theorem 4.3.4. *In the case of LTV dynamics with well-conditioned costs, suppose Assumptions 4.3.1 and 4.3.2 hold. The prediction power of the predictor with parameter θ is lower bounded by $P(\theta) \geq \sum_{t=0}^{T-1} \mu_u \sigma_t$, where σ_t is defined in* (4.17).

As a remark, the lower bound of the prediction power in Theorem 4.3.4 shows that even weak predictions (*i.e.*, small $\lambda_t(\theta)$ in Assumption 4.3.2) can help improve the control cost compared with the no-prediction baseline. Although Assumption 4.3.2 limits $V_t(\theta)$ to be only correlated with W_t, we provide a roadmap about how to relax it so $V_t(\theta)$ can depend on all future $W_{t:T-1}$ in Section 4.D.

4.4 Related Works

Online control with predictions. Our work is closely related to the line of works that study how to use predictions in online control. Our definition of the prediction power is inspired by Yu et al. (2020): the authors first define the prediction power as the maximum control cost improvement enabled by k steps of accurate predictions about future disturbances and characterize it in a time-invariant LQR setting. Compared with Yu et al. (2020), we extend the notion of prediction power to allow general dependencies between predictions and disturbances and characterize it under more general dynamics/costs. Rather than focusing on the prediction power, many works study the power of a certain policy class such as MPC (Yu et al., 2022; Lin, Hu, Shi, et al., 2021; Zhang, Li, and Li, 2021; Lin, Hu, Qu, et al., 2022), Averaging Fixed Horizon Control (AFHC) (Chen, Agarwal, et al., 2015; Chen, Comden, et al., 2016), Receding Horizon Gradient Descent (RHGD) (Li, Qu, and Li, 2018; Li, Chen, and Li, 2019), and others (Lin, Goel, and Wierman, 2020). While one can say the power of MPC equals to the prediction power in the LQR setting (Yu et al., 2020) (generalized in Section 4.3), we show they are not the same in general (see Section 4.B).

Decision-focused learning. Our work is, in part, motivated by both empirical and theoretical findings in the decision-focused learning (DFL) literature that multiple prediction models may have the same prediction accuracy, yet their predictions can lead to very different decision costs (see Mandi et al., 2024 for a recent survey). Research on DFL typically considers predictions given as point estimates of some uncertain input to decision-makers modeled as optimization problems, such as stochastic optimization (Donti, Amos, and Kolter, 2017), linear programs (Elmachtoub and Grigas, 2022), or model predictive control (Amos et al., 2018), although more recent works have started exploring other forms of predictions such as prediction sets (Yeh et al., 2024; Wang et al., 2023). In contrast, our work does not require any particular form of decision-maker; instead, our main result characterizes the benefit of optimally leveraging predictions, for whatever form an optimal controller may take. Whereas DFL aims to design procedures for training prediction models that reduce downstream control costs, our contribution answers a more fundamental question: how much gain in performance is even possible with better predictions? We believe that it may be possible to leverage our theoretical insights about prediction power to design more general decision-focused learning algorithms in future work.

4.A Proof of Theorem 4.2.3

Since we assume x_0 is the initial state (deterministic) and π^θ is the optimal policy under the predictor with parameter θ, we have

$$\mathbb{E}\left[C_0^{\pi^\theta}(x_0;\Xi)\right] = J^{\pi^\theta}(\theta) = J^*(\theta).$$

Similarly, we also have that

$$\mathbb{E}\left[C_0^{\bar{\pi}}(x_0;\Xi)\right] = J^{\bar{\pi}}(\mathbf{0}) = J^*(\mathbf{0}).$$

Let $\{\bar{X}_{0:T}, \bar{U}_{0:T-1}\}$ be the trajectory of the baseline controller $\bar{\pi}_{0:T-1}$ under instance Ξ starting from $\bar{X}_0 = x_0$. First, we will prove by backwards induction that the difference in cumulative costs between the optimal controller π^θ and $\bar{\pi}$ has the following decomposition:

$$C_0^{\pi^\theta}(x_0;\Xi) - C_0^{\bar{\pi}}(x_0;\Xi) = \sum_{t=0}^{T-1}\left(C_t^{\pi^\theta}(\bar{X}_t;\Xi) - Q_t^{\pi^\theta}(\bar{X}_t,\bar{U}_t;\Xi)\right). \tag{4.18}$$

For the base case at time $T-1$, we apply the definition of $C_{T-1}^{\bar{\pi}}$ to get

$$C_{T-1}^{\pi^\theta}(\bar{X}_{T-1};\Xi) - C_{T-1}^{\bar{\pi}}(\bar{X}_{T-1};\Xi) = C_{T-1}^{\pi^\theta}(\bar{X}_{T-1};\Xi) - Q_{T-1}^{\pi^\theta}(\bar{X}_{T-1},\bar{U}_{T-1};\Xi).$$

For the inductive step, suppose that

$$C_{\tau+1}^{\pi^\theta}(\bar{X}_{\tau+1};\Xi) - C_{\tau+1}^{\bar{\pi}}(\bar{X}_{\tau+1};\Xi) = \sum_{t=\tau+1}^{T-1}\left(C_t^{\pi^\theta}(\bar{X}_t;\Xi) - Q_t^{\pi^\theta}(\bar{X}_t,\bar{U}_t;\Xi)\right).$$

Note that for any $t < T$,

$$Q_t^{\bar{\pi}}(\bar{X}_t,\bar{U}_t;\Xi) = Q_t^{\pi^\theta}(\bar{X}_t,\bar{U}_t;\Xi) - \left(C_{t+1}^{\pi^\theta}(\bar{X}_{t+1};\Xi) - C_{t+1}^{\bar{\pi}}(\bar{X}_{t+1};\Xi)\right).$$

Therefore,

$$\begin{aligned}
&C_\tau^{\pi^\theta}(\bar{X}_\tau;\Xi) - C_\tau^{\bar{\pi}}(\bar{X}_\tau;\Xi)\\
&= C_\tau^{\pi^\theta}(\bar{X}_\tau;\Xi) - Q_\tau^{\bar{\pi}}(\bar{X}_\tau,\bar{U}_\tau;\Xi)\\
&= C_\tau^{\pi^\theta}(\bar{X}_\tau;\Xi) - \left[Q_\tau^{\pi^\theta}(\bar{X}_\tau,\bar{U}_\tau;\Xi) - \left(C_{\tau+1}^{\pi^\theta}(\bar{X}_{\tau+1};\Xi) - C_{\tau+1}^{\bar{\pi}}(\bar{X}_{\tau+1};\Xi)\right)\right]\\
&= C_\tau^{\pi^\theta}(\bar{X}_\tau;\Xi) - Q_\tau^{\pi^\theta}(\bar{X}_\tau,\bar{U}_\tau;\Xi) + \sum_{t=\tau+1}^{T-1}\left(C_t^{\pi^\theta}(\bar{X}_t;\Xi) - Q_t^{\pi^\theta}(\bar{X}_t,\bar{U}_t;\Xi)\right)\\
&= \sum_{t=\tau}^{T-1}\left(C_t^{\pi^\theta}(\bar{X}_t;\Xi) - Q_t^{\pi^\theta}(\bar{X}_t,\bar{U}_t;\Xi)\right).
\end{aligned}$$

This completes the induction. Next, define $U_t \coloneqq \pi_t^\theta(\bar{X}_t; I_t(\theta))$. Note that U_t is $\mathcal{F}_t(\theta)$-measurable, and $\bar{U}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable and therefore also $\mathcal{F}_t(\theta)$-measurable. Because we assume the matrices $M_{0:T-1}$ satisfy Condition 4.2.1,

$$\mathbb{E}\left[C_t^{\pi^\theta}(\bar{X}_t; \Xi) \mid I_t(\theta)\right] \leq \mathbb{E}\left[Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \mid I_t(\theta)\right] - \mathrm{Tr}\left\{M_t(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right\}. \tag{4.19}$$

Let $\tilde{U}_t \coloneqq \mathbb{E}\left[U_t \mid I_t(\mathbf{0})\right]$. We see that

$$\begin{aligned}
&\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \mid I_t(\mathbf{0})\right] \\
&= \mathbb{E}\left[(\tilde{U}_t - U_t)(\tilde{U}_t - U_t)^\top \mid I_t(\mathbf{0})\right] + \mathbb{E}\left[(\tilde{U}_t - U_t)(\bar{U}_t - \tilde{U}_t)^\top \mid I_t(\mathbf{0})\right] \\
&\quad + \mathbb{E}\left[(\bar{U}_t - \tilde{U}_t)(\tilde{U}_t - U_t)^\top \mid I_t(\mathbf{0})\right] + \mathbb{E}\left[(\bar{U}_t - \tilde{U}_t)(\bar{U}_t - \tilde{U}_t)^\top \mid I_t(\mathbf{0})\right] \\
&= \mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0})\right] + \mathbb{E}\left[\tilde{U}_t - U_t \mid I_t(\mathbf{0})\right](\bar{U}_t - \tilde{U}_t)^\top \\
&\quad + (\bar{U}_t - \tilde{U}_t)\mathbb{E}\left[\tilde{U}_t - U_t \mid I_t(\mathbf{0})\right]^\top + (\bar{U}_t - \tilde{U}_t)(\bar{U}_t - \tilde{U}_t)^\top \qquad (4.20a) \\
&= \mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0})\right] + (\bar{U}_t - \tilde{U}_t)(\bar{U}_t - \tilde{U}_t)^\top, \qquad (4.20b)
\end{aligned}$$

where we use $(\bar{U}_t - \tilde{U}_t)$ is $\mathcal{F}_t(\mathbf{0})$-measurable in (4.20a); we use the definition of $\tilde{U}_t$ in (4.20b).

Applying the towering rule in (4.18) and substituting in (4.19) gives that

$$\begin{aligned}
&\mathbb{E}\left[C_0^{\pi^\theta}(x_0; \Xi) - C_0^{\bar{\pi}}(x_0; \Xi)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[C_t^{\pi^\theta}(\bar{X}_t; \Xi) - Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[C_t^{\pi^\theta}(\bar{X}_t; \Xi) \mid I_t(\theta)\right] - \mathbb{E}\left[Q_t^{\pi^\theta}(\bar{X}_t, \bar{U}_t; \Xi) \mid I_t(\theta)\right]\right] \\
&\leq -\sum_{t=0}^{T-1} \mathbb{E}\left[\mathrm{Tr}\left\{M_t(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right\}\right], \\
&= -\sum_{t=0}^{T-1} \mathrm{Tr}\left\{M_t \mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right]\right\}. \qquad (4.21)
\end{aligned}$$

If the stronger Condition 4.2.2 (a) holds, by (4.20), since $\bar{X}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable, we have

$$\begin{aligned}
\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right] &= \mathbb{E}\left[\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \mid I_t(\mathbf{0})\right]\right] \\
&\succeq \mathbb{E}\left[\mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0})\right]\right] \succeq \Sigma_t. \qquad (4.22)
\end{aligned}$$

Therefore, we can apply (4.22) in (4.21) to obtain that the first statement of Theorem 4.2.3 holds:

$$\mathbb{E}\left[C_0^{\pi^\theta}(x_0;\Xi) - C_0^{\bar{\pi}}(x_0;\Xi)\right] \leq -\sum_{t=0}^{T-1} \mathrm{Tr}\{M_t\Sigma_t\}. \tag{4.23}$$

Else, if the weaker Condition 4.2.2 (b) holds, by (4.20), since $\bar{X}_t$ is $\mathcal{F}_t(\mathbf{0})$-measurable, we have

$$\begin{aligned}\mathrm{Tr}\left\{\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right]\right\} &= \mathbb{E}\left[\mathrm{Tr}\left\{\mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top \mid I_t(\mathbf{0})\right]\right\}\right]\\ &\geq \mathbb{E}\left[\mathrm{Tr}\left\{\mathbf{Cov}\left[\pi_t^\theta(\bar{X}_t; I_t(\theta)) \mid I_t(\mathbf{0})\right]\right\}\right] \geq \sigma_t. \end{aligned} \tag{4.24}$$

Note that for any positive semi-definite matrices A, B, C such that $A \succeq C \succeq 0$, we have

$$\mathrm{Tr}\{AB\} = \mathrm{Tr}\{CB\} + \mathrm{Tr}\{(A - C)B\} \geq \mathrm{Tr}\{CB\}.$$

Since $M_t \succeq \mu_{\min}(M_t)I$, we can apply (4.24) in (4.21) to obtain that

$$\begin{aligned}\mathbb{E}\left[C_0^{\pi^\theta}(x_0;\Xi) - C_0^{\bar{\pi}}(x_0;\Xi)\right] &\leq -\sum_{t=0}^{T-1} \mathrm{Tr}\left\{\mu_{\min}(M_t)I \cdot \mathbb{E}\left[(\bar{U}_t - U_t)(\bar{U}_t - U_t)^\top\right]\right\}\\ &\leq -\sum_{t=0}^{T-1} \mu_{\min}(M_t)\sigma_t.\end{aligned}$$

4.B Examples

Instantiation of Example 4.3.2

We instantiate Example 4.3.2 with the following parameters:

$$A = \begin{bmatrix}1 & 0.1\\ 0 & 1\end{bmatrix},\ B = \begin{pmatrix}0\\ 0.1\end{pmatrix},\ Q = \begin{pmatrix}1 & \\ & 1\end{pmatrix},\ R = (1),\ \text{and } \theta := \begin{bmatrix}1 & 0.8\\ 0 & 0.6\end{bmatrix}.$$

Under different values of coefficient ρ, we train a linear regressor to predict each entry of W_t from $V_t(\theta)$ (or $V_t(I)$) over a train dataset with $64,000$ independent samples. We plot in the MSE - ρ curve on a test dataset with $16,000$ independent samples in Figure 4.2. From the plot, we see that the predictors $V_t(\theta)$ and $V_t(I)$ achieve the same MSE when predicting each entry of W_t under each $\rho \in \{0, 0.1, \ldots, 0.7\}$.

Then, we use the trained linear regressors as $W_{t|t}^\theta$ and $W_{t|t}^I$ to implement the optimal policy in Proposition 4.3.1. We plot the averaged total cost over $16,000$ trajectories with horizon $T = 100$ in Figure 4.3. From the plot, we see that the optimal policies under the predictors $V_t(\theta)$ and $V_t(I)$ achieve significantly different control costs when $\rho > 0$. We also plot the theoretical expected control cost in Figure 4.3 to verify this cost difference.

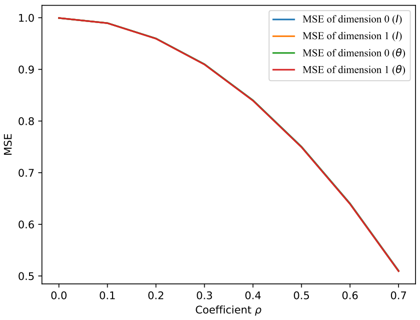

Figure 4.2: Example 4.3.2: MSE—ρ curve.

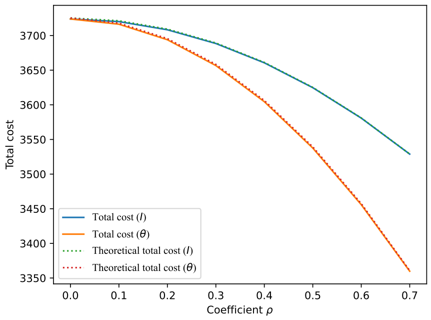

Figure 4.3: Example 4.3.2: Control cost—ρ curve.

An One-dimension Example

We also provide an example with $n = 1$, where the prediction $V_t(\theta)$ is correlated with two steps of future disturbances W_t and W_{t+1}.

Example 4.B.1. *Suppose the disturbance at each time step can be decomposed as $W_t = \sum_{i=0}^{2} W_t^{(i)}$, where the $\{W_t^{(i)}\}_{i=0}^{2}$ are independently sampled from three mean-zero distributions. We compare two predictors: $V_t(1) = \left(W_t^{(1)}, W_{t+1}^{(0)}\right)$ and $V_t(2) = P\left(W_t^{(0)} + W_t^{(1)}\right) + (A^\top - A^\top PH)PW_{t+1}^{(0)}$. They have the same prediction power when used in the control problem because*

$$\bar{u}_t^2(I_t(2)) = P\left(W_t^{(0)} + W_t^{(1)}\right) + (A^\top - A^\top PH)PW_{t+1}^{(1)} = \bar{u}_t^1(I_t(1)).$$

However, we know that $\mathcal{F}_t(1)$ is a strict super set of $\mathcal{F}_t(2)$, thus $V_t(1)$ can achieve a better MSE than $V_t(2)$ when predicting the disturbances. This is empirically verified in a 1D LQR problem with $A = B = Q = R = (1)$ and $W_t^{(i)} \overset{i.i.d.}{\sim} N(0, 1)$, as we plot in Figures 4.4. In the simulation, we train linear regressors to predict W_t and W_{t+1} with the history $I_t(1)$ or $I_t(2)$ for each time step $t < T = 100$ over a train dataset of size 160, 000*. Then, we plot the MSE - time curve on a test dataset of size* 40, 000*.*

Example: MPC can be suboptimal

We first highlight the challenge by showing that MPC can be suboptimal, i.e., only planning and optimizing based on the current information might be suboptimal when the cost functions are not quadratic.

Consider a 2-step optimal control problem (1-dimension):

$$X_1 = X_0 + U_0, \text{ and } X_2 = X_1 + U_1 + W_1.$$

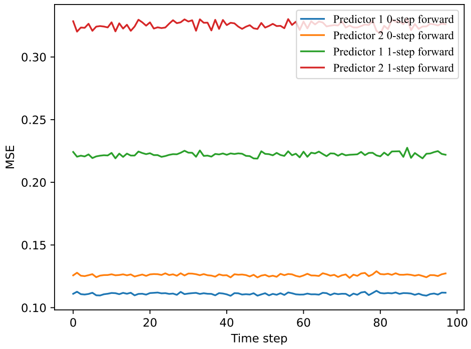

Figure 4.4: Example 4.B.1: MSE—time curve.

To construct the counterexample, we define the cost functions $h_0(x, u)$, $h_1(x, u)$, and h_2 as following:

$$h_0(x,u) = x^2 + u^2, \; h_1(x,u) = x^2 + u^2, \text{ and } h_2(x) = \begin{cases} x^2, & \text{if } x \le 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

Suppose W_1 is a random variable that satisfies $\mathbb{P}(W_1 = 1) = p$ and $\mathbb{P}(W_1 = 0) = 1-p$, where $0 < p < 1$. At time 0, we do not have any knowledge about W_1 (i.e., W_1 is independent with $I_0(\theta)$). However, at time 1, we can predict W_1 exactly, which means $\sigma(W_1) \subseteq \mathcal{F}_1(\theta)$.

Suppose the system starts at $x_0 = 0$. At time step 0, MPC (4.10) solves the optimization

$$\begin{aligned} &\min_{u_0,u_1} \mathbb{E}\left[h_0(X_0,u_0) + h_1(X_1,u_1) + h_2(X_2) \mid I_0(\theta)\right] \\ &\text{s.t. } X_0 = 0, \; X_1 = X_0 + u_0, \; X_2 = X_1 + u_1 + W_1. \end{aligned} \tag{4.25}$$

Since $I_0(\theta)$ is independent with W_1, the optimization problem can be expressed equivalently as

$$\begin{aligned} &\min_{u_0,u_1} u_0^2 + (u_0^2 + u_1^2) + \mathbb{E}\left[h_2(u_0 + u_1 + W_1)\right] \\ = &\min_{u_0,u_1} 2u_0^2 + u_1^2 + 1, \text{ s.t. } u_0 + u_1 = -1. \end{aligned}$$

The equation holds because the planned trajectory must avoid the huge cost at time step 2. Solving this gives that $u_0 = -\frac{1}{3}$. Thus, implementing MPC incurs a total cost that is at least $2u_0^2 = \frac{2}{9}$. In contrast, if one just pick $u_0 = 0$, the agent can pick u_1 based on the prediction revealed at time step 2:

$$u_1 = \begin{cases} 0 & \text{if } W_1 = 0, \\ -1 & \text{otherwise.} \end{cases}$$

In this case, the expected cost incurred is p. Thus, we can claim that MPC is not the optimal policy when $p < \frac{2}{9}$. The underlying reason that MPC is suboptimal is because it does not consider what information may be available when we make the decision in the future. In this specific example, since W_1 is revealed at time 1, we do not need to verify about the small probability event that leads to a huge loss.

We dive deeper into the reason why MPC (4.10) is optimal in the LQR setting (Section 4.3). Note that the expected optimal cost-to-go function at time step 1 is

$$\mathbb{E}\left[C_1^{\pi^\theta}(x;\Xi)\,\middle|\,I_1(\theta)\right] = \min_{u_1}\mathbb{E}\left[h_1(x,u_1) + h_2(X_2) \mid I_1(\theta)\right], \text{ s.t. } X_2 = x + u_1 + W_1. \tag{4.26}$$

Here, u_1 is $\mathcal{F}_1(\theta)$-measurable. And the true optimal policy at time 0 is decided by solving

$$\min_{u_0} h_0(x,u_0) + \mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \mid I_0(\theta)\right], \text{ s.t. } X_1 = x + u_0.$$

In general, we cannot use

$$\min_{u_1}\mathbb{E}\left[h_1(X_1,u_1) + h_2(X_2) \mid I_0(\theta)\right], \text{ s.t. } X_2 = X_1 + u_1 + W_1, \tag{4.27}$$

to replace $\mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi)\,\middle|\,I_0(\theta)\right]$ like what MPC does in (4.25) because here u_1 is $\mathcal{F}_0(\theta)$-measurable in (4.27). Recall that u_1 is $\mathcal{F}_1(\theta)$-measurable in (4.26) and $\mathcal{F}_0(\theta)$ is a subset of $\mathcal{F}_1(\theta)$. However, in the LQR setting, as the closed-form expression (4.28), the part of $\mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \mid I_0(\theta)\right]$ that depends on X_1 will not change even if $\mathcal{F}_1(\theta)$ changes. Thus, we can assume $\mathcal{F}_1(\theta) = \mathcal{F}_0(\theta)$ without affecting the optimal action at time 0. Therefore, MPC's replacement of $\mathbb{E}\left[C_1^{\pi^\theta}(X_1;\Xi) \mid I_0(\theta)\right]$ with (4.27) is valid in the LQR setting.

4.C Proofs for LTV Dynamics with Quadratic Costs

Proof of Proposition 4.3.1

To simplify notation, we introduce the shorthand

$$W_{\tau|t}^{\theta} = \mathbb{E}\left[W_\tau \mid I_t(\theta)\right].$$

We show by induction that

$$\begin{aligned}&\mathbb{E}\left[Q_t^{\pi^\theta}(x,u;\Xi) \mid I_t(\theta)\right]\\ &= \left(u + K_t x - \bar{u}_t^{\theta}(I_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t)\left(u + K_t x - \bar{u}_t^{\theta}(I_t(\theta))\right) + \psi_t^{\pi^\theta}(x; I_t(\theta)),\end{aligned}$$

$$\pi_t^\theta(x; I_t(\theta)) = -K_t x + \bar{u}_t^\theta(I_t(\theta)),$$

together with the expression of the optimal cost-to-go function

$$\mathbb{E}\left[C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta)\right] = x^\top P_t x + 2\left(\sum_{\tau=t}^{T-1} \Phi_{\tau+1,t}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top x + \Psi_t(I_t(\theta)), \quad (4.28)$$

where recall that for $t_2 > t_1$,

$$\begin{aligned}\Phi_{t_2,t_1}^\top &:= (A_{t_1} - B_{t_1}K_{t_1})^\top \cdots (A_{t_2-1} - B_{t_2-1}K_{t_2-1})^\top \\ &= (A_{t_1}^\top - A_{t_1}^\top P_{t_1+1} H_{t_1}) \cdots (A_{t_2-1}^\top - A_{t_2-1}^\top P_{t_2} H_{t_2-1}),\end{aligned}$$

and $\Psi_t(I_t(\theta))$ is a function of the history observations/predictions which does not depend on x. Note that (4.28) holds when $t = T$ because $C_T^{\pi^\theta}(x;\Xi) = x^\top P_T x$.

Suppose that (4.28) holds for $t+1$. Then, we have

$$\begin{aligned}&\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t;\Xi) \mid I_t(\theta)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t;\Xi) \mid I_{t+1}(\theta)\right] \mid I_t(\theta)\right] \\ &= \mathbb{E}\left[\left.(x+W_t)^\top P_{t+1} (x+W_t)\right| I_t(\theta)\right] + 2\mathbb{E}\left[\left.\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t+1}^\theta \right| I_t(\theta)\right]^\top x \\ &\quad + 2\mathbb{E}\left[\left.\left(\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t+1}^\theta\right)^\top W_t \right| I_t(\theta)\right] + \mathbb{E}\left[\Psi_{t+1}(I_{t+1}(\theta)) \mid I_t(\theta)\right] \\ &= x^\top P_{t+1} x + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top x + \mathrm{Tr}\{P_{t+1} \cdot \mathbf{Cov}\left[W_t \mid I_t(\theta)\right]\} \\ &\quad + 2\mathbb{E}\left[\left.\left(\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t+1}^\theta\right)^\top W_t \right| I_t(\theta)\right] + \mathbb{E}\left[\Psi_{t+1}(I_{t+1}(\theta)) \mid I_t(\theta)\right].\end{aligned}$$

To simplify the notation, let

$$\begin{aligned}\bar{\psi}_{t+1}(I_t(\theta)) :=\ & \mathrm{Tr}\{P_{t+1} \cdot \mathbf{Cov}\left[W_t \mid I_t(\theta)\right]\} + 2\mathbb{E}\left[\left.\left(\sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t+1}^\theta\right)^\top W_t \right| I_t(\theta)\right] \\ & + \mathbb{E}\left[\Psi_{t+1}(I_{t+1}(\theta)) \mid I_t(\theta)\right].\end{aligned}$$

We see that the expected Q function is given by

$$\mathbb{E}\left[Q_t^{\pi^\theta}(x,u;\Xi) \mid I_t(\theta)\right]$$

$$
\begin{aligned}
&= x^\top Q_t x + u^\top R_t u + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid I_t(\theta)\right] \\
&= x^\top Q_t x + u^\top R_t u + (A_t x + B_t u)^\top P_{t+1}(A_t x + B_t u) \\
&\quad + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top (A_t x + B_t u) + \bar{\psi}_{t+1}(I_t(\theta)) \\
&= u^\top (R_t + B_t^\top P_{t+1} B_t) u + 2\left(P_{t+1} A_t x + P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top B_t u \\
&\quad + x^\top (Q_t + A_t^\top P_{t+1} A_t) x + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top A_t x + \bar{\psi}_{t+1}(I_t(\theta)) \\
&= \left(u + K_t x - \bar{u}_t^\theta(I_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t)\left(u + K_t x - \bar{u}_t^\theta(I_t(\theta))\right) + \psi_t^{\pi^\theta}(x; I_t(\theta)),
\end{aligned}
$$

where $\psi_t^{\pi^\theta}(x; I_t(\theta))$ is given by

$$
\begin{aligned}
&x^\top (Q_t + A_t^\top P_{t+1} A_t) x + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top A_t x + \bar{\psi}_{t+1}(I_t(\theta)) \\
&\quad + \left(K_t x - \bar{u}_t^\theta(I_t(\theta))\right)^\top (R_t + B_t^\top P_{t+1} B_t)\left(K_t x - \bar{u}_t^\theta(I_t(\theta))\right).
\end{aligned}
$$

Using the expected Q function, we know that the optimal policy will pick the action

$$
\pi_t(x; I_t(\theta)) = \arg\min_u \mathbb{E}\left[Q_t^{\pi^\theta}(x, u; \Xi) \mid I_t(\theta)\right] = -K_t x + \bar{u}_t^\theta(I_t(\theta)).
$$

Therefore, we see the optimal cost-to-go function at time step t is given by

$$
\begin{aligned}
&\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right] \\
&= x^\top Q_t x + (K_t x - \bar{u}_t^\theta(I_t(\theta)))^\top R_t (K_t x - \bar{u}_t^\theta(I_t(\theta))) \\
&\quad + ((A_t - B_t K_t) x + B_t \bar{u}_t^\theta(I_t(\theta)))^\top P_{t+1}((A_t - B_t K_t) x + B_t \bar{u}_t^\theta(I_t(\theta))) \\
&\quad + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top ((A_t - B_t K_t) x + B_t \bar{u}_t^\theta(I_t(\theta))) + \bar{\psi}_{t+1}(I_t(\theta)) \\
&= x^\top (Q_t + K_t^\top R_t K_t + (A_t - B_t K_t)^\top P_{t+1}(A_t - B_t K_t)) x - 2\bar{u}_t^\theta(I_t(\theta))^\top R_t K_t x \\
&\quad + 2\bar{u}_t^\theta(I_t(\theta))^\top B_t^\top P_{t+1}(A_t - B_t K_t) x \\
&\quad + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top (A_t - B_t K_t) x \\
&\quad + \bar{u}_t^\theta(I_t(\theta))^\top (R_t + B_t^\top P_{t+1} B_t) \bar{u}_t^\theta(I_t(\theta)) \\
&\quad + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top B_t \bar{u}_t^\theta(I_t(\theta)) + \bar{\psi}_{t+1}(I_t(\theta)).
\end{aligned}
$$

Note that the term $-2\bar{u}_t^\theta(I_t(\theta))^\top R_t K_t x$ and the term $+2\bar{u}_t^\theta(I_t(\theta))^\top B_t^\top P_{t+1}(A_t - B_t K_t)x$ cancel out because $R_t K_t = B_t^\top P_{t+1}(A_t - B_t K_t)$. We also note that the matrix in the first quadratic term can be simplified to

$$\begin{aligned}
&Q_t + K_t^\top R_t K_t + (A_t - B_t K_t)^\top P_{t+1}(A_t - B_t K_t) \\
&= Q_t + K_t^\top B_t^\top P_{t+1}(A_t - B_t K_t) + (A_t - B_t K_t)^\top P_{t+1}(A_t - B_t K_t) \\
&= Q_t + A_t^\top P_{t+1}(A_t - B_t K_t) \\
&= Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} B_t K_t \\
&= Q_t + A_t^\top P_{t+1} A_t - A_t^\top P_{t+1} H_t P_{t+1} A_t \\
&= P_t,
\end{aligned}$$

where the last equation follows by the definition of P_t in (4.8).

Therefore, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta)\right] \\
&= x^\top P_t x + 2\left((A_t^\top - A_t^\top P_{t+1} H_t)(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta)\right)^\top x \\
&\quad + \bar{u}_t^\theta(I_t(\theta))^\top (R_t + B_t^\top P_{t+1} B_t)\bar{u}_t^\theta(I_t(\theta)) \\
&\quad + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top B_t \bar{u}_t^\theta(I_t(\theta)) + \bar{\psi}_{t+1}(I_t(\theta)) \\
&= x^\top P_t x + 2\left(\sum_{\tau=t}^{T-1} \Phi_{\tau,t}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top x + \bar{\psi}_t(I_t(\theta)),
\end{aligned}$$

where the residual term $\bar{\psi}_t(I_t(\theta))$ is given by

$$\begin{aligned}
\bar{\psi}_t(I_t(\theta)) &= \bar{u}_t^\theta(I_t(\theta))^\top (R_t + B_t^\top P_{t+1} B_t)\bar{u}_t^\theta(I_t(\theta)) \\
&\quad + 2\left(P_{t+1} W_{t|t}^\theta + \sum_{\tau=t+1}^{T-1} \Phi_{\tau,t+1}^\top P_{\tau+1} W_{\tau|t}^\theta\right)^\top B_t \bar{u}_t^\theta(I_t(\theta)) + \bar{\psi}_{t+1}(I_t(\theta)).
\end{aligned}$$

Thus, we have shown the statement of Proposition 4.3.1 and 4.28 by induction.

Proof of Theorem 4.3.1

By Proposition 4.3.1, we see Condition 4.2.1 holds with equality:

$$\begin{aligned}
&\mathbb{E}\left[Q_t^{\pi^\theta}(x,u;\Xi) - C_t^{\pi^\theta}(x;\Xi) \mid I_t(\theta) = \iota_t(\theta)\right] \\
&= (u - \pi_t^\theta(x;\iota_t(\theta)))^\top M_t (u - \pi_t^\theta(x;\iota_t(\theta))), \qquad (4.29)
\end{aligned}$$

where M_t is defined in Theorem 4.3.1. We also see that Condition 4.2.2 (a) holds with equality:

$$\mathbb{E}\left[\mathbf{Cov}\left[\pi_t^\theta(X; I_t(\theta)) \mid \mathcal{F}_t(\mathbf{0})\right]\right] = \Sigma_t. \tag{4.30}$$

By Theorem 4.2.3, we obtain that $P(\theta) \geq \sum_{t=0}^{T-1} \mathrm{Tr}\{M_t \Sigma_t\}$, but the lower bound is tight in this case. To see this, we go through the proof of Theorem 4.2.3 in Section 4.A and check each inequality: (4.19) holds with equality because of (4.29), which implies that (4.19) also holds with equality. (4.22) holds with equality because of (4.30) and the relationship that

$$\begin{aligned}
\bar{U}_t = & -K_t \bar{X}_t - (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} W_{\tau|t}^{\mathbf{0}} \\
= & -K_t \bar{X}_t - (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} \mathbb{E}\left[W_{\tau|t}^\theta \,\middle|\, I_t(\mathbf{0})\right] \\
= & \, \mathbb{E}\left[U_t \mid I_t(\mathbf{0})\right] = \tilde{U}_t.
\end{aligned}$$

Therefore, we know that (4.23) also holds with equality, so $P(\theta) = \sum_{t=0}^{T-1} \mathrm{Tr}\{M_t \Sigma_t\}$.

Proof of the MPC form

In the LQR setting, we can further simplify the MPC policy (4.10) to be *planning according to $w_{\tau|t}^\theta$*:

$$\begin{aligned}
\underset{u_{t:T-1}}{\arg\min} \quad & \sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T) \\
\text{s.t.} \quad & x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_{\tau|t}^\theta), \text{ for } \tau \geq t, \text{ and } x_t = x.
\end{aligned} \tag{4.31}$$

We show the MPC policies defined in (4.10) and (4.31) are equivalent to the optimal policy in Proposition 4.3.1.

To simplify the notation, we define the large vectors

$$\vec{x} := \begin{bmatrix} x_t \\ x_{t+1} \\ \vdots \\ x_T \end{bmatrix}, \ \vec{u} := \begin{bmatrix} u_t \\ u_{t+1} \\ \vdots \\ u_{T-1} \end{bmatrix}, \ \text{and } \vec{w} := \begin{bmatrix} w_t \\ w_{t+1} \\ \vdots \\ w_{T-1} \end{bmatrix}.$$

Follow the approach of system level thesis, we know the constraints that

$$x_{\tau+1} := A_\tau x_\tau + B_\tau u_\tau + w_\tau, \text{ for } \tau \geq t, \text{ and } x_t = x$$

can be expressed equivalently by the affine relationship

$$\vec{x} := \Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}.$$

Let $\vec{Q} = \mathrm{Diag}(Q_t, \ldots, Q_{T-1}, P_T)$ and $\vec{R} = \mathrm{Diag}(R_t, \ldots, R_{T-1})$. We know the objective function (with equality constraints)

$$\begin{aligned} &\sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T) \\ \text{s.t.} \quad & x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_t), \text{ for } \tau \geq t, \text{ and } x_t = x, \end{aligned} \tag{4.32}$$

can be written equivalently in the unconstrained form

$$(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w})^\top \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}) + \vec{u}^\top \vec{R} \vec{u}. \tag{4.33}$$

We introduce the notations

$$\vec{W} := \begin{bmatrix} W_t \\ W_{t+1} \\ \vdots \\ W_{T-1} \end{bmatrix}, \ \vec{W}^\theta_{\cdot|t} := \begin{bmatrix} W^\theta_{t|t} \\ W^\theta_{t+1|t} \\ \vdots \\ W^\theta_{T-1|t} \end{bmatrix}, \text{ and } \vec{w}^\theta_{\cdot|t} := \begin{bmatrix} w^\theta_{t|t} \\ w^\theta_{t+1|t} \\ \vdots \\ w^\theta_{T-1|t} \end{bmatrix}.$$

The MPC policy in (4.10) can be expressed as

$$\min_{\vec{u}} \mathbb{E}\left[(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W})^\top \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W}) + \vec{u}^\top \vec{R} \vec{u} \,\middle|\, I_t(\theta) = \iota_t(\theta) \right].$$

Because the objective function can be reduced to

$$\begin{aligned} &\mathbb{E}\left[(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W})^\top \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W}) + \vec{u}^\top \vec{R} \vec{u} \,\middle|\, I_t(\theta) = \iota_t(\theta) \right] \\ &= (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}^\theta_{\cdot|t})^\top \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{w}^\theta_{\cdot|t}) + \vec{u}^\top \vec{R} \vec{u} \\ &\quad + \mathbb{E}\left[(\Phi_w (\vec{W} - \vec{W}^\theta_{\cdot|t}))^\top \vec{Q} \Phi_w (\vec{W} - \vec{W}^\theta_{\cdot|t}) \,\middle|\, I_t(\theta) = \iota_t(\theta) \right], \end{aligned}$$

where the last term is independent with x and $\vec{u}$. Thus, the MPC policy in (4.10) is equivalent to

$$\mathbb{E}\left[(\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W})^\top \vec{Q} (\Phi_x x + \Phi_u \vec{u} + \Phi_w \vec{W}) + \vec{u}^\top \vec{R} \vec{u} \,\middle|\, I_t(\theta) = \iota_t(\theta) \right],$$

which is the MPC policy in (4.31).

Now, we show that (4.31) is equivalent to the optimal policy in Proposition 4.3.1. For any sequence $w_{t:T-1}$, let $\texttt{MPC}(x, w_{t:T-1})$ denote the first entry of the solution to

$$\arg\min_{u_{t:T-1}} \sum_{\tau=t}^{T-1} h_\tau(x_\tau, u_\tau) + h_T(x_T)$$

$$\text{s.t.} \quad x_{\tau+1} = f_\tau(x_\tau, u_\tau; w_t), \ \text{for } \tau \geq t, \ \text{and } x_t = x. \tag{4.34}$$

To show that (4.31) is equivalent to the optimal policy in Proposition 4.3.1, we only need to show that

$$\texttt{MPC}(x, w_{t:T-1}) = -K_t x - (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} w_t \tag{4.35}$$

holds for any sequence $w_{t:T-1}$. To see this, we consider the case when $w_{t:T-1}$ are deterministic disturbances on and after time step t, i.e., the agent knows $w_{t:T-1}$ exactly at time step t. In this scenario, we know the optimal policy is to follow the planned trajectory according to MPC in (4.32). On the other hand, by Proposition 4.3.1, we know the optimal action to take at time t is $-K_t x - (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} w_t$. Therefore, the first step planned by MPC must be identical with $-K_t x - (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top \sum_{\tau=t}^{T-1} \Phi_{\tau+1,t+1}^\top P_{\tau+1} w_t$. Thus, (4.35) holds. And replacing $w_{t:(T-1)}$ with $w_{t:(T-1)|t}^\theta$ finishes the proof.

Evaluation of the Expected Conditional Covariance

For two general random variables X and Y, we follow a standard procedure to evaluate the expectation of their conditional covariance $\mathbb{E}\left[\mathbf{Cov}\left[Y \mid X\right]\right]$ using a dataset $\{(x_n, y_n)\}$ that is independently sampled from the joint distribution of (X, Y) (Algorithm 5). The algorithm first train a regressor ψ that approximates the conditional expectation $\mathbb{E}\left[X \mid Y\right]$, where we use the definition:

$$\mathbb{E}\left[Y \mid X\right] = \min_{\psi \text{ is any function.}} \mathbb{E}\left[\|Y - \psi(X)\|_2^2\right].$$

Then, ψ is used for evaluating the conditional covariance. During training, we split the dataset to the train, validation, and test datasets in order to prevent overfitting.

```
Algorithm 5: Expected Conditional Covariance Estimator (ECCE)
Require: Dataset $D$ that consists input/output pair $(x_n, y_n)$.
Split the dataset $D$ to $D_{\text{train}}, D_{\text{val}}$, and $D_{\text{test}}$.
Initialize a regressor $\psi$ with input $x$ and target output $y$.
Fit $\psi$ to $D_{\text{train}}$ with MSE and use $D_{\text{val}}$ to prevent over-fit.
return $\Sigma := \frac{1}{|D_{test}|} \sum_{n \in D_{test}} (y_n - \psi(x_n))(y_n - \psi(x_n))^\top$
```

4.D Proofs for LTV Dynamics with General Costs

Infimal Convolution Properties

The first result states that the variant of infimal convolution preserves the strong convexity/smoothness of the input functions.

Lemma 4.D.1. *Consider a variant of infimal convolution where the optimization variable is multiplied by a matrix B:*

$$(f \square_B \omega)(x) = \min_u \{f(u) + \omega(x - Bu)\}, \tag{4.36}$$

where $f : \mathbb{R}^m \to \mathbb{R}$, $\omega : \mathbb{R}^n \to \mathbb{R}$, and $B \in \mathbb{R}^{n \times m}$ is a matrix. Suppose that f is a μ_f-strongly convex function, and ω is a μ_ω-strongly convex and ℓ_ω-smooth function. Then, $f \square_B \omega$ is a $\left(\frac{\mu_\omega \mu_f}{\mu_f + \|B\|^2 \mu_\omega}\right)$-strongly convex and ℓ_ω-smooth function. We also have $\nabla(f \square_B \omega)(x) = \nabla\omega(x - Bu(x))$.

The second result is about the optimal solution of the variant of infimal convolution. It states that for some distributions, the covariance on the input will induce a variance on the optimal solution. We state it in Lemma 4.D.2 and provide the proof later in this section.

Lemma 4.D.2. *Let $u_{(f \square_B \omega)}(x)$ denote the solution to the optimization problem* (4.15)*. Suppose function f is μ_f-strongly convex. Function ω is μ_ω-strongly convex and ℓ_ω-smooth. Suppose X is a random vector with bounded mean and* $\mathbf{Cov}[X] = \Sigma \succeq \sigma_0 I$*. Further, there exists a constant $C > 0$ such that for any positive integer N, X can be decomposed as $X = \sum_{i=1}^N X_i$ for i.i.d. random vectors X_i that satisfies $\mathbb{E}\left[\|X_i\|^4\right] \le C \cdot N^{-2}$. Then,*

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f \square_B \omega)}(X)\right]\right\} \ge \frac{n\sigma_0 \mu_\omega^2 \cdot \sigma_{min}(B)^2}{2(\ell_f + \ell_\omega \|B\|)^2}.$$

As a remark, examples of X that satisfies the assumptions include:

- Normal distribution $X \sim N(0, \Sigma)$. We have $X_i \sim N(0, \Sigma/N)$, thus $\mathbb{E}\left[\|X_i\|^4\right] \le 3\,\mathrm{Tr}\{\Sigma\} N^{-2}$.
- Poisson distribution (1D) with parameter a. We have $\mathbf{Var}[X] = a$ and X_i follows Poisson distribution with parameter a/N. Thus, $\mathbb{E}\left[X_i^4\right] = a^4 N^{-4}$.

The next result (Lemma 4.D.3) considers the case when there is an additional input w to function ω in the infimal convolution. When this additional parameter causes a covariance on the gradient $\nabla_1 \omega(x, W)$, the optimal solution of the infimal convolution will also have a nonzero variance.

Lemma 4.D.3. *Suppose that $\omega(x, w)$ satisfies that $\omega(\cdot, w)$ is an ℓ_ω-smooth convex function for all w. For a random variable W, suppose that the following inequality holds for arbitrary fixed vector $x \in \mathbb{R}^n$,*

$$\mathbf{Cov}\left[\nabla_1 \omega(x, W)\right] \succeq \sigma_0 I.$$

Suppose that $f : \mathbb{R}^m \to \mathbb{R}$ is a μ_f-strongly convex and ℓ_f-smooth function ($m \leq n$). Let B be a matrix in $\mathbb{R}^{n\times m}$. Then, the optimal solution of the infimal convolution

$$u_{(f\Box_B\omega)}(x, w) \coloneqq \arg\min_u \left(f(u) + \omega(x - Bu, w)\right)$$

satisfies that

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f\Box_B\omega)}(x, W)\right]\right\} \geq \frac{n\sigma_0 \cdot \sigma_{min}(B)^2}{2(\ell_f + \ell_\omega \|B\|)^2}$$

holds for arbitrary fixed vector x, where $\sigma_{min}(B)$ denotes the minimum singular value of B.

Lemma 4.D.3 is useful for showing Lemma 4.D.2. We provide its proof later in this section.

Proof of Lemma 4.3.3

We use induction to show that $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta) = \iota_t(\theta)\right]$ is a μ_t-strongly convex and ℓ_t-smooth function for any $\iota_t(\theta)$, where the coefficients μ_t and ℓ_t are defined recursively in (4.13). To simplify the notation, we will omit "$I_t(\theta) =$" in the conditional expectations throughout this proof when conditioning on a realization of the history $\iota_t(\theta)$.

Note that the statement holds for $t = T$, because $\mathbb{E}\left[C_T^{\pi^\theta}(x; \Xi) \mid \iota_T(\theta)\right] = h_T^x(x)$ and the terminal cost h_T^x is μ_x-strongly convex and ℓ_x-smooth.

Suppose the statement holds for $t + 1$. We see that

$$\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid \iota_t(\theta)\right] = h_t^x(x) + \min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid \iota_t(\theta)\right]\right).$$

By the induction assumption, we know that $\mathbb{E}\left[C_{t+1}^{\pi^\theta}(\cdot; \Xi) \mid \iota_{t+1}(\theta)\right]$ is a μ_{t+1}-strongly convex and ℓ_{t+1}-smooth function for any $\iota_{t+1}(\theta)$. Thus, $\mathbb{E}\left[C_{t+1}^{\pi^\theta}(\cdot + W_t; \Xi) \mid \iota_t(\theta)\right]$ is also a μ_{t+1}-strongly convex and ℓ_{t+1}-smooth function. Therefore,

$$\min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + B_t u + W_t; \Xi) \mid \iota_t(\theta)\right]\right)$$

is a $\frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}$-strongly convex and ℓ_{t+1}-smooth function of x by Lemma 4.D.1. By changing the variable from x to $A_t x$, we see that

$$\min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid \iota_t(\theta)\right]\right)$$

is a $\mu_A \cdot \frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}$-strongly convex and $\ell_A \cdot \ell_{t+1}$-smooth function by Assumption 4.3.1. Since h_t^x is a μ_x-strongly convex and ℓ_x-smooth function, we see that $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid \iota_t(\theta)\right]$ is also a μ_t-strongly convex and ℓ_t-smooth function because

$$\mu_t = \mu_x + \mu_A \cdot \frac{\mu_u \mu_{t+1}}{\mu_u + b^2 \mu_{t+1}}, \text{ and } \ell_t = \ell_x + \ell_A \cdot \ell_{t+1}.$$

Proof of Theorem 4.3.4

Note that the optimal action at time step t is determined by

$$\pi_t^\theta(x; I_t(\theta)) \coloneqq \arg\min_u \left(h_t^u(u) + \mathbb{E}\left[C_{t+1}^{\pi^\theta}(A_t x + B_t u + W_t; \Xi) \mid I_t(\theta)\right]\right). \tag{4.37}$$

This can be further simplified to

$$\pi_t^\theta(x; I_t(\theta)) \coloneqq \arg\min_u \left(h_t^u(u) + \bar{C}_{t+1}^{\pi^\theta}(A_t x + B_t u + W_{t|t}^\theta)\right).$$

The additional input $I_t(\theta)$ is not required for $\bar{C}_{t+1}^{\pi^\theta}$ because the function $\bar{C}_{t+1}^{\pi^\theta}(x; \iota_t(\theta))$ does not change with the history $\iota_t(\theta)$ under Assumption 4.3.2. The reason is that $W_t - W_{t|t}^\theta$ and all future predictions and disturbances $W_{t+1:T-1}, V_{t+1:T-1}^\theta$ are independent with the history $I_t(\theta)$.

By (4.16), we see that

$$\pi_t^\theta(x; I_t(\theta)) = u_{(h_t^u \square_{-B_t} \bar{C}_{t+1}^{\pi^\theta})}(A_t x + W_{t|t}^\theta).$$

Under Assumption 4.3.2, we see that

$$\mathbf{Cov}\left[W_{t|t}^\theta\right] = \mathbf{Cov}\left[W_t\right] - \mathbf{Cov}\left[W_t \mid V_t(\theta)\right] \succeq \lambda_t(\theta) I$$

and $W_{t|t}^\theta$ is Gaussian. Therefore, we can apply Lemma 4.D.2 to obtain that

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[\pi_t^\theta(x; I_t(\theta)) \mid \mathcal{F}_t(0)\right]\right\} \geq \sigma_t \coloneqq \frac{n \lambda_t(\theta) \mu_{t+1}^2 \cdot \mu_B}{2(\ell_u + \ell_{t+1}\sqrt{\ell_B})^2}.$$

Thus, Condition 4.2.2 (b) holds with σ_t.

On the other hand, Condition 4.2.1 holds with $M_t = \mu_t I$ by Lemma 4.3.3. Therefore, by Theorem 4.2.3, we obtain that $P(\theta) \geq \sum_{t=0}^{T-1} \mu_u \sigma_t$.

Proof of Lemma 4.D.1

By the definition of conjugate, we see that

$$\begin{aligned}
(f\square_B\omega)^*(y) &= \max_x \left\{\langle y,x\rangle - \min_u \{f(u)+\omega(x-Bu)\}\right\} && \text{(4.38a)}\\
&= \max_x \max_u \{\langle y,x\rangle - f(u) - \omega(x-Bu)\}\\
&= \max_x \max_u \{\langle y,x-Bu\rangle + \langle y,Bu\rangle - f(u) - \omega(x-Bu)\}\\
&= \max_u \max_x \left\{(\langle y,x-Bu\rangle - \omega(x-Bu)) + \left(\langle B^\top y,u\rangle - f(u)\right)\right\}\\
& && \text{(4.38b)}\\
&= \max_u \left\{\max_x \{\langle y,x-Bu\rangle - \omega(x-Bu)\} + \langle B^\top y,u\rangle - f(u)\right\}\\
&= \max_u \left\{\omega^*(y) + \langle B^\top y,u\rangle - f(u)\right\} && \text{(4.38c)}\\
&= \omega^*(y) + f^*(B^\top y), && \text{(4.38d)}
\end{aligned}$$

where we use the definition of $f\square_B\omega$ in (4.38a); we change the order of taking the maximum and use $\langle y,Bu\rangle = \langle B^\top y,u\rangle$ in (4.38b); we use the definition of ω^* in (4.38c); we use the definition of f^* in (4.38d).

Since $f\square_B\omega$ is convex, by Theorem 4.8 in Beck, 2017, we know that

$$(f\square_B\omega)(y) = \left(\omega^*(y) + f^*(B^\top y)\right)^*. \tag{4.39}$$

Since ω is a μ_ω-strongly convex and ℓ_ω-smooth function, we know ω^* is an $\frac{1}{\ell_\omega}$-strongly convex and $\frac{1}{\mu_\omega}$-smooth function by the conjugate correspondence theorem (Beck, 2017). Similarly, we know that f^* is a $\frac{1}{\mu_f}$-smooth convex function. Thus, we know that $\omega^*(y) + f^*(B^\top y)$ is an $\frac{1}{\ell_\omega}$-strongly convex and $\left(\frac{1}{\mu_\omega} + \frac{\|B\|^2}{\mu_f}\right)$-smooth function. Therefore, by the conjugate correspondence theorem, we know that $f\square_B\omega$ is a $\left(\frac{\mu_\omega\mu_f}{\mu_f+\|B\|^2\mu_\omega}\right)$-strongly convex and ℓ_ω-smooth function.

Now, we show that

$$\nabla(f\square_B\omega)(x) = \nabla\omega(x-Bu(x)). \tag{4.40}$$

Following a similar approach with the proof of Theorem 5.30 in Beck, 2017, we define $z = \nabla\omega(x-Bu(x))$. Define function $\phi(\xi) := (f\square_B\omega)(x+\xi) - (f\square_B\omega)(x) - \langle\xi,z\rangle$. We see that

$$\begin{aligned}
\phi(\xi) &= (f\square_B\omega)(x+\xi) - (f\square_B\omega)(x) - \langle\xi,z\rangle\\
&\le \omega(x+\xi-Bu(x)) - \omega(x-Bu(x)) - \langle\xi,z\rangle && \text{(4.41a)}
\end{aligned}$$

$$\begin{aligned}
&\leq \langle \xi, \nabla\omega(x+\xi-Bu(x))\rangle - \langle \xi, z\rangle && \text{(4.41b)}\\
&= \langle \xi, \nabla\omega(x+\xi-Bu(x)) - \nabla\omega(x-Bu(x))\rangle &&\\
&\leq \|\xi\| \cdot \|\nabla\omega(x+\xi-Bu(x)) - \nabla\omega(x-Bu(x))\| && \text{(4.41c)}\\
&\leq \ell_\omega \|\xi\|^2, && \text{(4.41d)}
\end{aligned}$$

where in (4.41a), we use

$$(f\square_B\omega)(x+\xi) \leq f(u(x)) + \omega(x+\xi-Bu(x)), \text{ and}$$
$$(f\square_B\omega)(x) = f(u(x)) + \omega(x-Bu(x));$$

we use the convexity of ω in (4.41b); we use the Cauchy-Schwarz inequality in (4.41c); we use the assumption that ω is ℓ_ω-smooth in (4.41d).

Since $(f\square_B\omega)$ is a convex function, ϕ is also convex, thus we see that

$$\phi(\xi) \geq 2\phi(0) - \phi(-\xi) = -\phi(-\xi) \geq -\ell_\omega\|\xi\|^2.$$

Combining this with (4.41), we conclude that $\lim_{\|\xi\|\to 0} |\phi(\xi)|/\|\xi\| = 0$. Thus, (4.40) holds.

Proof of Lemma 4.D.2

By Theorem 4.D.5, we see that

$$\mathbf{Cov}\left[\nabla\omega(X)\right] \geq \sigma_0\mu_\omega^2.$$

Then, we apply Lemma 4.D.3 with the second function input to the infimal convolution as $\tilde{\omega}(x,w) := \omega(x+w)$. In the context of Lemma 4.D.3, we set $W = X$, so the assumption about the covariance of the gradient holds with

$$\mathbf{Cov}\left[\nabla_1\tilde{\omega}(x,W)\right] \succeq \sigma_0\mu_\omega^2.$$

Note that for any fixed w, $\tilde{\omega}(\cdot,w)$ is μ_ω-strongly convex. Therefore, we obtain that

$$\mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f\square_B\omega)}(X)\right]\right\} = \mathrm{Tr}\left\{\mathbf{Cov}\left[u_{(f\square_B\tilde{\omega})}(0,W)\right]\right\} \geq \frac{n\sigma_0\mu_\omega^2\cdot\sigma_{\min}(B)^2}{2(\ell_f+\ell_\omega\|B\|)^2}.$$

Proof of Lemma 4.D.3

Because function c is ℓ_c-smooth, we have

$$\|\nabla c(u(x,w)) - \nabla c(u(x,w'))\| \leq \ell_c\|u(x,w) - u(x,w')\|. \tag{4.42}$$

Using the assumption that function f is ℓ_f-smooth, we obtain the following inequalities:

$$\begin{aligned}
&\left\|B^\top \nabla_1 f(x - Bu(x,w), w) - B^\top \nabla_1 f(x - Bu(x,w'), w')\right\| \\
\geq{}& \left\|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w')\right\| \\
&- \left\|B^\top \nabla_1 f(x - B \cdot u(x,w), w) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w)\right\| \\
&- \left\|B^\top \nabla_1 f(x - B \cdot u(x,w'), w') - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w')\right\| \quad (4.43a) \\
\geq{}& \left\|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w')\right\| \\
&- \ell_f \|B\| \cdot \left(\|u(x,w) - \mathbb{E}_W\left[u(x,W)\right]\| + \|u(x,w') - \mathbb{E}_W\left[u(x,W)\right]\|\right), \quad (4.43b)
\end{aligned}$$

where we use the triangle inequality in (4.43a); we use the smoothness of f in (4.43b).

Note that by the first-order optimality condition, we have

$$\nabla c(u(x,w)) - B^\top \nabla_1 f(x - B \cdot u(x,w), w) = 0.$$

Therefore, for any w, w', we have that

$$\begin{aligned}
&\nabla c(u(x,w)) - \nabla c(u(x,w')) \\
={}& B^\top \nabla_1 f(x - B \cdot u(x,w), w) - B^\top \nabla_1 f(x - B \cdot u(x,w'), w'). \quad (4.44)
\end{aligned}$$

By combining (4.44) with (4.42) and (4.43), we obtain that

$$\begin{aligned}
&\ell_c \|u(x,w) - u(x,w')\| \\
&+ \ell_f \cdot \|B\| \cdot \left(\|u(x,w) - \mathbb{E}_W\left[u(x,W)\right]\| + \|u(x,w') - \mathbb{E}_W\left[u(x,W)\right]\|\right) \\
\geq{}& \left\|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], w')\right\|
\end{aligned}$$

holds for arbitrary w and w'. Let W' be a random vector independent of W and have the same distribution. By replacing w/w' with W/W', respectively, we see

$$\begin{aligned}
&\ell_c \|u(x,W) - u(x,W')\| \\
&+ \ell_f \cdot \|B\| \cdot \left(\|u(x,W) - \mathbb{E}_W\left[u(x,W)\right]\| + \|u(x,W') - \mathbb{E}_W\left[u(x,W)\right]\|\right) \\
\geq{}& \left\|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], W) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], W')\right\|,
\end{aligned}$$

which implies

$$\begin{aligned}
&\left(\ell_c + \ell_f \|B\|\right) \left(\|u(x,W) - \mathbb{E}_W\left[u(x,W)\right]\| + \|u(x,W') - \mathbb{E}_W\left[u(x,W)\right]\|\right) \\
\geq{}& \left\|B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], W) - B^\top \nabla_1 f(x - B \cdot \mathbb{E}_W\left[u(x,W)\right], W')\right\| \\
&\qquad (4.45)
\end{aligned}$$

by the triangle inequality. Taking the square of both sides of (4.45) and applying the AM-GM inequality gives that

$$\begin{aligned}&2(\ell_c+\ell_f\|B\|)^2\|u(x,W)-\mathbb{E}_W\left[u(x,W)\right]\|^2\\&+2(\ell_c+\ell_f\|B\|)^2\|u(x,W')-\mathbb{E}_W\left[u(x,W)\right]\|^2\\\geq{}&\left\|B^\top\nabla_1 f(x-B\cdot\mathbb{E}_W\left[u(x,W)\right],W)-B^\top\nabla_1 f(x-B\cdot\mathbb{E}_W\left[u(x,W)\right],W')\right\|^2.\end{aligned}\tag{4.46}$$

Let $Y:=\nabla_1 f(x-B\cdot\mathbb{E}_W\left[u(x,W)\right],W)-\nabla_1 f(x-B\cdot\mathbb{E}_W\left[u(x,W)\right],W')$. Note that the right-hand side of (4.46) can be expressed as $\|B^\top Y\|^2=\mathrm{Tr}\{B^\top(YY^\top)B\}$. By taking the expectations of both sides, we obtain that

$$\begin{aligned}&4(\ell_c+\ell_f\|B\|)^2\,\mathrm{Tr}\{\mathbf{Cov}\left[u(x,W)\right]\}\\\geq{}&2\,\mathrm{Tr}\left\{B^\top\mathbf{Cov}\left[\nabla_1 f(x-B\mathbb{E}_W\left[u(x,W)\right],W)\right]B\right\}\\\geq{}&2n\sigma_0\sigma_{\min}(B)^2.\end{aligned}$$

In the last inequality, we use the property that the trace of a positive semi-definite matrix equals the sum of its eigenvalues. Thus, it is greater than or equal to n times the smallest eigenvalue $\sigma_0\sigma_{\min}(B)^2$. Rearranging the terms finishes the proof.

Useful Technical Results

In this section, we summarize some useful technical results about the random variables. We first state a lemma that justifies the decomposition

$$\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta))\mid\mathcal{F}_t(\mathbf{0})\right]\right]=\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi)\mid I_t(\mathbf{0})\right]\right]-\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^*(\Xi)\mid I_t(\theta)\right]\right]$$

that is used to derive the prediction power in the last step of Algorithm 4. This decomposition is helpful because otherwise, we would need to evaluate the conditional expectation inside another conditional expectation. Specifically, $\bar{u}_t^\theta(I_t(\theta))$ needs to be approximated by a learned regressor (say, ϕ) that takes $I_t(\theta)$ as an input. Then, to evaluate $\mathbb{E}\left[\mathbf{Cov}\left[\bar{u}_t^\theta(I_t(\theta))\mid\mathcal{F}_t(\mathbf{0})\right]\right]$, we would need to train another regressor to predict the output of ϕ. Our decomposition avoids this hierarchical dependence.

Lemma 4.D.4. *For any random variables X and two σ-algebras $\mathcal{F}\subseteq\mathcal{F}'$, the following equation holds*

$$\mathbb{E}\left[\mathbf{Cov}\left[\mathbb{E}\left[X\mid\mathcal{F}'\right]\mid\mathcal{F}\right]\right]=\mathbb{E}\left[\mathbf{Cov}\left[X\mid\mathcal{F}\right]\right]-\mathbb{E}\left[\mathbf{Cov}\left[X\mid\mathcal{F}'\right]\right].$$

Proof of Lemma 4.D.4. By the law of total covariance, we see that

$$\mathbf{Cov}\left[X|\mathcal{F}\right]=\mathbf{Cov}\left[\mathbb{E}\left[X\mid\mathcal{F}'\right]\mid\mathcal{F}\right]+\mathbb{E}\left[\mathbf{Cov}\left[X|\mathcal{F}'\right]|\mathcal{F}\right].$$

Taking expectation on both sides, we obtain the following equation, which is equivalent to the statement of Lemma 4.D.4:

$$\mathbb{E}\left[\mathbf{Cov}\left[X \mid \mathcal{F}\right]\right] = \mathbb{E}\left[\mathbf{Cov}\left[\mathbb{E}\left[X \mid \mathcal{F}'\right] \mid \mathcal{F}\right]\right] + \mathbb{E}\left[\mathbf{Cov}\left[X \mid \mathcal{F}'\right]\right].$$

□

We state a useful result about what functions can pass the covariance of its input to the output in Theorem 4.D.5.

Theorem 4.D.5. *Suppose that a function $g : \mathbb{R}^d \to \mathbb{R}^d$ satisfies*

$$\begin{aligned} &\langle g(x) - g(x'), x - x' \rangle \geq \gamma \|x - x'\|^2, \text{ and} \\ &\|g(x) - g(x')\| \leq L\|x - x'\|, \ \forall x, x' \in \mathbb{R}^d. \end{aligned} \tag{4.47}$$

Additionally, there exists a positive constant ℓ such that

$$-\ell I \preceq \nabla^2 g_i(x) \preceq \ell I, \ \forall x \in \mathbb{R}^d, i \in [d]. \tag{4.48}$$

Suppose X is a random vector that satisfies $|\mathbb{E}[X]| < \infty$ and $\mathbf{Cov}[X] = \Sigma \succeq \mu I$. Further, there exists a constant $C > 0$ such that for any positive integer N, X can be decomposed as $X = \sum_{i=1}^N X_i$ for i.i.d. random vectors X_i that satisfies $\mathbb{E}\left[\|X_i\|^4\right] \leq C \cdot N^{-2}$. Then, we have

$$\mathbf{Cov}\left[g(X)\right] \succeq \mu\gamma^2 I.$$

As a remark, the gradient of a well-conditioned function satisfies the conditions in (4.47).

Proof of Theorem 4.D.5. Without any loss of generality, we assume $\mathbb{E}[X] = 0$ because we can view $g(\mathbb{E}[X] + \cdot)$ as the function and subtract the mean from the random variables. The assumptions about g and X in Theorem 4.D.5 still hold.

For any $i \in [d]$ and $\epsilon \in \mathbb{R}^d$, we have the Taylor series expansion Lagrangian form (see Chapter 3.2 of Marsden and Tromba, 2003)

$$g_i(x + \epsilon) = g_i(x) + \nabla g_i(x)^\top \epsilon + \frac{1}{2}\epsilon^\top \nabla^2 g_i(\bar{x}^{(i)})\epsilon, \tag{4.49}$$

where $\bar{x}^{(i)}$ is a point on the line segment between x and $x + \epsilon$. For notational convenience, let

$$\nabla g(x) := \begin{bmatrix} \nabla g_1(x)^\top \\ \vdots \\ \nabla g_d(x)^\top \end{bmatrix} \in \mathbb{R}^{d \times d}, \text{ and } v_1(x, \epsilon) := \begin{bmatrix} \epsilon^\top \nabla^2 g_1(\bar{x}^{(1)})\epsilon \\ \vdots \\ \epsilon^\top \nabla^2 g_d(\bar{x}^{(d)})\epsilon \end{bmatrix} \in \mathbb{R}^d.$$

Using the above notation, Eq. (4.49) can be equivalently written in the following vector form:

$$g(x+\epsilon) - g(x) = \nabla g(x) \cdot \epsilon + \frac{1}{2} v_1(x, \epsilon). \tag{4.50}$$

From Eq. (4.48), we know that $|v(x,\epsilon)_i| \leq \ell \|\epsilon\|^2$, which implies

$$\|v_1(x,\epsilon)\| \leq \ell\sqrt{d}\|\epsilon\|^2. \tag{4.51}$$

In addition, by Eq. (4.47), we see that

$$\langle g(x+\epsilon) - g(x), \epsilon\rangle \geq \gamma\|\epsilon\|^2.$$

Substituting Eq. (4.50) into the above equation and rearranging the terms, we obtain

$$\epsilon^\top \cdot \nabla g(x) \cdot \epsilon \geq \gamma\|\epsilon\|^2 - \epsilon^\top \cdot v_1(x,\epsilon),$$

which is equivalent to

$$\epsilon^\top \cdot \frac{\nabla g(x) + \nabla g(x)^\top}{2} \cdot \epsilon \geq \gamma\|\epsilon\|^2 - \epsilon^\top \cdot v_1(x,\epsilon).$$

Observe that the term subtracted from the right-hand side satisfies $|\epsilon^\top \cdot v_1(x,\epsilon)| \leq \ell\sqrt{d}\|\epsilon\|^3$, which follows from Cauchy–Schwarz inequality and Eq. (4.51). Therefore, since the previous inequality holds for any $\epsilon \in \mathbb{R}^d$, taking $\epsilon \to 0$ gives that

$$\frac{\nabla g(x) + \nabla g(x)^\top}{2} \succeq \gamma I. \tag{4.52}$$

Before we proceed, we first state and prove a lemma that can convert the summation in Eq. (4.52) into a product form.

Lemma 4.D.6. *Let $M \in \mathbb{R}^{d\times d}$ be a real-valued matrix satisfying $M + M^\top \succeq 2\gamma I$. Then, for any positive definite matrix $\Sigma \succeq \mu I$, we have $M\Sigma M^\top \succeq \mu\gamma^2 I$.*

Proof of Lemma 4.D.6. Since $M + M^\top \succeq 2\gamma I$, we have for any $x \in \mathbb{R}^d$ that

$$\begin{aligned} 2\gamma\|x\|^2 &\leq 2x^\top M^\top x = 2x^\top \Sigma^{-1/2}\Sigma^{1/2}M^\top x \leq 2\|\Sigma^{-1/2}x\|\|\Sigma^{1/2}M^\top x\| \\ &\leq 2\mu^{-1/2}\|x\|\|\Sigma^{1/2}M^\top x\|, \end{aligned}$$

where the last inequality follows from $\Sigma \succeq \mu I \implies \|\Sigma^{-1/2}x\| = \sqrt{x^\top\Sigma^{-1}x} \leq \mu^{-1/2}\|x\|$. Rearranging terms, we obtain

$$\gamma\mu^{1/2}\|x\| \leq \|\Sigma^{1/2}M^\top x\|.$$

Squaring both sides concludes the proof. □

Next, we state and prove a lemma about the lower bound of the covariance induced by an additive random noise on the input that is useful when the noise is sufficiently small.

Lemma 4.D.7. *Let ε be a mean-zero random vector in $\mathbb{R}^d$ that satisfies $\underline{\delta} I \preceq \mathbf{Cov}\left[\varepsilon\right]$ and $\mathbb{E}\left[\|\varepsilon\|^4\right] \leq \overline{\gamma}$. Let g be a function that satisfies* (4.47) *and* (4.48)*. Then, for arbitrary fixed real vector $x \in \mathbb{R}^d$, we have*

$$\mathbf{Cov}\left[g(x+\varepsilon)\right] \succeq \left(\gamma^2 \underline{\delta} - 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} - \ell^2 d \overline{\gamma}\right) I.$$

Proof of Lemma 4.D.7. We first derive bounds on the i th moment of $\|\varepsilon\|$ ($i = 1, 2, 3$). By Jensen's inequality, we have

$$\mathbb{E}\left[\|\varepsilon\|^2\right] = \mathbb{E}\left[\left(\|\varepsilon\|^4\right)^{\frac{1}{2}}\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|^4\right]\right)^{\frac{1}{2}} \leq \overline{\gamma}^{\frac{1}{2}}. \tag{4.53}$$

Using Jensen'e inequality again, we obtain that

$$\mathbb{E}\left[\|\varepsilon\|\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|^2\right]\right)^{\frac{1}{2}} \leq \overline{\gamma}^{\frac{1}{4}}. \tag{4.54}$$

Lastly, by the Cauchy-Schwartz inequality, we see that

$$\mathbb{E}\left[\|\varepsilon\|^3\right] \leq \left(\mathbb{E}\left[\|\varepsilon\|^4\right] \cdot \mathbb{E}\left[\|\varepsilon\|^2\right]\right)^{\frac{1}{2}} \leq \overline{\gamma}^{\frac{3}{4}}. \tag{4.55}$$

Note that by (4.50), we have

$$\mathbf{Cov}\left[g(x+\varepsilon)\right] = \mathbf{Cov}\left[g(x+\varepsilon) - g(x)\right] = \mathbf{Cov}\left[\nabla g(x) \cdot \varepsilon + \frac{1}{2} v_1(x,\varepsilon)\right]. \tag{4.56}$$

Since $\mathbb{E}\left[\varepsilon\right] = 0$, we can further decompose (4.56) as

$$\begin{aligned}
&\mathbf{Cov}\left[\nabla g(x) \cdot \varepsilon + \frac{1}{2} v_1(x,\varepsilon)\right] \\
&= \mathbb{E}\Bigg[\left(\nabla g(x) \cdot \varepsilon + \frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right) \cdot \\
&\quad \left(\nabla g(x) \cdot \varepsilon + \frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^{\top}\Bigg] \\
&= \nabla g(x) \cdot \mathbf{Cov}\left[\varepsilon\right] \cdot \nabla g(x)^{\top} + \nabla g(x) \cdot \mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^{\top}\right] \\
&\quad + \mathbb{E}\left[\left(\frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right) \cdot \varepsilon^{\top}\right] \cdot \nabla g(x)^{\top} + \frac{1}{4}\mathbf{Cov}\left[v_1(x,\varepsilon)\right].
\end{aligned} \tag{4.57}$$

By Lemma 4.D.6 and (4.52), we know that the first term in (4.57) can be lower bounded by

$$\nabla g(x) \cdot \mathbf{Cov}\left[\varepsilon\right] \cdot \nabla g(x)^\top \succeq \gamma^2 \underline{\delta} I. \tag{4.58}$$

Define the residual term as the sum of the last 3 terms in (4.57):

$$\begin{aligned} R \coloneqq\ & \nabla g(x) \cdot \mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right] \\ & + \mathbb{E}\left[\left(\frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right) \cdot \varepsilon^\top\right] \cdot \nabla g(x)^\top + \frac{1}{4}\mathbf{Cov}\left[v_1(x,\varepsilon)\right]. \end{aligned} \tag{4.59}$$

To show Lemma 4.D.7, we only need to show

$$\|R\| \leq 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} + \ell^2 d\overline{\gamma}. \tag{4.60}$$

To see this, note that

$$\begin{aligned} & \left\|\nabla g(x) \cdot \mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]\right\| \\ \leq\ & \|\nabla g(x)\| \cdot \left\|\mathbb{E}\left[\varepsilon \cdot \left(\frac{1}{2} v_1(x,\varepsilon) - \frac{1}{2}\mathbb{E}\left[v_1(x,\varepsilon)\right]\right)^\top\right]\right\| && \text{(4.61a)} \\ \leq\ & \frac{L}{2}\left(\left\|\mathbb{E}\left[\varepsilon \cdot v_1(x,\varepsilon)^\top\right]\right\| + \left\|\mathbb{E}\left[\varepsilon\right] \cdot \mathbb{E}\left[v_1(x,\varepsilon)\right]^\top\right\|\right) && \text{(4.61b)} \\ \leq\ & \frac{L}{2}\left(\mathbb{E}\left[\left\|\varepsilon \cdot v_1(x,\varepsilon)^\top\right\|\right] + \|\mathbb{E}\left[\varepsilon\right]\| \cdot \|\mathbb{E}\left[v_1(x,\varepsilon)\right]\|\right) && \text{(4.61c)} \\ \leq\ & \frac{L}{2}\left(\mathbb{E}\left[\|\varepsilon\| \cdot \|v_1(x,\varepsilon)\|\right] + \mathbb{E}\left[\|\varepsilon\|\right] \cdot \mathbb{E}\left[\|v_1(x,\varepsilon)\|\right]\right) && \text{(4.61d)} \\ \leq\ & \frac{L\ell\sqrt{d}}{2} \cdot \left(\mathbb{E}\left[\|\varepsilon\|^3\right] + \mathbb{E}\left[\|\varepsilon\|\right] \cdot \mathbb{E}\left[\|\varepsilon\|^2\right]\right) && \text{(4.61e)} \\ \leq\ & L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}}, && \text{(4.61f)} \end{aligned}$$

where we use the definition of the induced matrix norm in (4.61a); we use (4.47) and the triangle inequality in (4.61b); we use the Jensen's inequality and the definition of the induced matrix norm in (4.61c) and (4.61d); we use (4.51) in (4.61e); we use the bounds on the moments of $\|\varepsilon\|$ (4.53), (4.54), and (4.55) in (4.61f).

On the other hand, we know that $\mathbf{Cov}\left[v_1(x,\varepsilon)\right]$ is a positive semi-definite matrix that satisfies

$$\begin{aligned} \mathbf{Cov}\left[v_1(x,\varepsilon)\right] &= \mathbb{E}\left[v_1(x,\varepsilon) v_1(x,\varepsilon)^\top\right] - \mathbb{E}\left[v_1(x,\varepsilon)\right] \cdot \mathbb{E}\left[v_1(x,\varepsilon)\right]^\top \\ &\preceq \mathbb{E}\left[v_1(x,\varepsilon) v_1(x,\varepsilon)^\top\right]. \end{aligned}$$

Therefore, we see that its induced matrix norm can be upper bounded by the expectation of squared norm:

$$\begin{aligned}\|\mathbf{Cov}\left[v_1(x,\varepsilon)\right]\| &\leq \left\|\mathbb{E}\left[v_1(x,\varepsilon)v_1(x,\varepsilon)^\top\right]\right\| \leq \mathbb{E}\left[\left\|v_1(x,\varepsilon)v_1(x,\varepsilon)^\top\right\|\right] \\ &\leq \mathbb{E}\left[\|v_1(x,\varepsilon)\|^2\right].\end{aligned}$$

Using the bound of $\|v_1(x,\varepsilon)\|$ in (4.51) and the 4 th moment bound of $\|\varepsilon\|$, we obtain that

$$\|\mathbf{Cov}\left[v_1(x,\varepsilon)\right]\| \leq \ell^2 d\mathbb{E}\left[\|\varepsilon\|^4\right] \leq \ell^2 d\overline{\gamma}. \tag{4.62}$$

Note that the norm of R (Equation (4.60)) can be upper bounded by the sum of the norms of the 3 separate terms. Thus, by combining the (4.61) and (4.62), we see that (4.60) holds. □

Lastly, we consider the case when the input of g can be expressed as the sum of a sequence of mutual independent random vectors.

Lemma 4.D.8. *Let $\{X_i\}_{1\leq i\leq N}$ be a sequence of mean-zero random vectors in $\mathbb{R}^d$ that are mutually independent and satisfies $\underline{\delta} I \preceq \mathbf{Cov}\left[X_i\right]$ and $\mathbb{E}\left[\|X_i\|^4\right] \leq \overline{\gamma}$. Let g be a function that satisfies* (4.47) *and* (4.48)*. Then, for any positive integer N, we have*

$$\mathbf{Cov}\left[g\left(\sum_{i=1}^N X_i\right)\right] \succeq N\left(\gamma^2\underline{\delta} - 2L\ell d^2 \cdot \overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right) I. \tag{4.63}$$

Proof of Lemma 4.D.8. We use an induction norm on N to show that (4.63) holds.

When $N = 1$, (4.63) holds by setting $x = 0$ and $\varepsilon = X_1$ in Lemma 4.D.7.

Suppose (4.63) holds for $N-1$. Then, for N, by the law of total variance, we see that

$$\mathbf{Cov}\left[g\left(\sum_{i=1}^N X_i\right)\right] = \mathbf{Cov}\left[\mathbb{E}\left[g\left(\sum_{i=1}^N X_i\right)\middle|\sum_{i=1}^{N-1} X_i\right]\right] + \mathbb{E}\left[\mathbf{Cov}\left[g\left(\sum_{i=1}^N X_i\right)\middle|\sum_{i=1}^{N-1} X_i\right]\right]. \tag{4.64}$$

For the first term in (4.64), we define a new function

$$\bar{g}(x) \coloneqq \mathbb{E}\left[g(x + X_N)\right].$$

Since the random variables $\{X_i\}_{1\leq i\leq N}$ are mutually independent, we observe that the conditional expectation can be written as

$$\mathbb{E}\left[g\left(\sum_{i=1}^{N} X_i\right)\middle|\sum_{i=1}^{N-1} X_i\right] = \bar{g}\left[\sum_{i=1}^{N-1} X_i\right].$$

One can verify that if g satisfies the conditions in (4.47) and (4.48), then $\bar{g}$ also satisfies the same conditions as g because

$$\begin{aligned}\|\bar{g}(x) - \bar{g}(x')\| = \|\mathbb{E}\left[g(x+X_N) - g(x'+X_N)\right]\| &\leq \mathbb{E}\left[\|g(x+X_N) - g(x'+X_N)\|\right]\\ &\leq L\|x-x'\|.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\langle\bar{g}(x) - \bar{g}(x'), x-x'\rangle &= \langle\mathbb{E}\left[g(x+X_N) - g(x'+X_N)\right], x-x'\rangle\\ &= \mathbb{E}\left[\langle g(x+X_N) - g(x'+X_N), x-x'\rangle\right] \geq \gamma\|x-x'\|^2.\end{aligned}$$

For the Hessian upper/lower bounds, because $\nabla^2\bar{g}_i(x) = \nabla^2\mathbb{E}\left[g_i(x+X_N)\right] = \mathbb{E}\left[\nabla^2 g_i(x+X_N)\right]$,

$$-\ell I \preceq \bar{g}_i(x) \preceq \ell I.$$

Therefore, by the induction assumption, we see that

$$\begin{aligned}\textbf{Cov}\left[\mathbb{E}\left[g\left(\sum_{i=1}^{N} X_i\right)\middle|\sum_{i=1}^{N-1} X_i\right]\right] &= \textbf{Cov}\left[\bar{g}\left[\sum_{i=1}^{N-1} X_i\right]\right]\\ &\succeq (N-1)\left(\gamma^2\underline{\delta} - 2L\ell d^2\cdot\overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right)I. \qquad (4.65)\end{aligned}$$

For the second term in (4.64), we note that for any realization x of $\sum_{i=1}^{N-1} X_i$, we have

$$\begin{aligned}\textbf{Cov}\left[g\left(\sum_{i=1}^{N} X_i\right)\middle|\sum_{i=1}^{N-1} X_i = x\right] = \textbf{Cov}\left[g\left(x+X_N\right)\middle|\sum_{i=1}^{N-1} X_i = x\right] &= \textbf{Cov}\left[g(x+X_N)\right]\\ \succeq \left(\gamma^2\underline{\delta} - 2L\ell d^2\cdot\overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right)I,&\end{aligned}$$

where the conditioning can be removed in the second step because the random variables $\{X_i\}_{1\leq i\leq N}$ are mutually independent, so $g(x+X_N)$ is independent with $\sum_{i=1}^{N-1} X_i$; and we use Lemma 4.D.7 in the last inequality. Therefore, we obtain that

$$\mathbb{E}\left[\textbf{Cov}\left[g\left(\sum_{i=1}^{N} X_i\right)\middle|\sum_{i=1}^{N-1} X_i\right]\right] \succeq \left(\gamma^2\underline{\delta} - 2L\ell d^2\cdot\overline{\gamma}^{\frac{3}{4}} - \ell^2 d\overline{\gamma}\right)I. \qquad (4.66)$$

Substituting (4.65) and (4.66) into (4.64) shows that (4.63) still holds for N. Thus, we have proved Lemma 4.D.8 by induction. □

Now we come back to the proof of Theorem 4.D.5. By the assumption, we know the distribution of X is identical with the distribution of $\sum_{i=1}^{N} X_i$, where X_i are i.i.d. random vectors that satisfies $\mathbb{E}\left[\|X_i\|^4\right] \leq C \cdot N^{-2}$. Thus, we have

$$\mathbf{Cov}\left[g(X)\right] = \mathbf{Cov}\left[g\left(\sum_{i=1}^{N} X_i\right)\right].$$

Note that each X_i satisfies that $\mathbf{Cov}\left[X_i\right] = \frac{1}{N}\mathbf{Cov}\left[X\right] \succeq \frac{\mu}{N} I$. Applying Lemma 4.D.8 gives that

$$\mathbf{Cov}\left[g(X)\right] \succeq \left(\mu\gamma^2 - \frac{C^{3/4}}{\sqrt{N}} \cdot 2L\ell d^2 - \frac{C}{N} \cdot \ell^2 d\overline{\gamma}\right) \cdot I.$$

By letting N tends to infinity in the above inequality, we finishes the proof of Theorem 4.D.5. □

Roadmap to Multi-step Prediction under Well-Conditioned Costs

A limitation of Assumption 4.3.2 in Section 4.3 is that it only allows the prediction $V_t(\theta)$ to depend on the disturbance W_t at time step t. A natural question is whether we can relax the assumption by allowing $V_t(\theta)$ to depend on all future disturbances $W_{t:(T-1)}$. In this section, we present a roadmap towards this generalization and discuss about the potential challenges.

First, we show that the expected cost-to-go function $\mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right]$ can be expressed as a function that only depends on the conditional expectations $W_{\tau|t}^{\theta}$ for all $\tau \geq t$, i.e., there exists a function $\tilde{C}_t^{\pi^\theta}$ that satisfies

$$\tilde{C}_t^{\pi^\theta}(x; W_{t:(T-1)|t}^{\theta}) = \mathbb{E}\left[C_t^{\pi^\theta}(x; \Xi) \mid I_t(\theta)\right]. \tag{4.67}$$

We show (4.67) by induction on $t = T, T-1, \ldots, 0$. Note that the statement holds for T. Suppose it holds for $t+1$, by (4.14), we have

$$\begin{aligned}\bar{C}_{t+1}^{\pi^\theta}(x; I_t(\theta)) &= \mathbb{E}\left[C_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^{\theta}; \Xi) \mid I_t(\theta)\right]\\ &= \mathbb{E}\left[\tilde{C}_{t+1}^{\pi^\theta}(x + W_t - W_{t|t}^{\theta}; W_{(t+1):(T-1)|t+1}^{\theta}) \,\middle|\, I_t(\theta)\right],\end{aligned}$$

where we use the induction assumption in the last equation. Define the random variables $\varepsilon_{t|t}^{\theta} \coloneqq W_t - W_{t|t}^{\theta}$ and $\varepsilon_{\tau|t}^{\theta} \coloneqq W_{\tau|(t+1)}^{\theta} - W_{\tau|t}^{\theta}$. Using the properties of joint Gaussian distribution, we know that $\varepsilon_{t:(T-1)|t}^{\theta}$ are independent with $I_t(\theta)$. Therefore,

$$\bar{C}_{t+1}^{\pi^\theta}(x; I_t(\theta)) = \mathbb{E}\left[\tilde{C}_{t+1}^{\pi^\theta}(x + \varepsilon_{t|t}^{\theta}; W_{(t+1):(T-1)|t}^{\theta} + \varepsilon_{(t+1):(T-1)|t}^{\theta}) \,\middle|\, I_t(\theta)\right]$$

$$= \mathbb{E}_{\varepsilon^{\theta}_{t:(T-1)|t}} \left[\tilde{C}^{\pi^{\theta}}_{t+1}(x + \varepsilon^{\theta}_{t|t}; W^{\theta}_{(t+1):(T-1)|t} + \varepsilon^{\theta}_{(t+1):(T-1)|t}) \right].$$

Thus, $\bar{C}^{\pi^{\theta}}_{t+1}(x; I_t(\theta))$ can be expressed as a function of x and $W^{\theta}_{(t+1):(T-1)|t}$, and we denote it as

$$\tilde{\bar{C}}_{t+1}(x; W^{\theta}_{(t+1):(T-1)|t}) \coloneqq \bar{C}^{\pi^{\theta}}_{t+1}(x; I_t(\theta)). \tag{4.68}$$

Therefore, we obtain that

$$\begin{aligned} \mathbb{E}\left[C^{\pi^{\theta}}_{t}(x; \Xi) \mid I_t(\theta) \right] &= h^x_t(x) + (h^u_t \square_{(-B_t)} \bar{C}^{\pi^{\theta}}_{t+1})(A_t x + W^{\theta}_{t|t}; I_t(\theta)) \\ &= h^x_t(x) + (h^u_t \square_{(-B_t)} \tilde{\bar{C}}^{\pi^{\theta}}_{t+1})(A_t x + W^{\theta}_{t|t}; W^{\theta}_{(t+1):(T-1)|t}). \end{aligned}$$

Therefore, $\mathbb{E}\left[C^{\pi^{\theta}}_{t}(x; \Xi) \mid I_t(\theta) \right]$ can also be expressed in the form $\tilde{C}^{\pi^{\theta}}_{t}(x; W^{\theta}_{t:(T-1)|t})$. Thus, we have shown (4.67) by induction, with (4.68) as an intermediate result.

Note that the optimal policy is given by

$$\begin{aligned} \pi^{\theta}_t(x; I_t(\theta)) &\coloneqq \arg\min_{u} \left(h^u_t(u) + \bar{C}^{\pi^{\theta}}_{t+1}(A_t x + B_t u + W^{\theta}_{t|t}; I_t(\theta)) \right) \\ &= \arg\min_{u} \left(h^u_t(u) + \tilde{\bar{C}}^{\pi^{\theta}}_{t+1}(A_t x + B_t u + W^{\theta}_{t|t}; W^{\theta}_{(t+1):(T-1)|t}) \right) \\ &= u_{(h^u_t \square_{-B_t} \tilde{\bar{C}}^{\pi^{\theta}}_{t+1})}(A_t x + W^{\theta}_{t|t}; W^{\theta}_{(t+1):(T-1)|t}). \end{aligned}$$

Therefore, by Lemma 4.D.3, we need to establish a covariance lower bound of the gradient

$$\nabla_x \tilde{\bar{C}}^{\pi^{\theta}}_{t+1}(x + W^{\theta}_{t|t}; W^{\theta}_{(t+1):(T-1)|t})$$

in order to derive a lower bound for the trace of the covariance matrix of $\pi^{\theta}_t(x; I_t(\theta))$. While this is relatively straightforward when we only have $W^{\theta}_{t|t}$ because it is added directly with x, it is much more challenging to also consider the covariance caused by $W^{\theta}_{(t+1):(T-1)|t}$. This is because they affect $\tilde{\bar{C}}^{\pi^{\theta}}_{t+1}$ through multiple steps of infimal convolutions. Nevertheless, we feel the approach that we describe here is promising if we can derive more properties that are preserved through the infimal convolution operators. We leave this direction as future work.

Part III

Policy Optimization

Chapter 5

SINGLE-TRAJECTORY ONLINE POLICY OPTIMIZATION

The results on prediction power in Part II provide strong motivations for studying policy optimization. First, a standard predictive policy cannot achieve near-optimal performance in all scenarios. For example, as prediction quality gets worse, we should not stick to the standard MPC policy, which trusts the given predictions completely when solving the predictive optimization problem. Second, even if the optimal predictive policy has a closed-form solution (e.g., Proposition 4.3.1 in Chapter 4), it may have complicated dependence on distributions and parameters, making it intractable to solve in practice. These challenges motivates us to study the problem of finding/tracking the optimal policy with limited feedback/observations online.

In this chapter, we study online policy optimization with time-varying costs and dynamics, which allows general policy classes that include predictive policies as a special case. We identify a critical property called *contractive perturbation* that makes the problem tractable and generalizes many existing results. When the Jacobians of the dynamics are known, we develop the Memoryless Gradient-based Adaptive Policy Selection (M-GAPS) algorithm together with an analytical framework that connects it with classic online optimization. When the Jacobians are unknown, we propose a meta-framework that can combine M-GAPS with an online estimator of the dynamical model. We demonstrate the effectiveness of M-GAPS by applying it in quadcopter control.

The results in this chapter are based on the following papers:

[Lin, Preiss, Anand, et al., 2023] Lin, Yiheng, James A. Preiss, Emile Anand, Yingying Li, Yisong Yue, and Adam Wierman. “Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations.” Advances in Neural Information Processing Systems 36 (2023): 53508-53521.

[Lin, Preiss, Xie, et al., 2024] Lin, Yiheng, James A. Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. “Online policy optimization in unknown nonlinear systems.” In The Thirty Seventh Annual Conference on Learning Theory, pp. 3475-3522. Proceedings of Machine Learning Research, 2024.

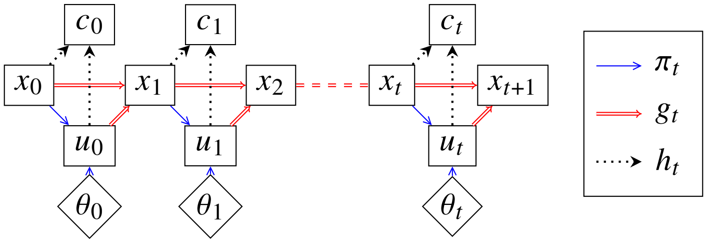

Figure 5.1: Diagram of the causal relationships between states, policy parameters, control inputs, and costs.

[Preiss et al., 2025] Preiss, James A., Fengze Xie, Yiheng Lin, Adam Wierman, and Yisong Yue. "Fast non-episodic adaptive tuning of robot controllers with model-based online policy optimization." Under submission.

5.1 Problem Setting

We consider online policy selection on a single trajectory. The setting is a discrete-time dynamical system with state $x_t \in \mathbb{R}^n$ for time index $t \in \mathcal{T} := [0 : T-1]$. At time step $t \in \mathcal{T}$, the policy picks a control action $u_t \in \mathbb{R}^m$, and the next state and the incurred cost are given by:

$$\text{Dynamics: } x_{t+1} = g_t(x_t, u_t), \qquad \text{Cost: } c_t := h_t(x_t, u_t),$$

respectively, where $g_t(\cdot,\cdot)$ is a time-varying dynamics function and $h_t(\cdot,\cdot)$ is a time-varying stage cost. The goal is to minimize the total cost $\sum_{t=0}^{T-1} c_t$.

We consider parameterized time-varying policies of the form of $u_t = \pi_t(x_t, \theta_t)$, where x_t is the current state at time step t and $\theta_t \in \Theta$ is the current policy parameter. Θ is a closed convex subset of $\mathbb{R}^d$. We assume the dynamics, cost, and policy functions $\{g_t, h_t, \pi_t\}_{t\in\mathcal{T}}$ are oblivious, meaning they are fixed before the game begins. The online policy selection algorithm optimizes the total cost by selecting θ_t sequentially. We illustrate how the policy parameter sequence $\theta_{0:T-1}$ affects the trajectory $\{x_t, u_t\}_{t\in\mathcal{T}}$ and per-step costs $c_{0:T-1}$ in Figure 5.1. The online algorithm has access to the partial derivatives of the dynamics g_t and cost h_t *along the visited trajectory*, but does not have oracle access to the g_t, h_t for arbitrary states and actions.

We provide two motivating examples for our setting. The first example is MPC with confidence coefficients, a generalization of Li, Qu, and Li (2021).

Example 5.1.1 (MPC with Confidence Coefficients)**.** *Consider a linear time-varying (LTV) system $g_t(x_t, u_t) = A_t x_t + B_t u_t + w_t$, with time-varying costs $h_t(x_t, u_t) = q(x_t, Q_t) + q(u_t, R_t)$. At time t, the policy observes*

$$\{A_{t:t+k-1}, B_{t:t+k-1}, Q_{t:t+k-1}, R_{t:t+k-1}, w_{t:t+k-1|t}\},$$

where $w_{\tau|t}$ is a (noisy) prediction of the future disturbance w_τ. Then, $\pi_t(x_t, \theta_t)$ commits the first entry of

$$\begin{aligned} &\operatorname*{arg\,min}_{u_{t:t+k-1|t}} \sum_{\tau=t}^{t+k-1} h_\tau(x_{\tau|t}, u_{\tau|t}) + q(x_{t+k|t}, \tilde{Q}) \\ &\text{s. t. } x_{t|t} = x_t, \quad x_{\tau+1|t} = A_\tau x_{\tau|t} + B_\tau u_{\tau|t} + \lambda_t^{[\tau-t]} w_{\tau|t} : \; t \le \tau < t+k, \end{aligned} \tag{5.1}$$

where $\theta_t = \left(\lambda_t^{[0]}, \lambda_t^{[1]}, \ldots, \lambda_t^{[k-1]}\right), \Theta = [0,1]^k$ and $\tilde{Q}$ is a fixed positive-definite matrix. Intuitively, $\lambda_t^{[i]}$ represents our level of confidence in the disturbance prediction i steps into the future at time step t, with entry 1 *being fully confident and* 0 *being not confident at all.*

The second example studies a nonlinear control model motivated by Li, Yang, Qu, Lin, et al., 2023; Qu, Yu, et al., 2021.

Example 5.1.2 (Linear Feedback Control in Nonlinear Systems)**.** *Consider a time-varying nonlinear control problem with dynamics $g_t(x_t, u_t) = Ax_t + Bu_t + \delta_t(x_t, u_t)$ and costs $h_t(x_t, u_t) = q(x_t, Q) + q(u_t, R)$. Here, the nonlinear residual δ_t comes from linearization and is assumed to be sufficiently small and Lipschitz. Inspired by Qu, Yu, et al., 2021, we construct an online policy based on the optimal controller $u_t = -\bar{K}x_t$ for the linear-quadratic regulator* $\mathrm{LQR}(A, B, Q, R)$*. Specifically, we let $\pi_t(x_t, \theta_t) = -K(\theta_t)x_t$ where K is a mapping from Θ to $\mathbb{R}^{n\times m}$ such that $\left\|K(\theta_t) - \bar{K}\right\|$ is uniformly bounded.*

Policy Class and Performance Metrics

In our setting, the state x_t at time t is uniquely determined by the combination of 1) a state x_τ at a previous time $\tau < t$, and 2) the parameter sequence $\theta_{\tau:t-1}$. Similarly, the cost at time t is uniquely determined by x_τ and $\theta_{\tau:t}$. Since we use these properties often, we introduce the following notation.

Definition 5.1.1 (Multi-Step Dynamics and Cost)**.** *The multi-step dynamics $g_{t|\tau}$ between two time steps $\tau \le t$ specifies the state x_t as a function of the previous state x_τ and previous policy parameters $\theta_{\tau:t-1}$. It is defined recursively, with the base case $g_{\tau|\tau}(x_\tau) \coloneqq x_\tau$ and the recursive case*

$$g_{t+1|\tau}(x_\tau, \theta_{\tau:t}) = g_t\left(z_t, \pi_t\left(z_t, \theta_t\right)\right), \; \forall t \ge \tau,$$

in which $z_t \coloneqq g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$.[1] *The multi-step cost* $h_{t|\tau}$ *specifies the cost* c_t *as function of* x_τ *and* $\theta_{\tau:t}$. *It is defined as* $h_{t|\tau}(x_\tau, \theta_{\tau:t}) \coloneqq h_t\left(z_t, \pi_t\left(z_t, \theta_t\right)\right)$.

In this paper, we frequently compare the trajectory of our algorithm against the trajectory that would arise from applying a fixed parameter θ since time step 0, which we denote as $\hat{x}_t(\theta) \coloneqq g_{t|0}(x_0, \theta_{\times t})$ and $\hat{u}_t(\theta) \coloneqq \pi_t(\hat{x}_t(\theta), \theta)$. A related concept that is heavily used is the *surrogate cost* F_t, which maps a single policy parameter to a real number.

Definition 5.1.2 (Surrogate Cost)**.** *The surrogate cost function is defined as* $F_t(\theta) \coloneqq h_t(\hat{x}_t(\theta), \hat{u}_t(\theta))$.

Figure 5.1 shows the overall causal structure, from which these concepts follow.

To measure the performance of an online algorithm, we adopt the objective of ***adaptive policy regret***, which has been used by Hazan and Seshadhri (2007) and Gradu, Hazan, and Minasyan (2023). It is a stronger benchmark than the static policy regret (Agarwal et al., 2019; Chen and Hazan, 2021) and is more suited to time-varying environments. We use $\{x_t, u_t, \theta_t\}_{t \in \mathcal{T}}$ to denote the trajectory of the online algorithm throughout the paper. The adaptive policy regret $R^A(T)$ is defined as the maximum difference between the cost of the online policy and the cost of the optimal fixed-parameter policy over any sub-interval of the whole horizon $\mathcal{T}$, i.e.,

$$R^A(T) \coloneqq \max_{I=[t_1:t_2]\subseteq\mathcal{T}} \left(\textstyle\sum_{t\in I} h_t(x_t, u_t) - \inf_{\theta\in\Theta} \sum_{t\in I} F_t(\theta)\right). \tag{5.2}$$

In contrast, the (static) policy regret defined in Chen and Hazan, 2021; Agarwal et al., 2019 restricts the time interval I to be the whole horizon $\mathcal{T}$. Thus, a bound on adaptive regret is strictly stronger than the same bound on static regret. Adaptive regret is particularly useful in time-varying environments like Examples 5.1.1 and 5.1.2 because an online algorithm must adapt quickly to compete against a comparator policy parameter that can change indefinitely with every time interval (Hazan, 2016, Section 10.2).

In the general case when surrogate costs $F_{0:T-1}$ are nonconvex, it is difficult (if not impossible) for online algorithms to achieve meaningful guarantees on classic regret metrics like $R^A(T)$ or static policy regret because they do not have oracle optimization solvers or even the exact knowledge of the surrogate costs. Therefore,

[1] z_t is an auxiliary variable to denote the state at t under initial state x_τ and parameters $\theta_{\tau:t}$.

we introduce the metric of ***local regret***, which bounds the sum of squared gradient norms over the whole horizon:

$$R^L(T) \coloneqq \textstyle\sum_{t=0}^{T-1} \|\nabla F_t(\theta_t)\|^2. \tag{5.3}$$

Similar metrics have been adopted by previous works on online nonconvex optimization (Hazan, Singh, and Zhang, 2017). Intuitively, $R^L(T)$ measures how well the online agent chases the (changing) stationary point of the surrogate cost sequence $F_{0:T-1}$. Since the surrogate cost functions are changing over time, the bound on $R^L(T)$ will depend on how much the system $\{g_t, f_t, \pi_t\}_{t\in\mathcal{T}}$ changes over the whole horizon $\mathcal{T}$. We defer the details to Section 5.3.

5.2 Contractive Perturbation and Stability

In this section, we introduce two key properties needed for our sub-linear regret guarantees in adaptive online policy selection. We define both with respect to trajectories generated by "slowly" time-varying parameters, which are easier to analyze than arbitrary parameter sequences.

Definition 5.2.1. *We denote the set of policy parameter sequences with ε-constrained step size by*

$$S_\varepsilon(t_1 : t_2) \coloneqq \{\theta_{t_1:t_2} \in \Theta^{t_2-t_1+1} \mid \|\theta_{\tau+1} - \theta_\tau\| \le \varepsilon, \forall \tau \in [t_1 : t_2 - 1]\}.$$

The first property we require is an exponentially decaying, or "contractive," perturbation property of the closed-loop dynamics of the system with the policy class. We now formalize this property.

Definition 5.2.2 (ε-Time-varying Contractive Perturbation)**.** *The ε-time-varying contractive perturbation property holds for $R_C > 0, C > 0$, $\rho \in (0, 1)$, and $\varepsilon \ge 0$ if, for any $\theta_{\tau:t-1} \in S_\varepsilon(\tau : t-1)$,*

$$\left\|g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g_{t|\tau}(x'_\tau, \theta_{\tau:t-1})\right\| \le C\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\|$$

holds for arbitrary $x_\tau, x'_\tau \in B_n(0, R_C)$ and time steps $\tau \le t$.

Intuitively, ε-time-varying contractive perturbation requires two trajectories starting from different states (in a bounded ball) to converge towards each other if they adopt the same slowly time-varying policy parameter sequence. We call the special case of $\varepsilon = 0$ *time-invariant contractive perturbation*, meaning the policy parameter is fixed. Although it may be difficult to verify the time-varying property directly since it allows the policy parameters to change, we show in Lemma 5.2.1 that time-invariant contractive perturbation implies that the time-varying version also holds for some small $\varepsilon > 0$.

The time-invariant contractive perturbation property is closely related to discrete-time incremental stability (e.g., Bayer, Bürger, and Allgöwer, 2013) and contraction theory (e.g., Tsukamoto, Chung, and Slotine, 2021), which have been studied in control theory. While some specific policies including DAC and MPC satisfy ε-time-varying contractive perturbation globally in linear systems, in other cases it is hard to verify. Our property is local and thus is easier to establish for broader applications in nonlinear systems (e.g., Example 5.1.2).

Besides contractive perturbation, another important property we need is the stability of the policy class, which requires $\pi_{0:T-1}$ can stabilize the system starting from the zero state as long as the policy parameter varies slowly. This property is stated formally below:

Definition 5.2.3 (ε-Time-varying Stability)**.** *The ε-time-varying stability property holds for $R_S > 0$ and $\varepsilon \geq 0$ if, for any $\theta_{\tau:t-1} \in S_\varepsilon(\tau : t-1)$, $\left\|g_{t|\tau}(0, \theta_{\tau:t-1})\right\| \leq R_S$ holds for any time steps $t \geq \tau$.*

Intuitively, ε-time-varying stability guarantees that the policy class $\pi_{0:T-1}$ can achieve stability if the policy parameters $\theta_{0:T-1}$ vary slowly.[2] Similarly to contractive perturbation, one only needs to verify time-invariant stability (i.e., $\varepsilon = 0$ and the policy parameter is fixed) to claim time-varying stability holds for some strictly positive ε (see Lemma 5.2.1). The reason we still use the time-varying contractive perturbation and stability in our assumptions is that they hold for $\varepsilon = +\infty$ in some cases, including DAC and MPC with confidence coefficients. Applying Lemma 5.2.1 for those systems will lead to a small, overly pessimistic ε.

Key Assumptions

We make two assumptions about the online policy selection problem to achieve regret guarantees.

Assumption 5.2.1. *The dynamics $g_{0:T-1}$, policies $\pi_{0:T-1}$, and costs $h_{0:T-1}$ are differentiable at every time step and satisfy that, for any convex compact sets*

[2] This property is standard in online control and is satisfied by DAC Agarwal et al., 2019; Hazan, Kakade, and Singh, 2020; Chen and Hazan, 2021; Simchowitz, Singh, and Hazan, 2020; Gradu, Hazan, and Minasyan, 2023 as well as Examples 5.1.1 & 5.1.2.

$\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{U} \subseteq \mathcal{R}^m$, one can find Lipschitzness/smoothness constants (can depend on $\mathcal{X}$ and $\mathcal{U}$) such that:

1. The dynamics $g_t(x, u)$ is $(L_{g,x}, L_{g,u})$-Lipschitz and $(\ell_{g,x}, \ell_{g,u})$-smooth in (x, u) on $\mathcal{X} \times \mathcal{U}$.
2. The policy function $\pi_t(x, \theta)$ is $(L_{\pi,x}, L_{\pi,\theta})$-Lipschitz and $(\ell_{\pi,x}, \ell_{\pi,\theta})$-smooth in (x, θ) on $\mathcal{X} \times \Theta$.
3. The stage cost function $h_t(x, u)$ is (L_h, L_h)-Lipschitz and $(\ell_{h,x}, \ell_{h,u})$-smooth in (x, u) on $\mathcal{X} \times \mathcal{U}$.

Assumption 5.2.1 is general because we only require the Lipschitzness/smoothness of g_t and h_t to hold for bounded states/actions within $\mathcal{X}$ and $\mathcal{U}$, where the coefficients may depend on $\mathcal{X}$ and $\mathcal{U}$. Similar assumptions are common in the literature of online control/optimization (Lin, Hu, Shi, et al., 2021; Shi et al., 2020; Li, Yang, Qu, Lin, et al., 2023).

Our second assumption is on the contractive perturbation and the stability of the closed-loop dynamics induced by a slowly time-varying policy parameter sequence.

Assumption 5.2.2. *Let $\mathcal{G}$ denote the set of all possible dynamics/policy sequences $\{g_t, \pi_t\}_{t \in \mathcal{T}}$ the environment/policy class may provide. For a fixed $\varepsilon \in \mathbb{R}_{\geq 0}$, the ε-time-varying contractive perturbation (Definition 5.2.2) holds with (R_C, C, ρ) for any sequence in $\mathcal{G}$. The ε-time-varying stability (Definition 5.2.3) holds with $R_S < R_C$ for any sequence in $\mathcal{G}$. We assume that the initial state satisfies $\|x_0\| < (R_C - R_S)/C$. Further, we assume that if $\{g, \pi\}$ is the dynamics/policy at an intermediate time step of a sequence in $\mathcal{G}$, then the time-invariant sequence $\{g, \pi\}_{\times T}$ is also in $\mathcal{G}$.*[3]

Note that Assumption 5.2.2 is on the joint properties of both the dynamical system and the policy class when composed together in a closed loop. The motivation is to generalize two key properties of linear systems under typical reasonable controllers: 1) the effect of past decisions on the current state decays exponentially fast, and 2) if the system is initialized near the origin, it remains near the origin. We generalize these properties via ε-time-varying contractive perturbation (Definition 5.2.2) and ε-time-varying stability (Definition 5.2.3), respectively. Although Assumption 5.2.2

[3] For $\{g, \pi\}_{\times T}$ to be in $\mathcal{G}$, it must satisfy other assumptions about contractive perturbation and stability that we impose on $\mathcal{G}$ but does not need to occur in real problem instances. We only use this assumption in the proof of Theorem 5.3.4, and it can be made without the loss of generality for time-invariant dynamics and policy classes.

may seem complicated to understand, it is less restrictive than the assumptions in the most closely related work (e.g., Agarwal et al., 2019; Hazan, Kakade, and Singh, 2020; Gradu, Hazan, and Minasyan, 2023) that focus on linear dynamics.

Compared to other settings where contractive perturbation holds globally (Agarwal et al., 2019; Simchowitz, Singh, and Hazan, 2020; Zhang, Li, and Li, 2021), Assumption 5.2.2 only requires it to hold locally in a bounded ball $B(0, R_C)$, which becomes important in nonlinear settings. This brings a new challenge because we need to guarantee that the starting state stays within $B(0, R_C)$ whenever we apply this property in the proof. Therefore, in Assumption 5.2.2, we assume $R_C > R_S + C\|x_0\|$. Similarly, to leverage the Lipschitzness/smoothness property, we require $\mathcal{X} \supseteq B(0, R_x)$ where $R_x \geq C(R_S + C\|x_0\|) + R_S$ and $\mathcal{U} = \{\pi(x, \theta) \mid x \in \mathcal{X}, \theta \in \Theta, \pi \in \mathcal{G}\}$. Since the coefficients in Assumption 5.2.1 depend on $\mathcal{X}$ and $\mathcal{U}$, we will set $\mathcal{X} = B(0, R_x)$ and $R_x = C(R_S + C\|x_0\|) + R_S$ by default when presenting these constants. The goal is to ensure that the controller never leaves the region where contractive perturbation applies, which is critical for our analysis and again generalizes properties found in the literature (e.g., Examples 5.1.1 and 5.1.2).

For some systems, verifying Assumption 5.2.2 is straightforward (e.g., Example 5.1.1). In other cases, we can rely on the following lemma, which can convert a time-invariant version of the property to general time-varying one. We defer its proof to Section 5.A.

Lemma 5.2.1. *Suppose Assumption 5.2.2 holds for $\varepsilon = 0$ and (R_C, C, ρ, R_S), which satisfies $R_C > (C + 1)R_S$. Suppose Assumption 5.2.1 also holds and let $\mathcal{X} := B(0, R_x)$, where $R_x = (C + 1)^2 R_S$. Then, Assumption 5.2.2 also holds for $\hat{\varepsilon} > 0$, $(\hat{R}_C, \hat{C}, \hat{\rho}, \hat{R}_S)$, and x_0 that satisfies $(\hat{R}_C - \hat{R}_S)/C$. Here, $\hat{R}_S, \hat{R}_C, \hat{\rho}$ are arbitrary constants that satisfies $R_S < \hat{R}_S < \hat{R}_C < R_C/(C + 1)$ and $\rho < \hat{\rho} < 1$. The positive constants $\hat{\varepsilon}$ and $\hat{C}$ are given detailed expressions in Section 5.A.*

Remark 5.2.2. *Lemma 5.2.1 can also be useful when applied to some parameterized controllers for time-invariant nonlinear systems. For example, the well-known "computed torque control" feedback linearization controllers for robotic manipulators (see, e.g., Slotine, Li, et al., 1991) renders the closed-loop dynamics exponentially stable about an equilibrium, and the feedback gains can be parameterized. Thus, it satisfies Assumption 5.2.2 in a neighborhood about the equilibrium, via Lemma 5.2.1. Even with time-invariant dynamics, the time-varying costs (such as tracking a trajectory determined online) provide a setting where selecting the policy parameters online can be beneficial.*

5.3 Memoryless Gradient-based Adaptive Policy Selection

Memoryless Gradient-Based Adaptive Policy Selection (M-GAPS) is inspired by the classic online gradient descent (OGD) algorithm (Hazan, 2016; Bansal and Gupta, 2019), with a novel approach for approximating the gradient of the surrogate stage cost F_t. It is an improved version of Gradient-Based Adaptive Policy Selection (GAPS) in (Lin, Preiss, Anand, et al., 2023) with a better computational complexity. In the context of online optimization, OGD works as follows. At each time t, the current stage cost describes how good the learner's current decision θ_t is. The learner updates its decision by taking a gradient step with respect to this cost. Mapping this intuition to online policy selection, the *ideal* OGD update rule would be the following.

Definition 5.3.1 (Ideal OGD Update)**.** *At time step t, update $\theta_{t+1} = \prod_\Theta(\theta_t - \eta\nabla F_t(\theta_t))$.*

This is because the surrogate cost F_t (Definition 5.1.2) characterizes how good θ_t is for time t if we had applied θ_t from the start, i.e., without the impact of other historical policy parameters $\theta_{0:t-1}$. However, since the complexity of computing ∇F_t exactly grows proportionally to t, the ideal OGD becomes intractable when the horizon T is large.

As outlined in Algorithm 6, M-GAPS uses G_t to approximate $\nabla F_t(\theta_t)$ efficiently. To see this, we compare the decompositions, with key differences highlighted in colored text:

$$\nabla F_t(\theta_t) = \sum_{b=0}^{t} \frac{\partial h_{t|0}}{\partial \theta_{t-b}}\bigg|_{x_0,(\theta_t)_{\times(t+1)}} \quad \text{and} \quad G_t = \sum_{b=0}^{t} \frac{\partial h_{t|0}}{\partial \theta_{t-b}}\bigg|_{x_0,\theta_{0:t}}. \tag{5.4}$$

GAPS efficiently approximates $\nabla F_t(\theta_t)$. by *replacing the ideal sequence $(\theta_t)_{\times(t+1)}$ by the actual sequence $\theta_{0:t}$*. This enables computing gradients along the actual trajectory experienced by the online policy without re-simulating the trajectory under θ_t. ε-time-varying contractive perturbation is the key to bound the bias of G_t: Intuitively, although θ_τ becomes more different with θ_t as τ decreases, its impact on $f_{t|0}$ decays more quickly (exponentially). We provide a rigorous bound of the bias in Theorem 5.3.1 and a proof outline in Section 5.B.

Although the expression of G_t in (5.4) decomposes it as the sum of $t+1$ partial derivatives, we can compute G_t efficiently by maintaining an auxiliary variable defined as the partial derivative of the current state with respect to all past policy

Algorithm 6: Memoryless Gradient-based Adaptive Policy Selection (M-GAPS, for `ALG`)

```
Parameters: Learning rate $\eta$, initial parameter $\theta_0$.
Initialize: Policy parameter $\theta_0$; Internal state $y_0 = \mathbf{0}$.
for $t = 0, 1, \dots, T-1$ do
    Take inputs $x_t, g_t, \pi_t, h_t$, and $f_t$.
    Update $y_{t+1} \leftarrow \frac{\partial g_{t+1|t}}{\partial x_t}\Big|_{x_t,\theta_t} \cdot y_t + \frac{\partial g_{t+1|t}}{\partial \theta_t}\Big|_{x_t,\theta_t}$.    /* Update partial
     derivatives accumulator.  */
    Let $G_t \leftarrow \frac{\partial h_{t|t}}{\partial x_t}\Big|_{x_t,\theta_t} \cdot y_t + \frac{\partial h_{t|t}}{\partial \theta_t}\Big|_{x_t,\theta_t}$.
    Update and output $\theta_{t+1} \leftarrow \prod_\Theta (\theta_t - \eta G_t)$.   /* $\Pi_\Theta$ is the Euclidean
     projection to $\Theta$.  */
end
```

parameters, i.e., $y_t \coloneqq \sum_{b=0}^{t} \frac{\partial g_{t|0}}{\partial \theta_{t-b}}\Big|_{x_0,\theta_{0:t}}$. Since we can update y_t with the chain rule, we provide the time- and space-efficient implementation of M-GAPS in Algorithm 6.

Compared to many previous online control algorithms that take a reduction approach based on OCO with Memory, our algorithm can be much more computationally efficient (see Section 5.B for an empirical comparison). Specifically, these works (Agarwal et al., 2019; Hazan, Kakade, and Singh, 2020; Chen and Hazan, 2021) take a different *finite-memory reduction* approach toward reducing the online control problem to OCO with Memory (Anava, Hazan, and Mannor, 2015) by completely removing the dependence on policy parameters before time step $t - B$ for a fixed memory length B. In the finite-memory reduction, one must "imaginarily" reset the state at time $t - B$ to be $\mathbf{0}$ and then use the B-step truncated multi-step cost function $h_{t|t-B}(\mathbf{0}, \theta_{t-B:t})$ in the OGD with Memory algorithm (Agarwal et al., 2019). When applied to our setting, this is equivalent to replacing G_t in (5.4) by

$$G_t' = \sum_{b=0}^{B-1} \frac{\partial h_{t|t-B}}{\partial \theta_{t-b}} |_{0,(\theta_t)_{\times(B+1)}} .$$

However, the estimator G_t' has limitations compared with G_t in M-GAPS. First, computing G_t' requires oracle access to the partial derivatives of the dynamics and cost functions for arbitrary state and actions. Second, even if those are available, G_t' is less computationally efficient than G_t in GAPS, especially when the policy is expensive to execute. Taking MPC (Example 5.1.1) as an example, computing G_t' at every time step requires solving B MPC optimization problems when re-simulating the system, where $B = \Omega(\log T)$. In contrast, computing G_t in GAPS only requires solving one MPC optimization problem and $O(1)$ matrix multiplications to update the partial derivatives.

Bounds on Truncation Error

We now present the first part of our main result, which states that the actual stage cost $h_t(x_t, u_t)$ incurred by GAPS is close to the ideal surrogate cost $F_t(\theta_t)$, and the approximated gradient G_t is close to the ideal gradient $\nabla F_t(\theta_t)$. In other words, GAPS mimics the ideal OGD update (Definition 5.3.1).

Theorem 5.3.1. *Suppose Assumptions 5.2.1 and 5.2.2 hold. Let $\{(x_t, u_t, \theta_t)\}_{t\in\mathcal{T}}$ denote the trajectory of M-GAPS with learning rate $\eta \leq \Omega\left((1-\rho)\varepsilon\right)$. Then, we have*

$$|h_t(x_t, u_t) - F_t(\theta_t)| = O\left((1-\rho)^{-3}\eta\right) \text{ and } \|G_t - \nabla F_t(\theta_t)\| = O\left((1-\rho)^{-5}\eta\right),$$

where $\Omega(\cdot)$ and $O(\cdot)$ hide the dependence on the Lipschitz/smoothness constants defined in Assumption 5.2.1 and the constant C in contractive perturbation.

We defer the proof of Theorem 5.3.1 to Section 5.B. Note that this result does not require any convexity assumptions on the surrogate cost F_t.

Regret Bounds for M-GAPS: Convex Surrogate Cost

The second part of our main result studies the case when the surrogate cost F_t is a convex function. This assumption is explicitly required or satisfied by the policy classes and dynamical systems in many prior works on online control and online policy selection (Agarwal et al., 2019; Hazan, Kakade, and Singh, 2020; Zhang, Li, and Li, 2021; Chen and Hazan, 2021).

The error bounds in Theorem 5.3.1 can reduce the problem of GAPS' regret bound in control to the problem of OGD's regret bound in online optimization, where the following result is well known: When the surrogate cost functions F_t are convex, the ideal OGD update (Definition 5.3.1) achieves the regret bound $\sum_{t=0}^{T-1} F_t(\theta_t) - \min_{\theta\in\Theta} \sum_{t=0}^{T-1} F_t(\theta) = O(\sqrt{T})$, when the step size η is of the order $1/\sqrt{T}$ (Hazan, 2016). By taking the biases on the stage costs and the gradients into consideration, we derive the adaptive regret bound in Theorem 5.3.2. Besides the adaptive regret, one can use a similar reduction approach to "transfer" other regret guarantees for OGD in online optimization to GAPS in control. We include the derivation of a dynamic regret bound as an example in Section 5.B.

Theorem 5.3.2. *Under the same assumptions as Theorem 5.3.1, if we additionally assume F_t is convex for every time t and $\mathsf{diam}(\Theta)$ is bounded by a constant D, then GAPS achieves adaptive regret*

$$R^A(T) = O\left(\eta^{-1} + (1-\rho)^{-5}\eta T + (1-\rho)^{-10}\eta^3 T\right),$$

where $O(\cdot)$ hides the same constants as in Theorem 5.3.1 and D.

We discuss how to choose the learning rate and the regret it achieves in the following corollary.

Corollary 5.3.3. *Under the same assumptions as Theorem 5.3.2, suppose the horizon length $T \gg \frac{1}{1-\rho}$. If we set $\eta = (1-\rho)^{\frac{5}{2}} T^{-\frac{1}{2}}$, then GAPS achieves adaptive regret $R^A(T) = O\left((1-\rho)^{-\frac{5}{2}} T^{\frac{1}{2}}\right)$.*

We defer the proof of Theorem 5.3.2 to Section 5.B. Compared to the (static) policy regret bounds of Agarwal et al. (2019) and Hazan, Kakade, and Singh (2020), our bound is tighter by a factor of $\log T$. The key observation is that the impact of a past policy parameter θ_{t-b} on the current stage cost c_t decays exponentially with respect to b (see Section 5.B for details). In comparison, the reduction-based approach first approximates c_t with $\hat{c}_t$ that depends on $\theta_{t-B+1:t}$, and then applies general OCO with memory results on $\hat{c}_t$ (Agarwal et al., 2019; Hazan, Kakade, and Singh, 2020). General OCO with memory cannot distinguish the different magnitudes of the contributions that $\theta_{t-B+1:t}$ make to $\hat{c}_t$, which leads to the regret gap of $B = O(\log T)$.

In the more restrictive setting of linear time-invariant dynamics with the DAC policy class, the results of a concurrent work (Kumar, Dean, and Kleinberg, 2023) can also be used to close the $\log T$ gap on static regret of online policy selection. In comparison, Theorem 5.3.2 considers more general time-varying dynamics and adopts the stronger metric of adaptive regret. As a practical matter, the follow-the-regularized-leader type of algorithm used by Kumar, Dean, and Kleinberg (2023) is often (much) less computationally efficient than a gradient-based algorithm like GAPS. Nevertheless, Kumar, Dean, and Kleinberg (2023) made distinct contributions by allowing the state space to be a general Banach space and providing a lower bound for OCO with unbounded memory.

Regret Bounds for M-GAPS: Nonconvex Surrogate Cost

The third part of our main result studies the case when the surrogate cost F_t is nonconvex. Before presenting the result, we formally define the variation intensity that measures how much the system changes over the whole horizon.

Definition 5.3.2 (Variation Intensity)**.** *Let $\{g_t, \pi_t, h_t\}_{t\in\mathcal{T}}$ be a sequence of dynamics/policy/cost functions that the environment provides. The variation intensity V of this sequence is defined as*

$$\sum_{t=1}^{T-1}\left(\sup_{x\in\mathcal{X},u\in\mathcal{U}}\|g_t(x,u)-g_{t-1}(x,u)\|+\sup_{x\in\mathcal{X},\theta\in\Theta}\|\pi_t(x,\theta)-\pi_{t-1}(x,\theta)\|\right.$$
$$\left.+\sup_{x\in\mathcal{X},u\in\mathcal{U}}|h_t(x,u)-h_{t-1}(x,u)|\right).$$

Variation intensity is used as a measure of hardness for changing environments in the literature of online optimization that often appear in regret upper bounds (see Mokhtari et al., 2016 for an overview). Definition 5.3.2 generalizes one of the standard definitions to online policy selection. Using this definition, we present our main result for GAPS applied to nonconvex surrogate costs using the metric of local regret (5.3).

Theorem 5.3.4. *Under the same assumptions as Theorem 5.3.1, if we additionally assume that $\Theta = \mathbb{R}^d$ for some integer d, then M-GAPS satisfies local regret*

$$R^L(T) = O\left(\frac{1+V}{(1-\rho)^3\eta}+\frac{\eta T}{(1-\rho)^6}+\frac{\eta^3 T}{(1-\rho)^{13}}\right),$$

where $O(\cdot)$ hides the same constants as in Theorem 5.3.1.

We defer the detailed expressions and the proof of Theorem 5.3.4 to Theorem 5.B.13 in Section 5.B. Note that the local regret will be sublinear in T if the variation intensity $V = o(T)$. To derive the local regret guarantee in Theorem 5.3.4, we address additional challenges compared to the convex case. First, we derive a local regret guarantee for OGD in online nonconvex optimization. We cannot directly apply results from the literature because they do not use ordinary OGD, and it is difficult to apply algorithms like Follow-the-Perturbed-Leader (e.g., Suggala and Netrapalli, 2020) to online policy selection due to constraints on information and step size. Then, to transfer the regret bound from online optimization to online policy selection, we show how to convert the measure of variation defined on $F_{0:T-1}$ to our variation intensity V defined on $\{g_t, \pi_t, h_t\}_{t\in\mathcal{T}}$.

A limitation of Theorem 5.3.4 is that we need to assume Θ is a whole Euclidean space so that GAPS will not converge to a point at the boundary of Θ that is not a stationary point.

5.4 Meta-Framework for Unknown Dynamics

We consider online policy optimization in a discrete-time dynamical system that varies over time with dynamics $x_{t+1} = g_t\big(x_t, u_t, f_t(x_t, a_t^*)\big) + w_t$, where $x_t \in \mathbb{R}^n$ denotes the system state, $u_t \in \mathbb{R}^m$ denotes the control input, and g_t is the dynamical function. Here, $f_t(x_t, a_t^*) \in \mathbb{R}^k$ is a nonlinear residual term of which the online agent can make (noisy) observations. It has a known function form f_t and an unknown parameter $a_t^* \in \mathcal{A} \subseteq \mathbb{R}^p$. The disturbance term $w_t \in \mathcal{W} \subseteq \mathbb{R}^n$ does not depend on the states or the control inputs.

To control this system, the online agent adopts a time-varying control policy π_t that is parameterized by a policy parameter θ_t and depends on its current estimation of the nonlinear residual. Specifically, the online agent picks the control input from the policy class $u_t = \pi_t\big(x_t, \theta_t, f_t(x_t, \hat{a}_t)\big)$. Here, function $f_t(\cdot, \hat{a}_t)$ reflects the online agent's current estimation of the ground true nonlinear residual function $f_t(\cdot, a_t^*)$ at time step t. Intuitively, we assume the policy class π_t cares about predicting the true nonlinear residual $f_t(x_t, a_t^*)$ rather than the unknown model parameter a_t^*. The objective of the online agent is to minimize the total cost $\sum_{t=0}^{T-1} c_t$ incurred over a finite horizon, where the stage cost at time step t is given by $c_t = h_t(x_t, u_t, \theta_t)$.

We provide a simple nonlinear control example that can be captured by our online policy optimization framework to help the readers understand the concepts we discussed.

Example 5.4.1. *Consider the problem of controlling a scalar discrete-time nonlinear dynamical system:*

$$x_{t+1} = x_t + \Delta\left(u_t + f_t(x_t, a_t^*) + w_t\right), \text{ where } f_t(x_t, a_t^*) = \phi(x_t) \cdot a_t^*. \tag{5.5}$$

In this equation, Δ is the discretization step size. The nonlinear residual takes the form $\phi(x_t) \cdot a_t^*$, where $\phi : \mathbb{R} \to \mathbb{R}^k$ is a (nonlinear) feature map and a_t^* is the unknown model parameter. To control this system, the online agent with an estimated model parameter $\hat{a}_t$ can adopt the policy class:

$$u_t = -k_t x_t - f_t(x_t, \hat{a}_t), \text{ where } f_t(x_t, \hat{a}_t) = \phi(x_t) \cdot \hat{a}_t, \text{ and } k_t = \theta_t. \tag{5.6}$$

Here, the goal of the second term $-f_t(x_t, \hat{a}_t)$ is to cancel out the true nonlinear residual $f_t(x_t, a_t^*)$. In an ideal case where the online agent has access to the true

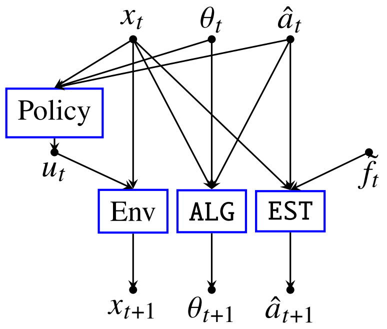

Figure 5.2: The meta-framework.

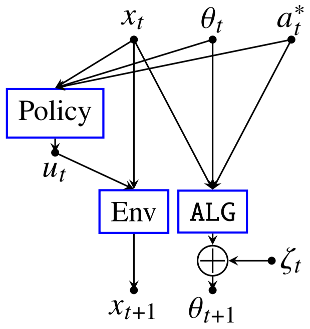

Figure 5.3: Theoretical comparand: ALG* with perturbations.

model parameter a_t^*, policy (5.6) *achieves the effect of removing the nonlinear residual and directly doing feedback control, resulting in the closed-loop dynamics* $x_{t+1} = x_t + \Delta\left(-k_t x_t + w_t\right)$. *In this case, the problem reduces to finding the optimal policy parameters (gains)* $\{\theta_t\}$ *in a known time-varying dynamical system.*

Performance Metrics

Although local regret is useful for measuring the performance of an online policy optimization algorithm under nonconvex surrogate costs, a limitation of applying it alone to our setting with unknown dynamical models is that the surrogate cost F_t is defined in terms of ALG's behavior with known true dynamics. To address this limitation, in addition to bounding the local regret of the policy parameters $\theta_{0:T-1}$, we also bound the distance between the actual trajectory of the online agent and the trajectory it would achieve with the same policy parameters $\theta_{0:T-1}$ and exact knowledge of true model parameters $a_{0:T-1}^*$.

Main Results

Our approach is outlined in Algorithm 7, where two modules ALG and EST work together to update the policy and estimated model parameter at each time step (see Figure 5.2 for an illustration). ALG and EST are responsible for optimizing the policy parameters $\theta_{0:T-1}$ and learning the unknown model parameters $a_{0:T-1}^*$ of the nonlinear residual terms, respectively:

- **ALG:** At time step t, ALG receives the current state x_t, policy parameter θ_t, and the known part of the time-varying system π_t, g_t, h_t, f_t. It also receives the current estimation $\hat{a}_t$ of the unknown model parameter a_t^*. Then, ALG outputs the new policy parameter θ_{t+1}. Note that we allow ALG to leverage/memorize history by maintaining an internal state y_t.

- **EST:** At time step t, EST receives the current state x_t and a (noisy) observation $\tilde{f}_t$ of the unknown component $f_t(x_t, a_t^*)$. Then, EST outputs the new estimation $\hat{a}_{t+1}$. Like ALG, we allow EST to keep internal state/memory (e.g., to memorize historical input data). We require EST to minimize the *trajectory-dependent model mismatches*:

$$\text{Zeroth-order model mismatch: } \varepsilon_t(x_t, \hat{a}_t, a_t^*) \coloneqq \left\| f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*) \right\|, \quad (5.7a)$$

$$\text{First-order model mismatch: } \varepsilon_t'(x_t, \hat{a}_t, a_t^*) \coloneqq \left\| \nabla_x f_t(x_t, \hat{a}_t) - \nabla_x f_t(x_t, a_t^*) \right\|_F. \quad (5.7b)$$

We adopt the shorthand $\varepsilon_t = \varepsilon_t(x_t, \hat{a}_t, a_t^*)$ and $\varepsilon_t' = \varepsilon_t'(x_t, \hat{a}_t, a_t^*)$ when the context is clear.

Algorithm 7: Meta-Framework

```
Require: ALG and EST
Require: Knowing functions $\{\pi_t, g_t, h_t, f_t\}$ at each time step $t$
Initialize: State $x_0$; Policy parameter $\theta_0$; Model parameter estimation $\hat{a}_0$.
for $t = 0, 1, \ldots, T-1$ do
    Decide control input $u_t = \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t))$.
    Incur stage cost $h_t(x_t, u_t, \theta_t)$.
    $\theta_{t+1} \leftarrow$ ALG.update$(x_t, \theta_t, \pi_t, g_t, h_t, f_t, \hat{a}_t)$. /* ALG can have internal memory. */
    System evolves to $x_{t+1} = g_t(x_t, u_t, f_t(x_t, a_t^*)) + w_t$.
    Receive a (noisy) observation $\tilde{f}_t$ of $f_t(x_t, a_t^*)$.
    $\hat{a}_{t+1} \leftarrow$ EST.update$(x_t, \tilde{f}_t, \hat{a}_t)$. /* EST can have internal memory. */
end
```

The key idea in analyzing our meta-framework (Algorithm 7) is to characterize how the inexact model estimations generated by EST affect the behavior ALG. We start by considering the "ideal" dynamics of applying ALG with exact model parameters $a_{0:T-1}^*$, which we denote as ALG*, and compare them with the actual dynamics of ALG that performs the update with estimated model parameters $\hat{a}_{0:T-1}$. We state the key insight of our analysis in the informal lemma below, which connects the performance of the meta-framework with ALG* and the model mismatches.

Lemma 5.4.2 (Informal)**.** *Suppose ALG* satisfies the desired properties in the next subsection. Then, the meta-framework (Algorithm 7) generates the same policy parameters as ALG* with perturbation ζ_t on the update of θ_{t+1} (see Figure 5.3). Further, $\sum_{t=0}^{T-1} \|\zeta_t\| = O\left(\sum_{t=0}^{T-1} \varepsilon_t + \sum_{t=0}^{T-1} \varepsilon_t'\right)$.*

The formal statement of Lemma 5.4.2 can be found in Theorem 5.4.3.

The rest of this section is organized as following: First, we specify the properties of ALG* that enables the meta-framework to be robust against inexact model parameters in the policy parameter update. Then, we formulate EST's task of learning $f_t(x_t, a_t^*)$ as an online optimization problem, where we view the state x_t as picked by an adaptive adversary. We also discuss how this problem reduces to existing results on online optimization.

Online Policy Optimization

In this section, we take a perspective that views the updates performed by ALG as part of a joint dynamics formed together with the original dynamical system. Compared to the common approach of analyzing ALG separately from the dynamical system to which it applies, our dynamical view enables us to compare the differences of applying ALG under different external inputs (i.e., different $\hat{a}_t$ estimates) more efficiently.

We consider the class of online policy optimization algorithms whose joint dynamics with the original system can be written in the following form: When the model parameter a_t is given as the input to ALG at time step t, the joint dynamics can be written as

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(x_t, y_t, \theta_t, a_t) = \begin{pmatrix} q_t^x(x_t, y_t, \theta_t, a_t) \\ q_t^y(x_t, y_t, \theta_t, a_t) \\ q_t^\theta(x_t, y_t, \theta_t, a_t) \end{pmatrix}, \text{ for } x_t \in \mathbb{R}^n, y_t \in \mathbb{R}^p, \theta_t \in \Theta \subset \mathbb{R}^d. \tag{5.8}$$

Here, $y_t \in \mathbb{R}^p$ is an auxiliary state that ALG can use to store something besides the system state x_t and the policy parameter θ_t to help it perform the update. For example, y_t can be a finite memory buffer that stores information from the past. It can also be the integral of past states in an integral controller. Thus, we introduce y_t to allow broader classes of online policy optimization algorithms, and a concrete example of y_t is the auxiliary state in M-GAPS.

The goal of formulating joint dynamics (5.8) is to compare the behaviors of the meta-framework and ALG* with perturbations on policy parameter updates. Specifically, recall that $\hat{a}_{0:T-1}$ denote the estimated model parameters of EST. The actual trajectory of the meta-framework is

$$\text{Meta-framework: } (x_{t+1}, y_{t+1}, \theta_{t+1})^\top = q_t(x_t, y_t, \theta_t, \hat{a}_t). \tag{5.9}$$

We compare it with the joint dynamics of ALG* (see Figure 5.3). Recall that ALG* denotes the scenario when ALG has access to exact model parameters $a^*_{0:T-1}$:

$$\text{ALG}^* \text{ with perturbations: } (x_{t+1}, y_{t+1}, \theta_{t+1})^\top = q_t(x_t, y_t, \theta_t, a_t^*) + (0, 0, \zeta_t)^\top. \tag{5.10}$$

Here, ζ_t is an additive perturbation on the update equation of policy parameter θ_{t+1}. To understand (5.10) intuitively, it is helpful to draw connections with the process of using a gradient-based optimizer to update the parameter θ_t in ML, where $\zeta_t \equiv 0$ corresponds to the case when exact gradients are available. In contrast, nonzero perturbations correspond to the more practical case when the optimizer can only use biased estimations of the gradient, which still performs well in general.

Note that the estimated model parameters $\hat{a}_{0:T-1}$ generated by EST may also depend on the state x_t and other parts of the dynamical system. Thus, a natural question is whether we should also incorporate the update rule of EST into the joint dynamical system in (5.9), where we include $\hat{a}_t$ as another element of the joint state. However, we still choose to model $\hat{a}_t$ as an external input in (5.9) and handle the update of $\hat{a}_t$ separately. This is because our approach requires comparing the actual joint dynamics with (5.10). Since a_t^* is an external input decided by the environment in (5.10), keeping the joint state space identical in (5.9) makes the comparison easier. Further, a strength of our proof framework based on the joint dynamics is that we can show the actual trajectory (5.9) will stay close to (5.10). However, we know that the estimated model parameter sequence $\{\hat{a}_t\}$ will not converge to the true sequence $\{a_t^*\}$ in general.

We state three important properties of the joint dynamics induced by ALG. The first property is about the Lipschitzness with respect to the model mismatches ε_t and ε_t'.

Property 5.4.1. *[Lipschitzness] For any $x_t, y_t, \theta_t, \hat{a}_t$ that satisfies $\|x_t\| \leq R_x$, $\|y_t\| \leq R_y$, $\theta_t \in \Theta$, $\hat{a}_t \in \mathcal{A}$, the following Lipschitzness conditions hold:*

$$\begin{aligned}
\left\|q_t^x(x_t, y_t, \theta_t, a_t^*) - q_t^x(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\leq \alpha_x \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_x \varepsilon_t'(x_t, \hat{a}_t, a_t^*),\\
\left\|q_t^y(x_t, y_t, \theta_t, a_t^*) - q_t^y(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\leq \alpha_y \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_y \varepsilon_t'(x_t, \hat{a}_t, a_t^*),\\
\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\leq \alpha_\theta \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_\theta \varepsilon_t'(x_t, \hat{a}_t, a_t^*).
\end{aligned}$$

Further, $q_t^\theta(x, y, \theta, a_t^*)$ is $(L_{\theta,x}, L_{\theta,y})$-Lipschitz in (x, y).

Intuitively, Property 5.4.1 says that the error brought by the inexact model parameters only “distort” the ideal joint dynamics (5.10) in the form of zeroth-order and first-order prediction errors. Therefore, to bound the error injected into the joint dynamics at every step, EST only needs to minimize ε_t and ε_t' on the actual state trajectory $x_{0:T-1}$ that the online agent visits. Note that this property can be viewed as a standard assumption about Lipschitzness if ALG is a gradient-based algorithm. This is because all terms that involve the unknown model parameter will take the form $f_t(x_t, \hat{a}_t)$ and $\nabla_x f_t(x_t, \hat{a}_t)$ in the joint dynamics.

The second property is about contraction stability of x_t and y_t under exact model parameters $a^*_{0:T-1}$. As we show in Theorem 5.4.3, this property guarantees that the dynamical updates of states x_t and y_t in the joint dynamics are robust to the model mismatches $\{\varepsilon_t, \varepsilon_t'\}_{0:T-1}$.

Property 5.4.2. *[Contraction Stability] For any sequence $\theta_{0:T-1}$ that satisfies the slowly time-varying constraint that $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ for all time step t, the partial dynamical system*

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, a_t^*), \qquad y_{t+1} = q_t^y(x_t, y_t, \theta_t, a_t^*) \tag{5.11}$$

satisfies that $\|x_t\| \le R_x^* < R_x$ and $\|y_t\| \le R_y^* < R_y$ always hold if the system starts from $(x_\tau, y_\tau) = (0, 0)$. Further, for some function $\gamma : \mathbb{Z}_{\ge 0} \to \mathbb{R}_{\ge 0}$ that satisfies $\sum_{t=0}^{\infty} \gamma(t) \le C$, from any initial states $(x_\tau, y_\tau), (x_\tau', y_\tau')$ that satisfy $\|x_\tau\|, \left\|x_\tau'\right\| \le R_x$ and $\|y_\tau\|, \left\|y_\tau'\right\| \le R_y$, the trajectory satisfies $\left\|(x_{\tau+t}, y_{\tau+t}) - (x_{\tau+t}', y_{\tau+t}')\right\| \le \gamma(t) \cdot \left\|(x_\tau, y_\tau) - (x_\tau', y_\tau')\right\|$.

Note that Property 5.4.2 is different with the contraction assumption of Lin, Preiss, Anand, et al. (2023) because it also considers the internal state y_t of ALG besides the system state x_t. The requirement that $\sum_{t=0}^{\infty} \gamma(t) \le C$ is also weaker than the exponential decay rate in Lin, Preiss, Anand, et al. (2023).

Intuitively, Property 5.4.2 guarantees that when the exact model parameters $\{a_t^*\}$ are replaced by inexact $\{\hat{a}_t\}$, the resulting trajectory $\{(x_t, y_t)\}$ still stays close to the trajectory that $\theta_{0:T-1}$ would achieve under exact predictions once the mismatch errors $\varepsilon_t, \varepsilon_t'$ are small or bounded. Property 5.4.2 can be viewed as an extension of the time-varying stability and contractive perturbation property in Lin, Preiss, Anand, et al., 2023 to include state y_t maintained by ALG. This is required in our framework because y_t can be affected by the prediction errors and it is involved in the dynamics of updating θ_t.

The third property we need is the robustness of the update rule of the policy parameter θ_t. Specifically, it requires the regret guarantee achieved by ALG to be robust against a certain level of adversarial disturbances $\{\zeta_t\}$ on the update dynamics of θ_t.

Property 5.4.3. *[Robustness] Consider the joint dynamics in* (5.10)*. When* $\|\zeta_t\| \le \bar{\zeta}$ *holds for all* t*, the resulting* $\{\theta_t\}$ *satisfies the slowly-time-varying constraint* $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ *for all time* t*. Further,* ALG* *with perturbations* (5.10) *can achieve a regret guarantee* $R(T, \sum_{t=0}^{T-1} \|\zeta_t\|)$ *that depends on the total magnitude of the perturbation sequence* $\zeta_{0:T-1}$*.*

To understand Property 5.4.3, we can think about online gradient descent (OGD) in online optimization problems without state or dynamics. It is known that this approach is robust to (biased) disturbances on the gradient estimation, and the total amount of added disturbances will affect the final regret bound (see, for example, Theorem 5.B.10).

Now, we present our main results about the stability of applying ALG with inexact model parameters and the regret bound in Theorem 5.4.3. Besides, Theorem 5.4.3 also bounds the distances between the actual trajectory and the trajectory achieved by applying the same policy parameter sequence with the exact model parameter sequence.

Theorem 5.4.3. *Suppose Properties 5.4.1, 5.4.2, and 5.4.3 hold. Let* $\xi = \{x_t, y_t, \theta_t\}$ *be the trajectory of the meta-framework (Algorithm 7). If the prediction errors* $\{\varepsilon_t, \varepsilon_t'\}_{0:T-1}$ *are uniformly bounded such that the following inequalities hold for all time step* t*:* $\alpha_\theta \varepsilon_t + \beta_\theta \varepsilon_t' \le \bar{\zeta}/2$*, and*

$$(\alpha_x + \alpha_y)\varepsilon_t + (\beta_x + \beta_y)\varepsilon_t' \le \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C}\right\},$$

then the trajectory ξ *satisfies* $\|x_t\| \le R_x$*,* $\|y_t\| \le R_y$*, and* $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ *for all time steps* t*. Further, define* $\tilde{\xi} \coloneqq \{\tilde{x}_t, \tilde{y}_t, \theta_t\}_{0:T-1}$*, where* $\{\tilde{x}_t, \tilde{y}_t\}_{0:T-1}$ *are obtained by implementing the policy parameters* $\theta_{0:T-1}$ *with exact model parameters* $a^*_{0:T-1}$*, i.e., the trajectory of partial joint dynamics* (5.11)*. The trajectory* $\tilde{\xi}$ *achieves the regret* $R(T, \sum_{t=0}^{T-1} \|\zeta_t\|)$ *with* $\sum_{t=0}^{T-1} \|\zeta_t\|$ *upper bounded by*

$$\left(\alpha_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\alpha_x + \alpha_y)\right)\sum_{t=0}^{T-1} \varepsilon_t$$
$$+ \left(\beta_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\beta_x + \beta_y)\right)\sum_{t=0}^{T-1} \varepsilon_t'.$$

The total distances between the states on the trajectories ξ *and* $\tilde{\xi}$ *satisfies that*

$$\sum_{t=1}^{T} \|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \le C\left((\alpha_x + \alpha_y)\sum_{t=0}^{T-1} \varepsilon_t + (\beta_x + \beta_y)\sum_{t=0}^{T-1} \varepsilon_t'\right).$$

We defer the proof of Theorem 5.4.3 to Section 5.E. Intuitively, Theorem 5.4.3 states that when the prediction error terms $\{\varepsilon_t, \varepsilon_t'\}_{0:T-1}$ are uniformly bounded, the actual trajectory ξ of applying `ALG` with inexact model parameters $\hat{a}_{0:T-1}$ will be uniformly bounded. Further, if the actual parameter sequence of $\theta_{0:T-1}$ is applied with exact model parameters $a^*_{0:T-1}$, the resulting trajectory $\tilde{\xi}$ achieves a regret guarantee that depends on the magnitudes of the prediction errors. It is worth noticing that the regret in Theorem 5.4.3 can be any regret that depends on the trajectory $\tilde{\xi}$. And as we discussed before, we evaluate the regret on trajectory $\tilde{\xi}$ rather than ξ because the metrics like the local regret are designed for evaluating the actual policy parameters $\theta_{0:T-1}$ rather than the whole trajectory ξ. The distances between ξ and $\tilde{\xi}$ are bounded in the last inequality in Theorem 5.4.3.

To show Theorem 5.4.3, the key idea is to fit the trajectory $\tilde{\xi}$ into the dynamical equation (5.10), where we design ζ_t to compensate the difference between the update rules $q_t(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*)$ and $q_t(x_t, y_t, \theta_t, \hat{a}_t)$. To leverage Property 5.4.3, we show the perturbations $\zeta_{0:T-1}$ we constructed are uniformly bounded by $\bar{\zeta}$. We bound ζ_t and the distances between ξ and $\tilde{\xi}$ by induction. The induction is important because the magnitude of ζ_t depends on the distance between $\{x_t, y_t\}_{0:T-1}$ and $\{\tilde{x}_t, \tilde{y}_t\}_{0:T-1}$ in the past time steps. On the other hand, to bound the distance between $\{x_t, y_t\}$ and $\{\tilde{x}_t, \tilde{y}_t\}$, we need to leverage the contraction property in Property 5.2.2, which relies on $\|\zeta_t\| \leq \bar{\zeta}$ so that $\theta_{0:T-1}$ is slowly time-varying. Lastly, we conclude the proof with the bounds on the distance between ξ and $\tilde{\xi}$ as well as the norm of ζ_t that depend on the model mismatches $\{\varepsilon_t, \varepsilon_t'\}_{0:T-1}$.

Online Parameter Estimation

The second part of our meta framework focuses on predicting the unknown model parameter based on possibly noisy observations of the true nonlinear residual $f_t(x_t, a_t^*)$. A critical difference with prior works on system identification or model-based learning (e.g., Dean et al., 2020) is that we only seek to optimize the zeroth-order and first-order model mismatches $\{\varepsilon_t, \varepsilon_t'\}$ (defined in (5.7)) on the actual trajectory that the online agent experiences. It is worth noticing that, although learning the ground-truth model parameter a_t^* is impossible for a general nonlinear residual, minimizing the sum of zeroth-order model mismatches incurred on the actual trajectory can be formulated as a classic online regression problem, which we discuss below:

Online regression problem: At the beginning, the environment commits a sequence of error functions $e_t : \mathbb{R}^n \times \mathcal{A} \to \mathbb{R}, t = 0, \ldots, T-1$, which are defined as

$e_t(x, a) \coloneqq f_t(x, a_t^*) - f_t(x, a)$ for $t = 0, \ldots, T - 1$. [4] The relationship between the error function e_t and the model mismatches $\{\varepsilon_t, \varepsilon_t'\}$ is $\varepsilon_t = \|e_t(x_t, \hat{a}_t)\|$, and $\varepsilon_t' = \|\nabla_x e_t(x_t, \hat{a}_t)\|$. At each time step t, the online parameter estimator $\mathtt{EST}$ predicts $\hat{a}_t = \mathtt{EST}(x_{0:t-1}, \hat{a}_{0:t-1}) \in \mathcal{A}$, which means the estimation $\hat{a}_t$ can be a general function of the historical states and estimations. Then, the environment reveals $x_t \in B_n(0, R_x)$ that can depend on the history $x_{0:t-1}$ and $\hat{a}_{0:t-1}$. We define the stage loss of $\mathtt{EST}$ as $\ell_t = \|e_t(x_t, \hat{a}_t)\|^2$, which is equal to the squared ℓ_2-norm of the model mismatch $e_t(x_t, \hat{a}_t)$.

Under different sets of assumptions on the error functions and the sequence of true model parameters $\{a_t^*\}$, existing online algorithms can achieve regret guarantees. We consider a general form of expected regret bound: $\mathbb{E}\left[\sum_{t=1}^T \ell_t\right] \leq R_0^\ell(T)$, where the expectation is taken over the randomness of implementing $\mathtt{EST}$ and generating x_t. While different assumptions and designs of $\mathtt{EST}$ can achieve different bounds on $R_0^\ell(T)$, we will provide an example later where a simple gradient estimator can achieve sublinear $R_0^\ell(T)$ under certain assumptions about the nonlinear residual and the path length $\sum_{t=1}^{T-1} \left\|a_{t+1}^* - a_t^*\right\|$.

While most prior works focus on minimizing the magnitude of the zeroth-order model mismatch $e_t(x_t, \hat{a}_t)$, we also need to bound the first-order model mismatch $\nabla_x e_t(x_t, \hat{a}_t)$ because it contributes to the regret bound in Theorem 5.4.3 (recall that $\|\nabla_x e_t(x_t, \hat{a}_t)\|_F = \varepsilon_t'$). Our main result in this section is about an automatic reduction from the regret bound $R_0^\ell(T)$ to a bound on the expected sum of the squared gradients $\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(x_t, \hat{a}_t)\|_F^2\right]$.

Remark 5.4.4. *Besides the online policy optimization problem for control, the regret bound that concerns $\|\nabla_x e_t(x_t, \hat{a}_t)\|$ can be of independent interest for the problem of online regression, because it characterizes how sensitive the regression loss is to any perturbations on the input sequence $x_{0:T-1}$ under the same estimations $\hat{a}_{0:T-1}$. Intuitively, if gradients of the error functions are small, the estimations $\hat{a}_{0:T-1}$ will be robust to small perturbations on the input sequence.*

To enable a reduction from the regret bound $R_0^\ell(T)$ to the gradient error bound, we employ Property 5.4.4 about the dynamical system that generates the state x_t. Specifically, we require there to be at least a small level of randomness when choosing x_t. Recall that $\hat{a}_{t+1}$ is decided based on the history $x_{0:t}$ and $\hat{a}_{0:t}$. We

[4] Thus, the error functions $e_{0:T-1}$ will not adapt to the inputs and online decisions.

define the filtrations $\mathcal{F}_t \coloneqq \sigma(x_{1:t}, \hat{a}_{1:t})$ and $\mathcal{F}_t' \coloneqq \sigma(x_{1:t}, \hat{a}_{1:t+1})$, which satisfy $\mathcal{F}_t \subseteq \mathcal{F}_t' \subseteq \mathcal{F}_{t+1}$.

Property 5.4.4. *There is a certain level of random disturbances when generating each state x_t, i.e., for some $\bar{\epsilon} > 0$ and $\underline{\sigma} > 0$, one can find a σ-algebra $\mathcal{G}_t$ such that $\mathcal{F}_t' \subseteq \mathcal{G}_t \subseteq \mathcal{F}_{t+1}$ and $\|x_{t+1} - \mathbb{E}[x_{t+1} \mid \mathcal{G}_t]\| \le \bar{\epsilon}$, $Cov(x_{t+1} \mid \mathcal{G}_t) \succeq \underline{\sigma} I$.*

Intuitively, the randomness enforced by Property 5.4.4 will "force" EST to also minimize the gradient of the error functions. To see this, suppose an input state x_t is given by $\bar{x}_t + v_t$, where $\bar{x}_t$ is the mean and v_t is a random disturbance. When the disturbance v_t is sufficiently small, we know that $e_t(x_t, \hat{a}_t) \approx e_t(\bar{x}_t, \hat{a}_t) + \nabla_x e_t(\bar{x}_t, \hat{a}_t) \cdot v_t$ by Taylor's expansion. Since we can pick v_t randomly in different directions, we know the zeroth-order loss $\mathbb{E}[e_t(x_t, \hat{a}_t)^2]$ cannot converge to zero unless the magnitude of the gradient $\nabla_x e_t(\bar{x}_t, \hat{a}_t)$ converges to zero. We follow this intuition to show the reduction from the regret bound $R_0^\ell(T)$ to the total gradient error in Theorem 5.4.5.

Theorem 5.4.5. *Suppose that for all time t, each dimension $i \in [k]$ of the error function satisfies*

$$\|\nabla_x e_t(x, a)_i\| \le \beta_e, \text{ and } \left\|\nabla_x^2 e_t(x, a)_i\right\| \le \gamma_e, \text{ for any } x \in B(0, R_x) \text{ and } a \in \mathcal{A}.$$

Suppose Property 5.4.4 holds with $\bar{\epsilon} \le \min\{\frac{1}{4}, \frac{1}{2\gamma_e}, \frac{1}{4\beta_e\gamma_e}\}$ and $\underline{\sigma} > 0$. If EST achieves the zeroth-order regret $\mathbb{E}\left[\sum_{t=1}^T \ell_t\right] \le R_0^\ell(T) \le \bar{\epsilon}^3 T$, the expected total squared gradient loss satisfies that

$$\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(x_t, \hat{a}_t)\|_F^2\right] \le \frac{2k}{\underline{\sigma}}(1 + \gamma_e + \beta_e\gamma_e)\bar{\epsilon}^3 T + 2k\gamma_e^2\bar{\epsilon}^2 T.$$

Recall that k is the dimension of the unknown component $f_t(x_t, a_t^*)$. We defer the proof of Theorem 5.4.5 to Section 5.E. We provide the following corollary to help the readers understand this result in a special case when $R_0^\ell(T)$ is $O(\sqrt{T})$. For example, a gradient estimator can achieve this regret bound if ℓ_t is convex in a.

Corollary 5.4.6. *Under the same assumptions as Theorem 5.4.5, if EST achieves $R_0^\ell(T) = O(\sqrt{T})$ and Property 5.4.4 holds with $\bar{\epsilon} = \theta(T^{1/6})$ and $\underline{\sigma} = \Omega(\bar{\epsilon}^2)$, then the expected total squared gradient loss is bounded by $\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(x_t, \hat{a}_t)\|_F^2\right] = O(kT^{5/6})$.*

In summary, with the help of Theorem 5.4.5, we reduce the problem of bounding the total squared first-order prediction errors $\sum_{t=0}^{T-1}(\varepsilon_t')^2$ to the standard online optimization problem. By substituting the bounds on $\varepsilon_{0:T-1}$ and $\varepsilon_{0:T-1}'$ into Theorem 5.4.3, one can derive the local regret bound for the actual joint dynamics and bound the distance between trajectories ξ and $\tilde{\xi}$.

Application: Matched Disturbance

In this section, we consider an instantiation of our setting to demonstrate the effectiveness of our meta-framework. Specifically, we study the matched-disturbance dynamics (Ferguson et al., 2020; Sinha et al., 2022; Garofalo, Ott, and Albu-Schäffer, 2012), where the controller can choose a control input to "cancel out" the nonlinear residual term $f_t(x_t, a_t^*)$ when the exact model parameter a_t^* is available. The dynamics have the form

$$x_{t+1} = g_t(x_t, u_t, f_t(x_t, a_t^*)) + w_t = \phi_t(x_t, u_t + f_t(x_t, a_t^*)) + w_t. \tag{5.12}$$

A ubiquitous application of the matched disturbance dynamics is the general joint-space dynamics of robotic manipulators (Siciliano et al., 2008) when the system has actuators for every joint. The matched-disturbance structure also appears in tilted-rotor rotorcraft (Rajappa et al., 2015; Zheng et al., 2020), which can move in six degrees of freedom. In both cases, due to the second-order structure of the rigid-body dynamics, all external disturbances are equivalent to additional joint torque (resp. rotor tilt/thrust) inputs. Our Example 5.4.1 also fits into the framework of (5.12). To control a matched-disturbance system, a natural policy class is to first cancel out the nonlinear residual with $-f_t(x_t, \hat{a}_t)$ and then apply an actuation term $\psi_t(x_t, \theta_t)$ to achieve the optimal costs. This policy class can be expressed as

$$u_t = \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t)) = -f_t(x_t, \hat{a}_t) + \psi_t(x_t, \theta_t). \tag{5.13}$$

To derive local regret for the meta-framework, we need assumptions (Assumptions 5.2.1-5.D.4) on the system that includes the dynamics, policy classes, and costs, which we discuss in detail in Section 5.E. Note that the matched-disturbance dynamics/policy class we consider can recover the setting (Lin, Preiss, Anand, et al., 2023) as a special case when f_t and w_t are always zero (so there is no need to estimate a_t^*). We recover the same regret bound as Lin, Preiss, Anand, et al., 2023 in that special case (see Lemma 5.4.7).

Online Policy Optimization. M-GAPS can serves as `ALG` in our meta-framework. M-GAPS use $\hat{a}_t$ to estimate how the current state x_t and policy parameter θ_t would affect the next state x_{t+1} and the current cost. The estimations are characterized by

$$\hat{g}_{t+1|t}(x_t, \theta_t) \coloneqq g_t(x_t, \pi_t(x_t, \theta_t, f_t(x_t, \hat{a}_t)), f_t(x_t, \hat{a}_t)), \text{ and} \tag{5.14a}$$

$$\hat{h}_{t|t}(x_t, \theta_t) \coloneqq h_t\left(x_t, \pi_t\left(x_t, \theta_t, f_t(x_t, \hat{a}_t)\right), \theta_t\right). \tag{5.14b}$$

Although M-GAPS can be applied to any online policy optimization problems that fit into the setting of Section 5.4 in general, we focus on its application to disturbance-matched dynamics and policy class for theoretical analysis. We verify that the joint dynamics of M-GAPS satisfy the required properties of our meta-framework to derive a concrete regret bound in Section 5.E.

Algorithm 8: Memoryless Gradient-based Adaptive Policy Selection (M-GAPS, for ALG)

```
Parameters: Learning rate $\eta$, initial parameter $\theta_0$.
Initialize: Policy parameter $\theta_0$; Internal state $y_0 = \mathbf{0}$.
for $t = 0, 1, \ldots, T-1$ do
    Take inputs $x_t, g_t, \pi_t, h_t, f_t$ and $\hat{a}_t$.    /* Inputs given when meta-framework calls ALG.update. */
    Use $\hat{a}_t$ to obtain $\hat{g}_{t+1|t}$ and $\hat{h}_{t|t}$.
    Update $y_{t+1} \leftarrow \frac{\partial \hat{g}_{t+1|t}}{\partial x_t}\Big|_{x_t,\theta_t} \cdot y_t + \frac{\partial \hat{g}_{t+1|t}}{\partial \theta_t}\Big|_{x_t,\theta_t}$.    /* Update partial derivatives accumulator. */
    Let $G_t \leftarrow \frac{\partial \hat{h}_{t|t}}{\partial x_t}\Big|_{x_t,\theta_t} \cdot y_t + \frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\Big|_{x_t,\theta_t}$.
    Update and output $\theta_{t+1} \leftarrow \prod_\Theta (\theta_t - \eta G_t)$.    /* $\Pi_\Theta$ is the Euclidean projection to $\Theta$. */
end
```

A key step of our proof shows that, when exact model parameters $a^*_{0:T-1}$ are available, M-GAPS is robust against perturbations on policy parameter updates as required by Property 5.4.3.

Lemma 5.4.7. *Under Assumptions 5.2.1 and 5.2.2, Property 5.4.3 holds when $\eta \leq \bar{\eta}$ for some positive constant $\bar{\eta}$ and*

$$R^L_\eta(T, \sum_{t=0}^{T-1} \|\zeta_t\|) = O\left(\frac{1}{\eta}(1 + V_{sys} + V_w) + \eta T + \eta^3 T + \frac{1}{\eta}\sum_{t=1}^{T-1}\|\zeta_t\|\right),$$

where the variation intensities are defined as $V_w = \sum_{t=1}^{T-1} \|w_t - w_{t-1}\|$ and

$$V_{sys} = \sum_{t=1}^{T-1} \Big(\sup_{x\in\mathcal{X}, u\in\mathcal{U}} \|\phi_t(x,u) - \phi_{t-1}(x,u)\| + \sup_{x\in\mathcal{X}, \theta\in\Theta} \|\psi_t(x,\theta) - \psi_{t-1}(x,\theta)\|$$
$$+ \sup_{x\in\mathcal{X}, u\in\mathcal{U}, \theta\in\Theta} |h_t(x,u,\theta) - h_{t-1}(x,u,\theta)|\Big).$$

The formal statement and proof of Lemma 5.4.7 can be found in Section 5.E. Note that in the special case of Lin, Preiss, Anand, et al., 2023, we have $\Theta = \mathbb{R}^d$, $V_w = 0$, and $\sum_{t=1}^{T-1} \|\zeta_t\| = 0$. The local regret bound $R_\eta^L(T, 0)$ of M-GAPS given by Lemma 5.4.7 matches the local regret bound of GAPS in Lin, Preiss, Anand, et al., 2023, because the projected gradients are identical with the gradients when $\Theta = \mathbb{R}^d$.

Online Parameter Estimation. In the application of matched-disturbance dynamics, we assume the online parameter estimator EST can make a noisy observation $\tilde{f}_t$ of the true nonlinear residual $f_t(x_t, a_t^*)$ after it decides $\hat{a}_t$ at each time step t. Recall that the prediction error function is defined as $e_t(x, a) \coloneqq f_t(x, a) - f_t(x, a_t^*)$ and the true prediction loss at time step t as $\ell_t \coloneqq \|e_t(x_t, \hat{a}_t)\|^2$. We instantiate EST with the gradient estimator (Algorithm 9), where $\tilde{f}_t$ is a (noisy) observation of $f_t(x_t, a_t^*)$ provided by the environment. It performs online gradient descent on an estimated prediction loss function constructed from $\tilde{f}_t$.

```
Algorithm 9: Gradient Estimator (for EST)
Parameters: Learning rate $\iota$; Initialize: Model parameter estimation $\hat{a}_0$.
for $t = 0, 1, \ldots, T - 1$ do
    Take inputs $x_t$ and $\tilde{f}_t$.        /* Inputs given when meta-framework
     calls EST.update.  */
    Incur loss $\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) \coloneqq \|f_t(x_t, \hat{a}_t) - \tilde{f}_t\|^2$.
    Update and output $\hat{a}_{t+1} \leftarrow \prod_{\mathcal{A}} \left(\hat{a}_t - \iota \cdot \partial\tilde{\ell}_t/\partial a_t|_{x_t, \hat{a}_t, \tilde{f}_t}\right)$.
end
```

Using Theorems 5.4.3 and 5.4.5, we show a local regret guarantee of our meta-framework in Theorem 5.4.8 and test it numerically in the setting of Example 5.4.1. Due to space limit, we defer the proof of Theorem 5.4.8 and the simulation results to Section 5.E.

Theorem 5.4.8. *Under Assumptions 5.2.1-5.D.4, if we use M-GAPS for* ALG *and Gradient Estimator for* EST*, the trajectory $\xi = \{x_t, y_t, \theta_t\}$ achieves an expected local regret* [5] *of*

$$O\left(\eta^{-1}(1 + V_{sys} + \bar{\epsilon} \cdot T) + \eta T + (\sqrt{m\bar{\epsilon}} + m\bar{\epsilon}) \cdot T\right),$$

where V_{sys} is the total variation of the system and $\bar{\epsilon}$ is the magnitude of the random disturbance w_t (see Section 5.E for detailed definitions). Under the same definition of $\tilde{\xi}$ as Theorem 5.4.3, the expected total distance between ξ and $\tilde{\xi}$ is bounded by

[5] We change the gradient $\nabla F_t(\theta_t)$ to the projected gradient $\nabla_{\eta,\Theta} F_t(\theta_t)$ (see Definition 5.C.2 in Section 5.E) in the local regret. This metric is introduced by Hazan, Singh, and Zhang, 2017 for online nonconvex optimization with constraints.

$$\mathbb{E}\left[\sum_{t=0}^{T-1}\left(\|x_t - \tilde{x}_t\| + \|y_t - \tilde{y}_t\|\right)\right] = O\left(T^{3/4} + \sqrt{m\bar{\epsilon}} \cdot T\right).$$

5.5 Application: Using Predictions Adaptively

Our example about Model Predictive Control (MPC) with Confidence Coefficients generalizes the λ-confident policy proposed by Li, Yang, Qu, Shi, et al. (2022). In this setting, some policy parameter $\theta^{(c)} \in \Theta$ can achieve near-optimal performance when the predictions of the future are accurate (consistency), and another policy parameter $\theta^{(r)} \in \Theta$ has a worst-case guarantee even when the predictions are unreliable (robustness). Minimizing regret in this setting implies that the online policy is both robust and consistent.

Recall that in this example, MPC selects the current control action by solving the optimization problem

$$\begin{aligned}
&\arg\min_{u_{t:t+k-1|t}} \sum_{\tau=t}^{t+k-1} f_\tau(x_{\tau|t}, u_{\tau|t}) + q(x_{t+k|t}, \tilde{Q}) \\
&\quad \text{s. t. } x_{t|t} = x_t, \\
&\quad\quad x_{\tau+1|t} = A_\tau x_{\tau|t} + B_\tau u_{\tau|t} + \lambda_t^{[\tau-t]} w_{\tau|t} : \ t \le \tau < t+k,
\end{aligned} \tag{5.15}$$

where $\theta_t = \left(\lambda_t^{[0]}, \lambda_t^{[1]}, \ldots, \lambda_t^{[k-1]}\right)$, $\Theta \subseteq [0,1]^k$. Thus, Θ is a convex compact set with diameter $\sqrt{k}$.

For this example to satisfy Assumptions 5.2.1 and 5.2.2, we make two standard assumptions that are also required by prior works on online control with MPC (Lin, Hu, Shi, et al., 2021; Lin, Hu, Qu, et al., 2022). The first assumption is about the uniform bounds on the dynamical matrices, cost function matrices, and disturbances.

Assumption 5.5.1. *For any time step $t \in \mathcal{T}$, we have $\|A_t\| \le a, \|B_t\| \le b, \|w_t\| \le \bar{w}$, and*

$$\mu I_n \preceq Q_t \preceq \ell I_n, \mu I_m \preceq R_t \preceq \ell I_m, \mu I_n \preceq \tilde{Q} \preceq \ell I_n.$$

We also assume that $\left\|w_{\tau|t}\right\| \le \bar{w}$ for predicted disturbances.

The second assumption is about the uniform controllability of the LTV system:

Assumption 5.5.2. *We define the transition matrix in this LTV system as*

$$\Phi(t_2, t_1) := \begin{cases} A_{t_2-1} \cdots A_{t_1}, & \textit{if } t_2 > t_1, \\ I, & \textit{otherwise.} \end{cases}$$

For any two time steps $t' > t$, we define the controllability matrix in this LTV system as following:

$$\Xi_{t,t'} = \left[\Phi(t', t+1)B_t, \Phi(t', t+2)B_{t+1}, \ldots, \Phi(t', t')B_{t'}\right].$$

We assume that there exists a positive integer d_0 such that the smallest singular value of the matrix $\Xi_{t,t'}$ is uniformly lower bounded by some positive constant σ when $t' \geq t + d_0$, i.e., $\sigma_{min}\left(\Xi_{t,t'}\right) \geq \sigma$ holds for any $t' \geq t + d_0$, where $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix.

Before we proceed to show that Assumptions 5.2.1 and 5.2.2 hold in this example, we first define an auxiliary parameterized optimal problem control problem that will be used in our analysis: For any two time steps $t < t'$, let $\psi_t^{t'}(y_t, v_{t:t'}; P_{t'})$ denote the optimal trajectory planned according to initial state y_t, disturbances $v_{t:t'}$, and terminal cost matrix $P_{t'}$, i.e.,

$$\begin{aligned}\psi_t^{t'}(y_t, v_{t:t'}; P_{t'}) = \operatorname*{arg\,min}_{x_{t:t'}, u_{t:(t'-1)}} & \sum_{\tau=t}^{t'-1} f_\tau\left(x_\tau, u_\tau\right) + \frac{1}{2} x_{t'}^\top P_{t'} x_{t'} \\ \text{s.t. } & x_{\tau+1} = A_\tau x_\tau + B_\tau u_\tau + v_\tau, \forall \tau \in [t : t'-1]; \\ & x_t = y_t.\end{aligned}$$

Using this notation, we can express MPC with confidence coefficients as

$$\pi_t(x_t, \theta_t) = \psi_t^{t+k}(x_t, \{\lambda_t^{[\tau-t]} w_{\tau|t}\}_{\tau\in[t:t+k-1]}; \tilde{Q})_{u_t},$$

where the index u_t denotes the corresponding entry in the predictive optimal solution. The perturbation bound in Lin, Hu, Shi, et al., 2021, Theorem 3.3 states that for any $\tau \in [t, t']$, we have

$$\left\|\psi_t^{t'}(y_t, v_{t:t'}; \tilde{Q})_{u_\tau} - \psi_t^{t'}(y_t', v_{t:t'}'; \tilde{Q})_{u_\tau}\right\| \leq C_0\left(\rho_0^{\tau-t}\left\|y_t - y_t'\right\| + \sum_{j=t}^{t'} \rho_0^{|\tau-j|}\left\|v_j - v_j'\right\|\right), \tag{5.16}$$

where $C_0 > 0, \rho_0 \in (0, 1)$ are constants that depends on the system parameters including a, b, μ, ℓ, σ, and d_0. And the inequality (5.16) still holds if we replace the index u_τ on the left hand side by x_τ. Therefore, we know that $\psi_t^{t+k}(y_t, v_{t:t+k}; \tilde{Q})_{u_t}$ is bounded Lipschitz in y_t and $v_{t:t+k}$. Note that $\pi_t(x_t, \theta_t)$ is an affine function in its inputs (x_t, θ_t), i.e., $\mathsf{ALG}_t(x_t, \theta_t)$ can be expressed equivalently as

$$\mathsf{ALG}_t(x_t, \theta_t) = -\bar{K}_t^{(k)} x_t - \sum_{\tau=t}^{t+k-1} \lambda_t^{[\tau-t]} \bar{K}_t^{(k,\tau)} w_{\tau|t}, \tag{5.17}$$

where the matrices $\bar{K}_t^{(k)}, \bar{K}_t^{(k,\tau)}$ only depends on $\{(A_t, B_t, Q_t, R_t)\}_{t\in\mathcal{T}}$ and $\tilde{Q}$ (Zhang, Li, and Li, 2021; Yu et al., 2020). The superscript k denotes the prediction horizon of the MPC we adopt, thus $k = T$ will give MPC future predictions all the way to the end of the online policy selection game. So the smoothness constants $\ell_{\pi,x} = \ell_{\pi,\theta} = 0$. Thus, we see that Assumption 5.2.1 holds. We also see that the surrogate cost function F_t is convex.

Next, we show a lemma about contractive perturbation and stability. We defer the proof of Lemma 5.5.1 to Section 5.F.

Lemma 5.5.1. *Suppose Assumptions 5.5.1 and 5.5.2 hold. Recall that C_0 and ρ_0 are given in* (5.16). *Then, for any $\rho \in (\rho_0, 1)$, if the prediction horizon satisfies*

$$k \geq \frac{1}{2} \log\left(C_0^3 ab\rho_0/(\rho - \rho_0)\right) / \log(1/\rho_0),$$

the MPC with confidence coefficients policy class satisfies ε-time-varying contractive perturbation with $\varepsilon = +\infty$, $R_C = +\infty$, $C = C_0$ and decay factor ρ. It also satisfies ε-time-varying stability with $\varepsilon = +\infty$ and $R_S = \frac{C_0(1-\rho_0+C_0)\bar{w}}{(1-\rho_0)(1-\rho)}$.

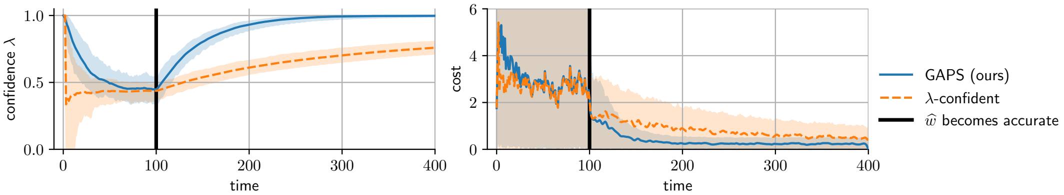

Figure 5.4: Comparing GAPS and baseline (Li, Yang, Qu, Shi, et al., 2022) for online adaptation of a confidence parameter for MPC with disturbance predictions. *Left:* Confidence parameter. *Right:* Per-step cost. Shaded bands show 10%-90% quantile range over randomized disturbance properties. See body for details.

MPC confidence parameter. We compare GAPS to the follow-the-leader-type method of Li, Yang, Qu, Shi, et al. (2022) for tuning a scalar confidence parameter in model-predictive control with noisy disturbance predictions. The setting is close to Example 5.1.1 but restricted to satisfy the conditions of the theoretical guarantees in Li, Yang, Qu, Shi, et al. (2022). We consider the scalar system $x_{t+1} = 2x_t + u_t + w_t$ under non-stochastic disturbances w_t with the cost $f_t(x_t, u_t) = x_t^2 + u_t^2$. For $t = 0$ to 100, the predictions of w_t are corrupted by a large amount of noise. After $t > 100$, the prediction noise is instantly reduced by a factor of 100. In this setup, an ideal algorithm should learn to decrease confidence level at first to account for the noise, but then increase to $\lambda \approx 1$ when the predictions become accurate.

Figure 5.4 shows the values of the confidence parameter λ and the per-timestep cost generated by each algorithm. Both methods are initialized to $\lambda = 1$. The method of Li, Yang, Qu, Shi, et al. (2022) rapidly adjusts to an appropriate confidence level at first, while GAPS adjusts more slowly but eventually reaches the same value. However, when the accuracy changes, GAPS adapts more quickly and obtains lower costs towards the end of the simulation. In other words, we see that GAPS behaves essentially like an instance of Ideal OGD with constant step size, which is consistent with our theoretical results (Theorem 5.3.1).

Recover the optimal policy under the stochastic prediction model. Consider a policy class that performs linear-feedback on the past predictions:

$$u_t = -Kx_t + \sum_{\tau=0}^{m} M_t^{(\tau)} \cdot v_{t-\tau}(\theta),$$

where the matrix $M_t^{(\tau)} \in \mathbb{R}^{2\times 2}$ is the policy parameter. To find the optimal policy parameter M^θ for a predictor $v_t(\theta)$, we can adopt M-GAPS to tune M_t. By the results of Lin, Preiss, Anand, et al., 2023, the averaged cost incurred by M-GAPS will converge towards the optimal average cost (up to an error of $O(\eta)$, where η is the learning rate). Thus, an alternative approach to compare two predictors is to run a policy optimizer directly and compare the average cost.

In Figures 5.5 and 5.6, we plot the effect of applying M-GAPS to Example 4.3.2 with $\rho = 0.5$ and the predictions $V_t(I)$. We see that M-GAPS can learn an approximation of the optimal predictive policy and the average cost converges to the optimal average cost over time (without the knowledge of the joint distribution). We repeat the same experiment with the predictions $V_t(\theta)$ and M-GAPS also finds a near-optimal predictive policy over time. As shown in Figure 5.9, the prediction power matches the long-term performance gain of online policy optimization with a stronger predictor.

5.6 Application: Quadcopter Control

Dynamics and Representation. We represent the quadcopter state at time step t as a tuple

$$(p_t, v_t, r_t, \omega_t), \text{ for } p_t \in \mathbb{R}^3, v_t \in \mathbb{R}^3, r_t \in \mathfrak{so}(3), \omega_t \in \mathfrak{so}(3),$$

where p_t is the position; v_t is the velocity; r_t is the logarithmic coordinate, so $\exp(r_t)$ rotates from the body frame to the inertial frame; and ω_t is the angular in the body frame. The control input is a tuple

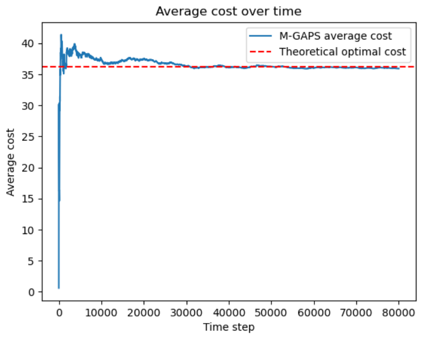

Figure 5.5: Example: Recover the optimal policy. Average cost over time. $\rho = 0.5$ and prediction $V_t(I)$. M-GAPS learning rate 1×10^{-4}.

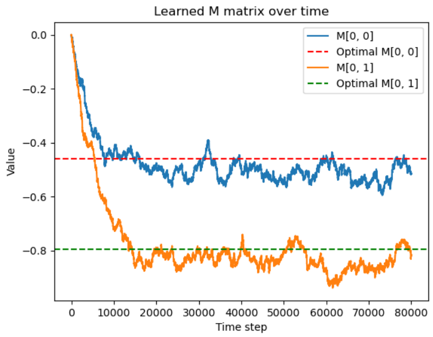

Figure 5.6: Example: Recover the optimal policy. Learned policy. $\rho = 0.5$ and prediction $V_t(I)$. M-GAPS learning rate 1×10^{-4}.

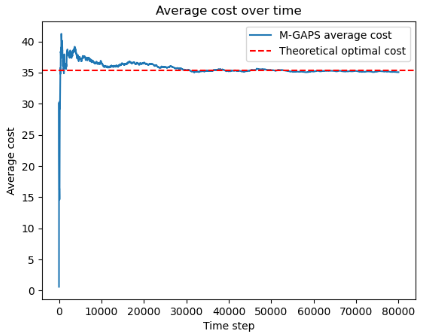

Figure 5.7: Example: Recover the optimal policy. Average cost over time. $\rho = 0.5$ and prediction $V_t(\theta)$. M-GAPS learning rate 1×10^{-4}.

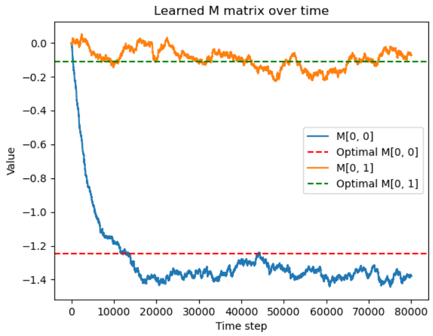

Figure 5.8: Example: Recover the optimal policy. Learned policy. $\rho = 0.5$ and prediction $V_t(\theta)$. M-GAPS learning rate 1×10^{-4}.

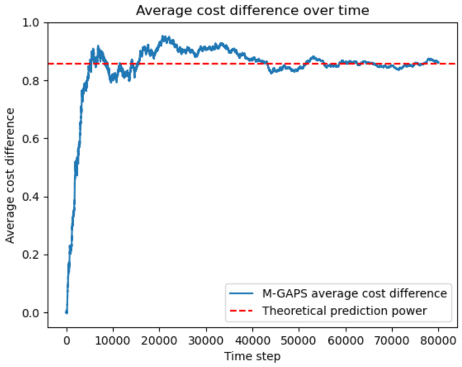

Figure 5.9: Prediction power and the average cost difference of running M-GAPS with predictions $V_t(I)$ or $V_t(\theta)$.

$$u_t = (\xi_t, \tau_t), \text{ for } \xi_t \in \mathbb{R}_{\geq 0} \text{ and } \tau_t = \left[\tau_t^{(r)}, \tau_t^{(p)}, \tau_t^{(y)}\right]^\top \in \mathbb{R}^3,$$

where ξ_t is a mass-normalized thrust and τ_t is the desired angular acceleration in the body frame. The discrete-time dynamics of the quadcopter are given by

$$p_{t+1} = p_t + \delta v_t, \tag{5.18a}$$

$$v_{t+1} = v_t + \delta(\xi_t \exp(r_t) e_z - g), \tag{5.18b}$$

$$r_{t+1} = \log\left(\exp(r_t)\exp(\delta\omega_t)\right), \tag{5.18c}$$

$$\omega_{t+1} = \omega_t + \delta\hat{\tau}_t, \tag{5.18d}$$

where $\delta > 0$ is the time interval for discretization and g is a 3D vector that denotes the gravitational constant (so the entries at dimensions x and y are zeros). Here, we use e_x, e_y, e_z to denote the standard basis vectors of $\mathbb{R}^3$. As a remark, elements of $\mathfrak{so}(3)$ are skew-symmetric matrices, i.e.,

$$\mathfrak{so}(3) = \{A \in \mathbb{R}^{3\times 3} \mid A = -A^\top\},$$

which we interpret as 3D angular velocities. In (5.18d), the map $v \longrightarrow \hat{v}$ is the canonical (natural) isomorphism from $\mathbb{R}^3$ to $\mathfrak{so}(3)$:

$$\hat{}: \mathbb{R}^3 \longrightarrow \mathfrak{so}(3), \; \hat{v} = \begin{pmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{pmatrix} \text{ for } v = (v_1, v_2, v_3)^\top.$$

Note that by expressing the dynamics as (5.18), we implicitly assume that the control input u_t can be realized.

We use p_t^d to denote the target position trajectory (and define v_t^d, a_t^d, and ω_t^d for the desired velocity, acceleration, and angular velocity similarly). To incorporate the integral of the tracking error into our policy class, we add a virtual state $\bar{i}_{t+1}$ with the dynamics

$$\bar{i}_{t+1} = \bar{i}_t + \delta(p_t - p_t^d) \text{ for } t \geq 0,$$

where the initial state $\bar{i}_0 = 0$. Therefore, the full state of the quadcopter is expressed as $x_t = (\bar{i}_t, p_t, v_t, r_t, \omega_t)$.

Policy Class. Our control policy is formed by an outer loop that calculates a desired mass-normalized thrust vector $z_t \in \mathbb{R}^3$ and an inner loop that orients the quadcopter's body towards z_t. The outer-loop law is given by

$$z_t = -K_t^{(i)}\bar{i}_t - K_t^{(p)}\left(p_t - p_t^d\right) - K_t^{(v)}(v_t - v_t^d) + a_t^d + g e_z, \tag{5.19}$$

where the gain matrices $K_t^{(i)}, K_t^{(p)}, K_t^{(v)}$ are diagonal and positive definite. We compute the thrust command $\xi_t \in \mathbb{R}_{\geq 0}$ by projecting z_t onto the body thrust axis:

$$\xi_t = z_t^\top \exp(r_t) e_z.$$

The inner loop first constructs a desired attitude $r_t^d \in \mathfrak{so}(3)$ as the shortest rotation that takes e_z to the direction of z, which is

$$r_t^d = \begin{cases} \cos^{-1}\left(e_z \cdot \frac{z_t}{\|z_t\|}\right) \widehat{\left(\frac{e_z \times z}{\|e_z \times z\|}\right)}, & \text{if } e_z \times z \neq \mathbf{0}, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Recall that $\hat{}$ maps $\mathbb{R}^3$ to $\mathfrak{so}(3)$. In our setting, $\|z\| > 0$ because the acceleration and error terms in (5.19) are sufficiently small. And we simplify the problem by not considering the desired heading (yaw). The controller decides the desired angular acceleration τ_t by

$$\tau_t' = -K_t^{(r)} \log\left(\exp(r_t)\exp\left(-r_t^d\right)\right) - K_t^{(\omega)}(\omega_t - \omega_t^d), \tag{5.20a}$$

$$\tau_t = \operatorname{softclamp}\left(\tau_t', [B^{(xy)}, B^{(xy)}, B^{(z)}]^\top\right), \tag{5.20b}$$

where the gain matrices $K_t^{(r)}$ and $K_t^{(\omega)}$ are diagonal and positive definite. The softclamp function

$$\operatorname{softclamp}(x, B) = B \tanh(x/B)$$

is critical for maintaining the stability, and it is applied elementwise in (5.20b).

In summary, the policy parameters that are updated in the online policy optimization include

$$\begin{aligned} &K_t^{(i)} = \operatorname{diag}(k_t^{(i,xy)}, k_t^{(i,xy)}, k_t^{(i,z)}),\ K_t^{(p)} = \operatorname{diag}(k_t^{(p,xy)}, k_t^{(p,xy)}, k_t^{(p,z)}), \\ &K_t^{(v)} = \operatorname{diag}(k_t^{(v,xy)}, k_t^{(v,xy)}, k_t^{(v,z)}),\ K_t^{(r)} = \operatorname{diag}(k_t^{(r,xy)}, k_t^{(r,xy)}, k_t^{(r,z)}), \\ &K_t^{(\omega)} = \operatorname{diag}(k_t^{(\omega,xy)}, k_t^{(\omega,xy)}, k_t^{(\omega,z)}), \end{aligned}$$

where we use equal gains for the two horizontal axes x and y. Thus, the "raw" policy parameter ϑ_t is

$$\vartheta_t = \left(k_t^{(i,xy)}, k_t^{(i,z)}, k_t^{(p,xy)}, k_t^{(p,z)}, k_t^{(v,xy)}, k_t^{(v,z)}, k_t^{(r,xy)}, k_t^{(r,z)}, k_t^{(\omega,xy)}, k_t^{(\omega,z)}\right) \in \mathbb{R}_{>0}^{10}.$$

In the experiment, we use the reparameterization $\theta_t = \log(\vartheta_t)$ and optimize the parameter θ_t because it improves the optimization landscape.

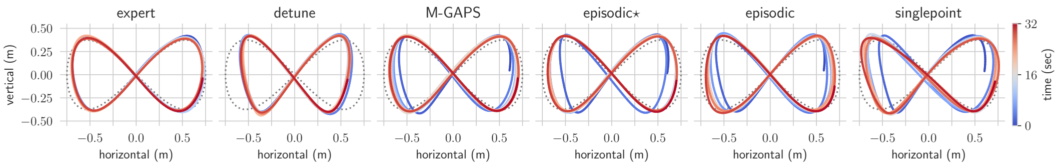

Figure 5.10: Trajectories of a quadrotor tracking an aggressive figure-8 trajectory under online policy optimization algorithms. Dotted line shows target. Color changes from blue (beginning) to red (end) over time. *Expert*: $\theta_t = \theta^m \ \forall t$. *Detune*: $\theta_t = \theta^m - \log 2 \ \forall t$. All algorithms initialized with *detune* parameter.

Experiments

In the experiment, we use a a Bitcraze Crazyflie 2.0 with the manufacturer's thrust upgrade bundle and an upgraded battery. We conduct three trajectory tracking experiments to compare M-GAPS with other policy optimization algorithms (DiffTune (Cheng et al., 2024) and ORPF (Zhang, Zhou, et al., 2024)) and benchmarks. In the first experiment, we start with a scenario where the disturbances are small. The expert-tuned policy parameter θ^m is near-optimal in this case, so we test if M-GAPS can improve the control performance from a bad initialization. Then, in the next two experiments, we want to demonstrate the necessity of online policy optimization by ruling out the possibility that θ^m is near-optimal in all scenarios. To achieve this goal, we introduce heavy payload or time-varying wind and test whether M-GAPS can outperform θ^m in such settings.

Suboptimal initialization. We simulate the process of controller tuning by initializing each online policy optimization algorithm with the "detuned" parameter $\theta_0 = \theta^m - \log 2$, which decrease the feedback gains by a half under the logarithmic policy parameterization. We plot the trajectories of the expert parameter θ^m, the detuned parameter θ_0, and each candidate online policy optimization algorithm in Figure 5.10. We note that M-GAPS starts near *Detune* but quickly gets closer to *Expert* after 2-3 laps.

In Figure 5.11, we plot the cost difference of each candidate/benchmark with the expert policy θ^m, which is near-optimal in this case. We see that the cumulative cost of M-GAPS converges towards a constant level, and it is lower than other episodic policy optimization algorithms.

Heavy payload. The original mass of the quadcopter is 39g. We attached an additional 23g steel weight near its center, which can be viewed as a large, near-constant disturbance in the dynamics. Note that θ^m is not tuned to handle this

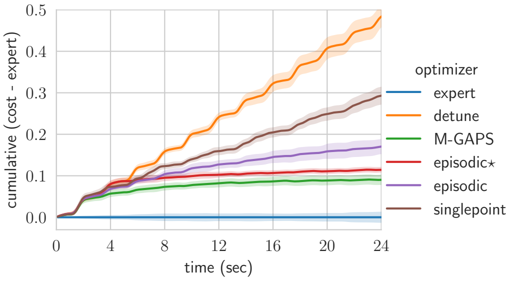

Figure 5.11: Cumulative cost difference versus expert-tuned parameters θ^m in figure-8 tracking experiment with detuned initialization. Error bars indicate ±1 standard deviation over 5 trials. See caption of Figure 5.10 for legend key.

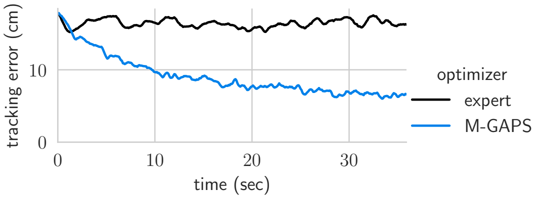

Figure 5.12: Position tracking error under M-GAPS for heavy payload disturbance.

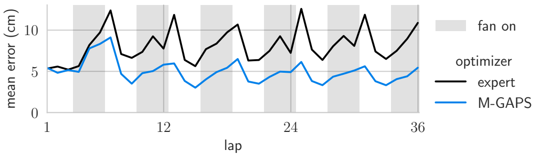

Figure 5.13: Position tracking error under M-GAPS for periodic fan disturbance. Error is averaged per "lap" due to substantial variance within each lap from fan airflow pattern.

scenario. We compare the position tracking error of M-GAPS and the export θ^m in Figure 5.12. The results verifies our conjecture that θ^m is suboptimal, and deploying an online policy optimization algorithm like M-GAPS can effectively reduce the tracking error compared with θ^m.

Time-varying wind. As an addition to the second experiment, we use periodic wind from three household box fans to test if M-GAPS can adapt quickly. We attach a cardboard panel to the quadcopter to magnify the effect of wind disturbances. The quadcopter flies back-and-forth pattern with each lap takes about 4 seconds. To create the periodic disturbance, we toggle the fan power every 12 seconds. We plot the tracking error of M-GAPS and compare it against the expert θ^m in Figure 5.13. The results show that M-GAPS outperforms the expert, and it confirms that M-GAPS can adapt quickly to environment changes on the scale of about ten seconds.

5.A Proof of Contractive Perturbation and Stability

Proof of Lemma 5.2.1

We first restate Lemma 5.2.1 with detailed coefficients in Lemma 5.A.1:

Lemma 5.A.1. *Suppose Assumption 5.2.2 holds for $\varepsilon = 0$ and (R_C, C, ρ, R_S), which satisfies $R_C > (C+1)R_S$. Suppose Assumption 5.2.1 also holds and let $\mathcal{X} \coloneqq B_n(0, R_x)$, where $R_x = (C+1)^2 R_S$. Then, Assumption 5.2.2 also holds for $\hat{\varepsilon} > 0$, $(\hat{R}_C, \hat{C}, \hat{\rho}, \hat{R}_S)$, and x_0 that satisfies the inequality $\|x_0\| \le (\hat{R}_C - \hat{R}_S)/C$. Here, $\hat{R}_S, \hat{R}_C, \hat{\rho}$ are arbitrary constants that satisfies $R_S < \hat{R}_S < \hat{R}_C < R_C/(C+1)$ and $\rho < \hat{\rho} < 1$. Other coefficients are given by*

$$\begin{aligned}\hat{C} = {} & \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u} L_{\pi,x}\right) + L_{g,u}\ell_{\pi,x}\right) \cdot \\ & \left(1 + L_{g,x} + L_{g,u}L_{\pi,x}\right)^{2h} \hat{\rho}^{-h}, \\ \hat{\varepsilon} = {} & \min\left\{\frac{(\hat{\rho}^h - C\rho^h)(1-\rho)^2}{C \cdot C'\rho\left(1 + L_{g,x} + L_{g,u}L_{\pi,x}\right)^{2h}}, \frac{(1-\rho)^2(\hat{R}_S - R_S)}{CL_{g,u}L_{\pi,\theta}}\right\}, \text{ where} \\ C' = {} & \Big((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u}\cdot(L_{\pi,x} + L_{\pi,\theta})\right) + L_{g,u}\cdot(\ell_{\pi,x} + \ell_{\pi,\theta})\Big)\cdot \\ & (L_{g,u}L_{\pi,\theta} + 1),\end{aligned}$$

where h is a constant integer that satisfies $C\rho^h < \min\{\hat{\rho}^h, 1 - \hat{R}_S/\hat{R}_C\}$.

Before showing Lemma 5.A.1, we first show that the composition of Lipschitz and smooth functions is still Lipschitz and smooth in the following technical lemma:

Lemma 5.A.2. *Suppose the sequence of functions $\iota_{1:t}$ satisfies that $\iota_i : D_i \to D_{i+1}$ is L-Lipschitz and ℓ-smooth for all $i \in \{1, 2, \cdots, t\}$. Then, their composition $(\iota_t \circ \iota_{t-1} \circ \cdots \circ \iota_1)$ is L^t-Lipschitz and $\ell(1+L)^{2t}$-smooth.*

Proof of Lemma 5.A.2. We show the conclusion by induction. For $t = 1$, ι_1 is L-Lipschitz and $\ell(1+L)$-smooth.

Suppose we have shown that $(\iota_t \circ \iota_{t-1} \circ \cdots \circ \iota_1)$ is L^t-Lipschitz and $\ell(1+L)^{2t}$-smooth for any t functions that satisfies the assumptions of Lemma 5.A.2. For $t+1$, we simplify the notation by defining $\hat{\iota} \coloneqq (\iota_{t+1} \circ \iota_t \circ \cdots \circ \iota_2)$. Our goal is to show $(\hat{\iota} \circ \iota_1)$ is L^{t+1}-Lipschitz and $\ell(1+L)^{2(t+1)}$-smooth. Note that

$$\|(\hat{\iota} \circ \iota_1)(x) - (\hat{\iota} \circ \iota_1)(x')\| \le L^t \|\iota_1(x) - \iota_1(x')\| \le L^{t+1}\|x - x'\|,$$

where we use the induction assumption in the first inequality and the assumption of Lemma 5.A.2 in the second inequality. We also see that

$$
\begin{aligned}
&\left\| \frac{\partial(\hat{\iota} \circ \iota_1)}{\partial x}\bigg|_{x} - \frac{\partial(\hat{\iota} \circ \iota_1)}{\partial x}\bigg|_{x'} \right\| \\
&= \left\| \frac{\partial \hat{\iota}}{\partial y}\bigg|_{\iota_1(x)} \cdot \frac{\partial \iota_1}{\partial x}\bigg|_{x} - \frac{\partial \hat{\iota}}{\partial y}\bigg|_{\iota_1(x')} \cdot \frac{\partial \iota_1}{\partial x}\bigg|_{x'} \right\| && (5.21a)\\
&\le \left\| \frac{\partial \hat{\iota}}{\partial y}\bigg|_{\iota_1(x)} - \frac{\partial \hat{\iota}}{\partial y}\bigg|_{\iota_1(x')} \right\| \cdot \left\| \frac{\partial \iota_1}{\partial x}\bigg|_{x} \right\| + \left\| \frac{\partial \hat{\iota}}{\partial y}\bigg|_{\iota_1(x')} \right\| \cdot \left\| \frac{\partial \iota_1}{\partial x}\bigg|_{x} - \frac{\partial \iota_1}{\partial x}\bigg|_{x'} \right\| \\
&\le \ell(1+L)^{2t}\|\iota_1(x) - \iota_1(x')\| \cdot L + L^t \cdot \ell\|x - x'\| && (5.21b)\\
&\le \ell\left(L^2(1+L)^{2t} + L^t\right)\|x - x'\| && (5.21c)\\
&\le \ell(1+L)^{2(t+1)}\|x - x'\|,
\end{aligned}
$$

where we use the chain rule decomposition in (5.21a); we use the induction assumption in (5.21b); we use the assumption that ι_1 is L-Lipschitz in (5.21c).

Therefore, we have shown Lemma 5.A.2 by induction. □

Now we are ready to show Lemma 5.A.1.

We first discuss the intuition behind the proof. Since the multi-step dynamics is differentiable under Assumption 5.2.1, we only need to show an upper bound of $\left\| \frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau, \theta_{\tau:t-1}} \right\|$ that is exponentially decaying. Intuitively, we use the chain rule to decompose the partial derivative $\frac{\partial g_{t|\tau}}{\partial x_\tau}\Big|_{x_\tau, \theta_{\tau:t-1}}$ as the product of multiple partial derivatives $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}}$, where each time interval $[t_i : t_{i+1} - 1]$ has (approximately) length h. By the time-invariant contractive property, we know the norm of $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}$ can be upper bounded by $C\rho^h$ once it is realized on the trajectory $\{x_{t_i}, (\theta_{t_i})_{\times h}\}$, where the policy parameter is repeating. By the smoothness guarantee derived in Lemma 5.A.2, we can show the difference between $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}}$ and $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i}, (\theta_{t_i})_{\times h}}$ is in the order of $O(\varepsilon)$. This implies that the norm of $\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\Big|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}}$ can be bounded by $\hat{\rho}^h$ once ε is sufficiently small with respect to the gap $\left(\hat{\rho}^h - C\rho^h\right)$.

We present the formal proof of Lemma 5.A.1 below.

Proof of Lemma 5.A.1. We first show that, under a time-invariant policy parameter θ, starting from an arbitrary x_τ that satisfies $\|x_\tau\| \leq R_C$, we have $\|x_t\| \leq R_S + C\rho^{t-\tau}\|x_\tau\|$ for all $t \geq \tau$.

To see this, note that by time-invariant contractive perturbation, we have

$$\begin{aligned}\left\|g_{t|\tau}(x_\tau, \theta_{\times(t-\tau)})\right\| &\leq \left\|g_{t|\tau}(x_\tau, \theta_{\times(t-\tau)}) - g_{t|\tau}(0, \theta_{\times(t-\tau)})\right\| + \left\|g_{t|\tau}(0, \theta_{\times(t-\tau)})\right\| \\ &\leq C\rho^{t-\tau}\|x_\tau\| + R_S.\end{aligned}$$

Now, we show that if starting from x_τ that satisfies $\|x_\tau\| \leq \hat{R}_C$, the trajectory induced by an $\hat{\varepsilon}$-time-varying parameter sequence satisfies that $\|x_t\| \leq \hat{R}_S + C\rho^{t-\tau}\|x_\tau\|$ for all $t \geq \tau$. Since this ball is contained in $B(0, R_C)$, we know the time-invariant contractive perturbation always apply.

We show this statement by induction. Suppose the statement holds for all time steps $\tau, \ldots, t-1$. We see that

$$\begin{aligned}&\left\|g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g_{t|\tau}(x_\tau, \theta_{\times(t-\tau)})\right\| \\ \leq &\sum_{j=\tau}^{t} \left\|g_{t|\tau}(x_\tau, \theta_{\tau:j}, (\theta_t)_{\times(t-j-1)}) - g_{t|\tau}(x_\tau, \theta_{\tau:j-1}, (\theta_t)_{\times(t-j)})\right\| \\ \leq &\, CL_{g,u}L_{\pi,\theta} \sum_{j=\tau}^{t} \rho^{t-j-1}\left\|\theta_t - \theta_j\right\| \leq \frac{CL_{g,u}L_{\pi,\theta}\hat{\varepsilon}}{(1-\rho)^2}.\end{aligned}$$

Therefore, we see that

$$\begin{aligned}\left\|g_{t|\tau}(x_\tau, \theta_{\tau:t-1})\right\| &\leq \left\|g_{t|\tau}(x_\tau, \theta_{\times(t-\tau)})\right\| + \frac{CL_{g,u}L_{\pi,\theta}\hat{\varepsilon}}{(1-\rho)^2} \\ &\leq C\rho^{t-\tau}\|x_\tau\| + R_S + \frac{CL_{g,u}L_{\pi,\theta}\hat{\varepsilon}}{(1-\rho)^2} \\ &= C\rho^{t-\tau}\|x_\tau\| + \hat{R}_S.\end{aligned}$$

This finishes the proof of ε-time-varying stability with $\hat{\varepsilon}$ and $\hat{R}_S$.

Note that we can decompose the partial derivative $\left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_{\tau:t-1}}$ as

$$\begin{aligned}&\left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} \\ = &\left.\frac{\partial g_{t_p|t_{p-1}}}{\partial x_{t_{p-1}}}\right|_{x_{t_{p-1}}, \theta_{t_{p-1}:t_p-1}} \cdot \left.\frac{\partial g_{t_{p-1}|t_{p-2}}}{\partial x_{t_{p-2}}}\right|_{x_{t_{p-2}}, \theta_{t_{p-2}:t_{p-1}-1}} \cdots \left.\frac{\partial g_{t_1|t_0}}{\partial x_{t_0}}\right|_{x_{t_0}, \theta_{t_0:t_1-1}},\end{aligned} \tag{5.22}$$

where $t_0 = \tau, t_p = t$; $t_i = t_{i-1} + h$ holds for $i = 1, \ldots, p-1$, and $t_{p-1} < t_p \leq t_{p-1} + h$.

For $i \in [0, p-2]$, we obtain the following bounds on the bias of the partial derivatives of the multi-step dynamics:

$$\begin{aligned}
&\left\| \left.\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\right|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\right|_{x_{t_i}, (\theta_{t_i}) \times h} \right\| \\
\leq\ & \sum_{j=0}^{h-1} \left\| \left.\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\right|_{x_{t_i}, (\theta_{t_i})_{\times j}, \theta_{t_i+j:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\right|_{x_{t_i}, (\theta_{t_i})_{\times (j+1)}, \theta_{t_i+j+1:t_{i+1}-1}} \right\| && \text{(5.23a)} \\
\leq\ & \sum_{j=0}^{h-1} \left\| \left.\frac{\partial g_{t_i+j}}{\partial x_{t_i}}\right|_{x_{t_i}, (\theta_{t_i})_{\times j}} \right\| \cdot \\
&\left\| \left.\frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i+j:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i}, \theta_{t_i+j+1:t_{i+1}-1}} \right\| && \text{(5.23b)} \\
\leq\ & C \sum_{j=0}^{h-1} \rho^j \left\| \left.\frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i+j:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i}, \theta_{t_i+j+1:t_{i+1}-1}} \right\|, && \text{(5.23c)}
\end{aligned}$$

where we use the shorthand notation $\bar{x}_{t_i+j} = g_{t_i+j|t_i}(x_{t_i}, (\theta_{t_i})_{\times j})$. We use the triangle inequality in (5.23a), the chain rule decomposition in (5.23b). In (5.23c), we can apply the time-invariant contractive perturbation property because $\|x_\tau\| \leq \hat{R}_C$, which implies that $\left\|x_{t_i}\right\| \leq R_C$. Note that

$$\begin{aligned}
&\left\| \left.\frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i+j:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i}, \theta_{t_i+j+1:t_{i+1}-1}} \right\| \\
\leq\ & \left\| \left.\frac{\partial g_{t_{i+1}|(t_i+j+1)}}{\partial x_{t_i+j+1}}\right|_{\bar{x}_{t_i+j+1}, \theta_{t_i+j+1:t_{i+1}-1}} \right\| \cdot \\
&\left\| \left.\frac{\partial g_{(t_i+j+1)|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i+j}} - \left.\frac{\partial g_{(t_i+j+1)|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i}} \right\| \\
&+ \left\| \left.\frac{\partial g_{t_{i+1}|(t_i+j+1)}}{\partial x_{t_i+j+1}}\right|_{\bar{x}_{t_i+j+1}, \theta_{t_i+j+1:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|(t_i+j+1)}}{\partial x_{t_i+j+1}}\right|_{\bar{x}'_{t_i+j+1}, \theta_{t_i+j+1:t_{i+1}-1}} \right\| \cdot \\
&\left\| \left.\frac{\partial g_{(t_i+j+1)|(t_i+j)}}{\partial x_{t_i+j}}\right|_{\bar{x}_{t_i+j}, \theta_{t_i+j}} \right\| \\
\leq\ & \left(L_{g,x} + L_{g,u} L_{\pi,x}\right)^{h-j-1} \cdot \left((1 + L_{\pi,x}) \ell_{g,u} L_{\pi,\theta} + L_{g,u} \ell_{\pi,\theta}\right) \left\|\theta_{t_i+j} - \theta_{t_i}\right\| \\
&+ \left((1 + L_{\pi,x}) \left(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x}\right) + L_{g,u} \cdot \ell_{\pi,x}\right) \cdot \\
&\left(1 + L_{g,x} + L_{g,u} L_{\pi,x}\right)^{2(h-j)} \cdot L_{g,u} L_{\pi,\theta} \left\|\theta_{t_i+j} - \theta_{t_i}\right\| \\
\leq\ & C' \left(1 + L_{g,x} + L_{g,u} L_{\pi,x}\right)^{2h} \cdot j\varepsilon,
\end{aligned}$$

where we adopt the shorthand $\bar{x}'_{t_i+j+1} \coloneqq g_{t_i+j+1|t_i+j}(\bar{x}_{t_i+j}, \theta_{t_i+j})$, and we define the coefficient C' as

$$\begin{aligned} C' = {} & \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u} \cdot (L_{\pi,x} + L_{\pi,\theta})\right) + L_{g,u} \cdot (\ell_{\pi,x} + \ell_{\pi,\theta})\right) \cdot \\ & (L_{g,u} L_{\pi,\theta} + 1). \end{aligned}$$

Here, we use the chain rule and the triangle inequality in the first inequality. We can apply Lemma 5.A.2 and

$$\left\|\bar{x}'_{t_i+j+1} - \bar{x}_{t_i+j+1}\right\| \le L_{g,u} L_{\pi,\theta} \left\|\theta_{t_i+j} - \theta_{t_i}\right\|$$

in the second inequality because the trajectory induced by $(\theta_{t_i})_{\times j}, \theta_{t_i+j:t_{i+1}-1}$ always stay within the ball $B(0, R_x)$ where the Lipschitzness/smoothness of dynamics/policies hold. Substituting this into (5.23) gives

$$\begin{aligned} & \left\| \left.\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\right|_{x_{t_i}, \theta_{t_i:t_{i+1}-1}} - \left.\frac{\partial g_{t_{i+1}|t_i}}{\partial x_{t_i}}\right|_{x_{t_i}, (\theta_{t_i})\times h} \right\| \\ & \le CC' \left(1 + L_{g,x} + L_{g,u} L_{\pi,x}\right)^{2h} \varepsilon \sum_{j=0}^{h-1} \rho^j \cdot j \\ & \le \frac{CC'\rho \left(1 + L_{g,x} + L_{g,u} L_{\pi,x}\right)^{2h}}{(1-\rho)^2} \cdot \varepsilon \\ & \le \hat{\rho}^h - C\rho^h. \end{aligned} \tag{5.24}$$

Therefore, by (5.22), we see that

$$\begin{aligned} & \left\| \left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} \right\| \\ \le {} & \left\| \left.\frac{\partial g_{t_p|t_{p-1}}}{\partial x_{t_{p-1}}}\right|_{x_{t_{p-1}}, \theta_{t_{p-1}:t_p-1}} \right\| \cdot \left\| \left.\frac{\partial g_{t_{p-1}|t_{p-2}}}{\partial x_{t_{p-2}}}\right|_{x_{t_{p-2}}, \theta_{t_{p-2}:t_{p-1}-1}} \right\| \cdots \left\| \left.\frac{\partial g_{t_1|t_0}}{\partial x_{t_0}}\right|_{x_{t_0}, \theta_{t_0:t_1-1}} \right\| \\ \le {} & \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x}\right) + L_{g,u} \cdot \ell_{\pi,x}\right) \cdot \left(1 + L_{g,x} + L_{g,u} L_{\pi,x}\right)^{2h} \\ & \cdot (\hat{\rho}^h)^{p-1} \qquad\qquad (5.25\text{a}) \\ \le {} & \hat{C}(\hat{\rho})^{t-\tau}, \qquad\qquad (5.25\text{b}) \end{aligned}$$

where we use (5.24) in (5.25a); we use the definition of $\hat{C}$ in (5.25b). This finishes the proof of ε-time-varying contractive perturbation with $\hat{\epsilon}, \hat{R}_C, \hat{C}, \hat{\rho}$. □

5.B Proof of M-GAPS

Regret of M-GAPS under Convex Surrogate Costs

In this section, we provide a proof outline of the adaptive regret bound for GAPS. As we discussed in Section 5.3, the intuition behind GAPS is to mimic the ideal OGD update $\theta_{t+1} = \prod_\Theta(\theta_t - \eta_t \nabla F_t(\theta_t))$ with limited memory size and computational complexity. While the existing literature of OCO guarantees that the ideal OGD update with constant step size η of the order $1/\sqrt{T}$ achieves a policy regret of $O\left(\sqrt{T}\right)$, GAPS incurs an approximation error at every time step since it uses G_t (Algorithm 6) instead of $\nabla F_t(\theta_t)$ to implement gradient descent. We characterize how a per-step bias in the gradient estimation may affect the regret guarantee of the OGD in Theorem 5.B.1. We provide the proof later in this section.

Theorem 5.B.1. *Consider the update rule $\theta_{t+1} = \prod_\Theta(\theta_t - \eta G_t)$. Suppose Θ is a convex compact set with diameter D. If F_t is convex and $\|\nabla F_t(\theta)\| \leq W$ for all $\theta \in \Theta$, and $\|\nabla F_t(\theta_t) - G_t\| \leq \alpha$ holds for all time steps t, then, for arbitrary $I = [r : s] \subseteq \mathcal{T}$,*

$$\sum_{t=r}^{s} F_t(\theta_t) - \min_{\theta_I \in \Theta} \sum_{t=r}^{s} F_t(\theta_I) \leq \alpha D T + (W^2 + \alpha^2)\eta T + \frac{D^2}{2\eta}.$$

With Theorem 5.B.1, obtaining the policy regret bounds for GAPS reduces to showing both $|h_t(x_t, u_t) - F_t(\theta_t)|$ and $\|\nabla F_t(\theta_t) - G_t\|$ are in the order of $O(1/\sqrt{T})$. Here, we only consider the order of magnitude with respect to the horizon T for clarity. As we will show in Theorem 5.B.5 and Theorem 5.B.6, both of these quantities are in the order of $O(\eta)$ when GAPS adopts the learning rate η.

To obtain these results, we first show a lemma about the stability of the trajectory achieved by an ε-time-varying policy parameter sequence.

Lemma 5.B.2. *Suppose Assumptions 5.2.1 and 5.2.2 hold. For any starting state $x_\tau \in B_n(0, R_S + C\|x_0\|)$ and $\theta_{\tau:t-1} \in S_\varepsilon(\tau : t-1)$, the final state $x_t := g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$ satisfies $\|x_t\| \leq C\rho^{t-\tau}\|x_\tau\| + R_S$.*

Proof of Lemma 5.B.2. By ε-time-varying contractive perturbation, we see that

$$\left\|x_t - g_{t|\tau}(0, \theta_{\tau:t-1})\right\| \leq C\rho^{t-\tau}\|x_\tau\|.$$

Thus, by the triangle inequality, we see that

$$\|x_t\| \leq \left\|x_t - g_{t|\tau}(0, \theta_{\tau:t-1})\right\| + \left\|g_{t|\tau}(0, \theta_{\tau:t-1})\right\| \leq C\rho^{t-\tau}\|x_\tau\| + R_S,$$

where we use ε-time-varying stability in the last inequality. □

Next, we show a lemma about the contractive property of the partial derivatives of the multi-step dynamics.

Lemma 5.B.3 (Lipschitzness/Smoothness of the Multi-Step Dynamics)**.** *Suppose Assumptions 5.2.1 and 5.2.2 hold. Given two time steps $t > \tau$, for any $x_\tau, x'_\tau \in B_n(0, R_S{+}C\|x_0\|)$ and $\theta_\tau, \theta'_\tau \in \Theta, \theta_{\tau+1:t-1} \in S_\varepsilon(\tau{+}1 : t{-}1)$, if $x'_{\tau+1} := g_{\tau+1|\tau}(x'_\tau, \theta'_\tau)$ is also in $B_n(0, R_S + C\|x_0\|)$, the multi-step dynamical function $g_{t|\tau}$ satisfies that*

$$\begin{aligned}
&\left\| \left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} \right\| \le C_{L,g,x}\rho^{t-\tau}, \text{ and} \\
&\left\| \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} \right\| \le C_{L,g,\theta}\rho^{t-\tau}, \forall \theta_{\tau:t-1} \in S_\varepsilon(\tau : t-1), \\
&\left\| \left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} - \left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x'_\tau, \theta'_\tau, \theta_{\tau+1:t-1}} \right\| \\
\le{}& C_{\ell,g,(x,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,g,(x,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|, \\
&\left\| \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} - \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x'_\tau, \theta'_\tau, \theta_{\tau+1:t-1}} \right\| \\
\le{}& C_{\ell,g,(\theta,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,g,(\theta,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,
\end{aligned}$$

where $C_{L,g,x} = C, C_{L,g,\theta} = \frac{C L_{g,u} L_{\pi,\theta}}{\rho}$, and

$$\begin{aligned}
C_{\ell,g,(x,x)} &= \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u}L_{\pi,x}\right) + L_{g,x}\ell_{\pi,x}\right) C^3\rho^{-1}(1-\rho)^{-1}, \\
C_{\ell,g,(x,\theta)} &= \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u}L_{\pi,x}\right) + L_{g,x}\ell_{\pi,x}\right) C^3 L_{g,u}L_{\pi,\theta}\cdot \\
&\quad \rho^{-1}(1-\rho)^{-1} + \left((1 + L_{\pi,x})\ell_{g,u}L_{\pi,\theta} + L_{g,u}\ell_{\pi,\theta}\right) C\rho^{-1}(1-\rho)^{-1}, \\
C_{\ell,g,(\theta,x)} &= \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u}L_{\pi,x}\right) + L_{g,x}\ell_{\pi,x}\right)\left(L_{g,x} + L_{g,u}L_{\pi,x}\right)\cdot \\
&\quad C^3 L_{g,u}L_{\pi,\theta}\rho^{-2}(1-\rho)^{-1} \\
&\quad + C\left(L_{\pi,\theta}(\ell_{g,x} + \ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x}\right)\rho^{-1}, \\
C_{\ell,g,(\theta,\theta)} &= \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u}\cdot L_{\pi,x}\right) + L_{g,x}\cdot \ell_{\pi,x}\right) L_{g,u}^2 L_{\pi,\theta}^2 C^3\cdot \\
&\quad \rho^{-2}(1-\rho)^{-1} + \left(L_{g,u}\ell_{\pi,\theta} + \ell_{g,u}L_{\pi,\theta}^2\right) C\rho^{-1}.
\end{aligned}$$

Intuitively, Lemma 5.B.3 shows that the dependence of the state x_t on the previous state x_τ and θ_τ decays exponentially with respect to their time distance $t - \tau$. Specifically, recall that the multi-step dynamics $g_{t|\tau}$ writes x_t as a function of x_τ and $\theta_{\tau:t-1}$. When other variables are fixed, the Lipschitzness and smoothness constants with respect to x_τ and θ_τ are both $O(\rho^{t-\tau})$. While the contractive Lipschitzness on x_τ is automatically guaranteed by ε-time-varying contractive perturbation (Definition 5.2.2), we use this property and the chain rule decomposition to show the Lipschitzness on θ_τ and the smoothness.

The first inequality in Lemma 5.B.3 directly follows from ε-time-varying contractive perturbation. To reflect the main technical difficulty, we show the third inequality here with the assumption that the first two inequalities hold. We provide the proof of other inequalities later in this section.

Proof of the 3rd inequality in Lemma 5.B.3. Note that we have the chain rule decomposition

$$\begin{aligned} \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} &= \frac{\partial g_{t|t-1}}{\partial x_{t-1}}\bigg|_{x_{t-1},\theta_{t-1}} \cdot \frac{\partial g_{t-1|t-2}}{\partial x_{t-2}}\,\big|_{x_{t-2},\theta_{t-2}} \cdots \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau}, \\ \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} &= \frac{\partial g_{t|t-1}}{\partial x_{t-1}}\bigg|_{x'_{t-1},\theta_{t-1}} \cdot \frac{\partial g_{t-1|t-2}}{\partial x_{t-2}}\bigg|_{x'_{t-2},\theta_{t-2}} \cdots \frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau}, \end{aligned} \tag{5.26}$$

where we use the notation $x_{\tau'} = g_{\tau'|\tau}(x_\tau, \theta_{\tau:\tau'-1})$ and $x'_{\tau'} = g_{\tau'|\tau}(x'_\tau, \theta'_\tau, \theta_{\tau+1:\tau'-1})$ for $\tau' \in [\tau+1 : t-1]$.

Note that for any $i \in [1 : t-\tau]$ and any $\theta'_{t-i} \in \Theta$, we have the decomposition

$$\begin{aligned} &\frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},\theta'_{t-i}} \\ ={}& \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} + \frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x_{t-i},u_{t-i}} \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} \\ &- \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},\theta'_{t-i}} \\ ={}& \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} + \left(\frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x_{t-i},u_{t-i}} - \frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}}\right) \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} \\ &+ \frac{\partial g_{t-i}}{\partial u_{t-i}}\bigg|_{x'_{t-i},u'_{t-i}} \left(\frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} - \frac{\partial \pi_{t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},\theta'_{t-i}}\right), \end{aligned}$$

where we use the notation $u_{t-i} = \pi_{t-i}(x_{t-i}, \theta_{t-i}), u'_{t-i} = \pi_{t-i}(x'_{t-i}, \theta'_{t-i})$. Taking norms on both sides of the equation and applying the triangle inequality gives

$$\begin{aligned} &\left\| \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\bigg|_{x_{t-i},\theta_{t-i}} - \frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\bigg|_{x'_{t-i},\theta'_{t-i}} \right\| \\ \leq{}& \ell_{g,x}\left\|x_{t-i} - x'_{t-i}\right\| + \ell_{g,u}\left\|\pi_{t-i}(x_{t-i}, \theta_{t-i}) - \pi_{t-i}(x'_{t-i}, \theta'_{t-i})\right\| \\ &+ L_{\pi,x}\left(\ell_{g,x}\left\|x_{t-i} - x'_{t-i}\right\| + \ell_{g,u}\left\|\pi_{t-i}(x_{t-i}, \theta_{t-i}) - \pi_{t-i}(x'_{t-i}, \theta'_{t-i})\right\|\right) \\ &+ L_{g,u} \cdot \left(\ell_{\pi,x}\left\|x_{t-i} - x'_{t-i}\right\| + \ell_{\pi,\theta}\left\|\theta_{t-i} - \theta'_{t-i}\right\|\right) &\text{(5.27a)} \\ \leq{}& \left((1 + L_{\pi,x})\left(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x}\right) + L_{g,u} \cdot \ell_{\pi,x}\right)\left\|x_{t-i} - x'_{t-i}\right\| \\ &+ \left((1 + L_{\pi,x})\ell_{g,u}L_{\pi,\theta} + L_{g,u}\ell_{\pi,\theta}\right)\left\|\theta_{t-i} - \theta'_{t-i}\right\|, &\text{(5.27b)} \end{aligned}$$

where we use Assumption 5.2.1 and the definition of u_{t-i}, u'_{t-i} in (5.27a); and Assumption 5.2.1 in (5.27b). Therefore, by (5.26) and (5.27), we see that

$$\begin{aligned}
&\left\| \left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_{\tau:t-1}} - \left.\frac{\partial g_{t|\tau}}{\partial x_\tau}\right|_{x'_\tau, \theta_{\tau:t-1}} \right\| \\
\le\ & \sum_{i=1}^{t-\tau-1} \left(\left\| \prod_{\tau'=1}^{i-1} \left.\frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\right|_{x'_{t-\tau'}, \theta_{t-\tau'}} \right\| \cdot \left\| \left.\frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\right|_{x_{t-i}, \theta_{t-i}} - \left.\frac{\partial g_{t-i+1|t-i}}{\partial x_{t-i}}\right|_{x'_{t-i}, \theta_{t-i}} \right\| \cdot \right. \\
& \left. \left\| \prod_{\tau'=i+1}^{t-\tau} \left.\frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\right|_{x_{t-\tau'}, \theta_{t-\tau'}} \right\| \right) \\
& + \left\| \prod_{\tau'=1}^{t-\tau-1} \left.\frac{\partial g_{t-\tau'+1|t-\tau'}}{\partial x_{t-\tau'}}\right|_{x'_{t-\tau'}, \theta_{t-\tau'}} \right\| \cdot \left\| \left.\frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\right|_{x_\tau, \theta_\tau} - \left.\frac{\partial g_{\tau+1|\tau}}{\partial x_\tau}\right|_{x'_\tau, \theta'_\tau} \right\| \\
\le\ & \sum_{i=1}^{t-\tau} (C\rho^{i-1}) \cdot \left((1 + L_{\pi,x}) \left(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x} \right) + L_{g,x} \cdot \ell_{\pi,x} \right) \left\| x_{t-i} - x'_{t-i} \right\| \cdot \\
& (C\rho^{t-\tau-i}) + C\rho^{t-\tau-1} \cdot \left((1 + L_{\pi,x}) \ell_{g,u} L_{\pi,\theta} + L_{g,u} \ell_{\pi,\theta} \right) \left\| \theta_\tau - \theta'_\tau \right\| \qquad (5.28a) \\
=\ & \left((1 + L_{\pi,x}) \left(\ell_{g,x} + \ell_{g,u} \cdot L_{\pi,x} \right) + L_{g,x} \cdot \ell_{\pi,x} \right) C^2 \cdot \rho^{t-\tau-1} \sum_{i=1}^{t-\tau} \left\| x_{t-i} - x'_{t-i} \right\| \\
& + C\rho^{t-\tau-1} \cdot \left((1 + L_{\pi,x}) \ell_{g,u} L_{\pi,\theta} + L_{g,u} \ell_{\pi,\theta} \right) \left\| \theta_\tau - \theta'_\tau \right\| \\
\le\ & C_{\ell,g,(x,x)} \cdot \rho^{t-\tau} \left\| x_\tau - x'_\tau \right\| + C_{\ell,g,(x,\theta)} \cdot \rho^{t-\tau} \left\| \theta_\tau - \theta'_\tau \right\|, \qquad (5.28b)
\end{aligned}$$

where we use the ε-time-varying contractive perturbation property and (5.27) in (5.28a); we use the first two inequalities to bound $\left\| x_{t-i} - x'_{t-i} \right\| \le C\rho^{t-i-\tau} \left\| x_\tau - x'_\tau \right\| + \frac{C L_{g,u} L_{\pi,\theta}}{\rho} \cdot \rho^{t-i-\tau} \left\| \theta_\tau - \theta'_\tau \right\|$ in (5.28b). □

Since we will need more general forms of policy sequence later to bound $\|G_t - \nabla F_t\|$ than the sequence with small step sizes, we state the contractive Lipschitzness and smoothness of the multi-step cost function $h_{t|\tau}$. This is an implication of Lemma 5.B.3 because for any $\tau < t$, the previous state x_τ and previous policy parameter θ_τ can only affect the current stage cost c_t by affecting the current state x_t. We formalize this result in Corollary 5.B.4 and provide the detailed proof later in this section.

Corollary 5.B.4 (Lipschitzness/Smoothness of the Multi-Step Costs)**.** *Under the same assumptions as Lemma 5.B.3, let*

$$x_t := g_{t|\tau}(x_\tau, \theta_{\tau:t-1}), u_t := \pi_t(x_t, \theta_t);\ \textit{and}$$
$$x'_t := g_{t|\tau}(x'_\tau, \theta'_\tau, \theta_{\tau+1:t-1}), u'_t := \pi_t(x'_t, \theta_t).$$

Then, the partial derivatives of the multi-step cost function $h_{t|\tau}$ satisfy the following inequalities:

$$\left\| \frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau, \theta_{\tau:t}} \right\| \le C_{L,h,x}\rho^{t-\tau}, \quad \left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:t}} \right\| \le C_{L,h,\theta}\rho^{t-\tau},$$

$$\left\| \frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau, \theta_\tau, \theta_{\tau+1:t}} - \frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau, \theta'_\tau, \theta_{\tau+1:t}} \right\|$$
$$\le C_{\ell,h,(x,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,h,(x,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,$$

$$\left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_\tau, \theta_{\tau+1:t}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\bigg|_{x'_\tau, \theta'_\tau, \theta_{\tau+1:t}} \right\|$$
$$\le C_{\ell,h,(\theta,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,h,(\theta,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|,$$

where $C_{L,h,x} = L_h C(1 + L_{\pi,x})$, $C_{L,h,\theta} = L_h \max\{C_{L,g,\theta}(1 + L_{\pi,x}), L_{\pi,\theta}\}$, and

$$\begin{aligned}
C_{\ell,h,(x,x)} &= L_h(1 + L_{\pi,x})C_{\ell,g,(x,x)} \\
&\quad + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_h\ell_{\pi,x})C_{L,g,x}^2, \\
C_{\ell,h,(x,\theta)} &= L_h(1 + L_{\pi,x})C_{\ell,g,(x,\theta)} + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) \\
&\quad + L_h\ell_{\pi,x})C_{L,g,x}C_{L,g,\theta}, \\
C_{\ell,h,(\theta,x)} &= L_h(1 + L_{\pi,x})C_{\ell,g,(\theta,x)} \\
&\quad + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_h\ell_{\pi,x})C_{L,g,x}C_{L,g,\theta}, \\
C_{\ell,h,(\theta,\theta)} &= L_h(1 + L_{\pi,x})C_{\ell,g,(\theta,\theta)} \\
&\quad + ((\ell_{f,x} + \ell_{f,u}L_{\pi,x})(1 + L_{\pi,x}) + L_h\ell_{\pi,x})C_{L,g,\theta}^2.
\end{aligned}$$

With the help of Lemma 5.B.3 and Corollary 5.B.4, we first bound the cost difference $|h_t(x_t, u_t) - F_t(\theta_t)|$ in Theorem 5.B.5. This inequality bounds the difference between the actual stage cost $h_t(x_t, u_t)$ incurred by GAPS and the ideal cost $F_t(\theta_t)$. Besides this inequality, in Theorem 5.B.5, we also bound the distance between GAPS' trajectory and the imaginary trajectory if the same policy parameter θ_t had been used from time 0 to time t, which will be useful for showing Theorem 5.B.6 later in this section.

Theorem 5.B.5. *Suppose Assumptions 5.2.1 and 5.2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of M-GAPS (Algorithm 6) with a constant learning rate $\eta \le \frac{(1-\rho)\varepsilon}{C_{L,h,\theta}}$. Then, both $\|G_t\|$ and $\|\nabla F_t(\theta_t)\|$ are upper bounded by $\frac{C_{L,h,\theta}}{1-\rho}$, and the following inequalities holds for any two time steps τ, t ($\tau \le t$):*

$$\|\theta_t - \theta_\tau\| \le \frac{C_{L,h,\theta}}{1-\rho} \cdot (t - \tau)\eta, \text{ and}$$

$$\|x_\tau - \hat{x}_\tau(\theta_t)\| \leq \frac{C_{L,h,\theta} C_{L,g,\theta} \rho}{(1-\rho)^2} \left((t-\tau) + \frac{1}{1-\rho}\right) \cdot \eta,$$

where we use the notation $\hat{x}_\tau(\theta) \coloneqq g_{0,\tau}(x_0, \theta_{\times(\tau+1)}), \forall \theta \in \Theta$. Further, we have that

$$|h_t(x_t, u_t) - F_t(\theta_t)| \leq \frac{C_{L,h,\theta} C_{L,g,\theta} L_h (1 + L_{\pi,x}) \rho}{(1-\rho)^3} \cdot \eta.$$

In addition, for any parameter sequence $\tilde{\theta}_{0:t} \in \Theta^{t+1}$, let $\tilde{x}_t$ and $\tilde{u}_t$ be the state/control action achieved by this sequence $\tilde{x}_t \coloneqq g_{t|0}(x_0, \tilde{\theta}_{0:t-1})$ and $\tilde{u}_t \coloneqq \pi_t(\tilde{x}_t, \tilde{\theta}_t)$. If $\|\tilde{x}_t\| \leq \min\{R_C, R_x\}$ holds for all t, then the following inequality holds for all time t:

$$\left|h_t(\tilde{x}_t, \tilde{u}_t) - F_t(\tilde{\theta}_t)\right| \leq \frac{C L_{\pi,\theta} L_{g,x} L_h (1 + L_{\pi,x})}{1-\rho} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \left\|\tilde{\theta}_{\tau+1} - \tilde{\theta}_\tau\right\|.$$

To show Theorem 5.B.5, we first derive a uniform upper bound on the norm the estimated gradient G_t, which implies that the policy parameter sequence does not vary too quickly, i.e., it is in the same order as the constant learning rate η. We then leverage strong contractive perturbation to bound $\|x_\tau - \hat{x}_\tau(\theta_t)\|$ and use it to bound $|h_t(x_t, u_t) - F_t(\theta_t)|$ by the Lipschitzness of h_t. We provide the detailed proof later in this section.

In Theorem 5.B.6 below, we bound the difference between the estimated gradient G_t used by GAPS and the ideal gradient $\nabla F_t(\theta_t)$ used by the ideal OGD.

Theorem 5.B.6 (Gradient Bias)**.** *Suppose Assumptions 5.2.1 and 5.2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t \in \mathcal{T}}$ denote the trajectory of M-GAPS (Algorithm 6) with learning rate $\eta \leq \frac{(1-\rho)\varepsilon}{C_{L,h,\theta}}$. Then, the following holds for all $\tau \leq t$:*

$$\left\| \frac{\partial h_{t|0}}{\partial \theta_\tau}\bigg|_{x_0, \theta_{0:t}} - \frac{\partial h_{t|0}}{\partial \theta_\tau}\bigg|_{x_0, (\theta_t)_{\times(t+1)}} \right\| = O\left(\left(\frac{1}{(1-\rho)^4} + \frac{t-\tau}{(1-\rho)^3} + \frac{(t-\tau)^2}{(1-\rho)^2}\right)\rho^{t-\tau} \cdot \eta\right),$$

Further, we see that

$$\|G_t - \nabla F_t(\theta_t)\| \leq O\left(\frac{\eta}{(1-\rho)^5}\right).$$

(See Theorem 5.B.9 for the detailed expressions.)

The key technique we used to show Theorem 5.B.6 is a sequential decomposition of the error based on the triangle inequality. Specifically, note that Corollary 5.B.4 only allow us to compare the partial derivatives when $\theta_{\tau+1:t}$ are fixed and the only perturbations are on x_τ and θ_τ. To compare the partial derivatives realized on two trajectory instances $(\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)})$ and $(x_\tau, \theta_{\tau:t})$, we change the parameters sequentially one by one, following the path

$$(\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)}) \longrightarrow (x_\tau, \theta_\tau, (\theta_t)_{\times(t-\tau)}) \longrightarrow (x_\tau, \theta_{\tau:\tau+1}, (\theta_t)_{\times(t-\tau-1)}) \longrightarrow \cdots \\ \longrightarrow (x_\tau, \theta_{\tau:t}).$$

The bounds in Theorems 5.B.5 and 5.B.6 show that we can achieve our desired bounds $|h_t(x_t, u_t) - F_t(\theta_t)| = O(1/\sqrt{T})$ and $\|\nabla F_t(\theta_t) - G_t\| = O(1/\sqrt{T})$ if we set the learning rate and the buffer length to be $O(1/\sqrt{T})$ and $O(\log T)$, respectively. Substituting these bounds into Theorem 5.B.1 with a more careful analysis on the order of the factor $1/(1-\rho)$ will finish the proof of Theorem 5.3.2. The detailed proof can be found in the next subsection.

Detailed Statement and Proofs of Theorems 5.3.1 and 5.3.2

We restate Theorem 5.3.1 with detailed expressions in Theorem 5.B.7.

Theorem 5.B.7. *Suppose Assumptions 5.2.1 and 5.2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of M-GAPS (Algorithm 6) with buffer size B and learning rate $\eta_t = \eta \leq \frac{(1-\rho)\varepsilon}{C_{L,h,\theta}}$, where $C_{L,h,\theta}$ is defined in Corollary 5.B.4. Then, we have*

$$\begin{aligned} |h_t(x_t, u_t) - F_t(\theta_t)| &\leq \frac{C_{L,h,\theta} C_{L,g,\theta} L_h (1 + L_{\pi,x})\rho}{(1-\rho)^3} \cdot \eta, \text{ and} \\ \|G_t - \nabla F_t(\theta_t)\| &\leq \left(\hat{C}_0 (1-\rho)^{-1} + (\hat{C}_1 + \hat{C}_2)(1-\rho)^{-2} + \hat{C}_2 (1-\rho)^{-3}\right)\eta, \end{aligned} \tag{5.29}$$

where $\hat{C}_0, \hat{C}_1, \hat{C}_2$ are defined in Theorem 5.B.9.

Proof of Theorem 5.B.7. Theorem 5.B.7 directly followed from Theorems 5.B.5 and 5.B.6. □

We restate Theorem 5.3.2 with detailed expressions in Theorem 5.B.8.

Theorem 5.B.8. *Under the same assumptions as Theorem 5.3.1, if we additionally assume the surrogate stage cost F_t is convex for every time step t, then M-GAPS achieves the adaptive regret bound*

$$R^A(T) \leq \left(\frac{C_{L,h,\theta}^2}{(1-\rho)^3} + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right) D\right)\eta T + \frac{D^2}{2\eta}$$

$$+ 2\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)^2 D^2\eta^3 T.$$

Proof of Theorem 5.B.8. The first two inequalities are shown in Theorem 5.B.5 and Theorem 5.B.6. Thus, we focus on the adaptive regret part here in the proof.

Fix a time interval $I = [r : s] \subseteq \mathcal{T}$ and let θ_I be an arbitrary policy parameter in Θ. By Theorem 5.B.1 and Theorem 5.B.6, we see that the sequence of policy parameters of the online policy satisfies that

$$\sum_{t=r}^{s} F_t(\theta_t) - \sum_{t=r}^{s} F_t(\theta_I) \leq (W^2 + \alpha^2)\eta T + \frac{D^2}{2\eta} + \alpha D T, \tag{5.30}$$

where $W = \frac{C_{L,h,\theta}}{1-\rho}$ by Theorem 5.B.5 and $\alpha = \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta$.

Note that by definition, we have $F_t(\theta_I) = h_t(\hat{x}_t(\theta_I), \hat{u}_t(\theta_I))$ and by Theorem 5.B.5, we have

$$|h_t(x_t, u_t) - F_t(\theta_t)| \leq \frac{C_{L,h,\theta} C_{L,g,\theta} L_h (1 + L_{\pi,x})\rho}{(1-\rho)^3} \cdot \eta. \tag{5.31}$$

Substituting these into (5.30) gives that

$$\begin{aligned}
&\sum_{t=r}^{s} h_t(x_t, u_t) - \sum_{t=r}^{s} h_t(\hat{x}_t(\theta_I), \hat{u}_t(\theta_I)) \\
\leq\ &\left(\sum_{t=r}^{s} F_t(\theta_t) - \sum_{t=r}^{s} F_t(\theta_I)\right) + \sum_{t=r}^{s} |h_t(x_t, u_t) - F_t(\theta_t)| \\
\leq\ &\frac{D^2}{2\eta} + \alpha D T + \left(W^2 + \alpha^2 + \frac{C_{L,h,\theta} C_{L,g,\theta} L_h (1 + L_{\pi,x})\rho}{(1-\rho)^3}\right) \cdot \eta T \\
\leq\ &\left(\frac{C_{L,h,\theta}^2}{(1-\rho)^3} + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right) D\right)\eta T + \frac{D^2}{2\eta} \\
&+ 2\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)^2 D^2\eta^3 T,
\end{aligned}$$

where we used (5.30) and (5.31) in the second inequality. □

Proof of Theorem 5.B.1

Our proof is inspired by the proof of Theorem 2.1 in Bansal and Gupta, 2019. For a fixed time interval $I = [r : s] \subseteq \mathcal{T}$ and $\theta_I \in \Theta$, we consider the potential function $\Phi_t = \frac{1}{2\eta}\|\theta_t - \theta_I\|^2$. Note that θ_I satisfies $\|\theta_r - \theta_I\| \leq D$ because we assume $\mathrm{diam}(\Theta) \leq D$. To simplify the notation, we define $\theta'_{t+1} = \theta_t - \eta G_t$.

By proposition 2.2 in Bansal and Gupta, 2019, we see that the potential change between two consecutive steps can be bounded by

$$\begin{aligned}\frac{1}{2}(\|\theta_{t+1}-\theta_I\|^2-\|\theta_t-\theta_I\|^2) &\le \frac{1}{2}(\left\|\theta'_{t+1}-\theta_I\right\|^2-\|\theta_t-\theta_I\|^2)\\ &= \langle\theta'_{t+1}-\theta_t,\theta_t-\theta_I\rangle+\frac{1}{2}\left\|\theta'_{t+1}-\theta_t\right\|^2\\ &= \eta\langle G_t,\theta_I-\theta_t\rangle+\frac{\eta^2}{2}\|G_t\|^2.\end{aligned}$$

Using this inequality, we see that

$$\begin{aligned}&F_t(\theta_t)-F_t(\theta_I)+\Phi_{t+1}-\Phi_t\\ &=F_t(\theta_t)-F_t(\theta_I)+\langle G_t,\theta_I-\theta_t\rangle+\frac{\eta}{2}\|G_t\|^2\\ &=F_t(\theta_t)-F_t(\theta_I)+\langle\nabla F_t(\theta_t)+(G_t-\nabla F_t(\theta_t)),\theta_I-\theta_t\rangle\\ &\quad+\frac{\eta}{2}\|\nabla F_t(\theta_t)+(G_t-\nabla F_t(\theta_t))\|^2\\ &\le F_t(\theta_t)-F_t(\theta_I)+\langle\nabla F_t(\theta_t),\theta_I-\theta_t\rangle+\langle G_t-\nabla F_t(\theta_t),\theta_I-\theta_t\rangle+\eta\|\nabla F_t(\theta_t)\|^2\\ &\quad+\eta\|G_t-\nabla F_t(\theta_t)\|^2 &\text{(5.32a)}\\ &\le 0+\|G_t-\nabla F_t(\theta_t)\|\cdot\|\theta_I-\theta_t\|+\eta\|\nabla F_t(\theta_t)\|^2+\eta\alpha^2 &\text{(5.32b)}\\ &\le \alpha D+W^2\eta+\eta\alpha^2, &\text{(5.32c)}\end{aligned}$$

where we used the triangle inequality and the AM-GM inequality in (5.32a); we used the assumption that F_t is convex, $\|G_t-\nabla F_t(\theta_t)\|\le\alpha$, and the Cauchy-Schwarz inequality in (5.32b); and we used the assumptions $\|G_t-\nabla F_t(\theta_t)\|\le\alpha$, $\mathrm{diam}(\Theta)\le D$, $\|\nabla F_t(\theta_t)\|\le W$ in (5.32c).

Summing (5.32) over the time interval $[r:s]$ gives that

$$\begin{aligned}\sum_{t=r}^{s}(F_t(\theta_t)-F_t(\theta_I)) &\le (s-r)\cdot\left(\alpha D+W^2\eta+\eta\alpha^2\right)+(\Phi_r-\Phi_{s+1})\\ &\le \left(\alpha D+W^2\eta+\eta\alpha^2\right)T+\frac{D^2}{2\eta},\end{aligned}$$

where we used $\mathrm{diam}(\Theta)\le D$ and $\Phi_{s+1}\ge 0$ in the last inequality. Since this inequality holds for any time interval $I=[r:s]$ and $\theta_I\in\Theta$, this finishes the proof of the first part of Theorem 5.B.1.

Proof of Inequalities 1,2, and 4 in Lemma 5.B.3

The first inequality directly follows from ε-time-varying contractive perturbation (Definition 5.2.2). For the second inequality, when $t=\tau+1$, note that $\frac{\partial g_{\tau+1|\tau}}{\partial\theta_\tau}\Big|_{x_\tau,\theta_\tau}=$

$\frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau}$, where $u_\tau = \pi_\tau(x_\tau, \theta_\tau)$. Taking norms of both sides of the equation gives

$$\left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| = \left\| \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| \le \left\| \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} \right\| \cdot \left\| \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| \le L_{g,u} L_{\pi,\theta}.$$

When $t > \tau + 1$, we see that

$$\begin{aligned}\left\| \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:t-1}} \right\| &= \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| \\ &\le \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| \le \frac{C_0 L_{g,u} L_{\pi,\theta}}{\rho} \cdot \rho^{t-\tau},\end{aligned}$$

where $x_{\tau+1} = g_{\tau+1|\tau}(x_\tau, \theta_\tau)$.

For the last inequality of Lemma 5.B.3, when $t = \tau + 1$, we see that

$$\begin{aligned}&\left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \\ &= \left\| \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x'_\tau,u'_\tau} \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \\ &\le \left\| \left(\frac{\partial g_\tau}{\partial u_\tau}\Big|_{x_\tau,u_\tau} - \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x'_\tau,u'_\tau} \right) \cdot \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\| + \left\| \frac{\partial g_\tau}{\partial u_\tau}\Big|_{x'_\tau,u'_\tau} \left(\frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial \pi_\tau}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right) \right\| \quad (5.33a) \\ &\le L_{\pi,\theta} \left(\ell_{g,x} \|x_\tau - x'_\tau\| + \ell_{g,u} \|u_\tau - u'_\tau\| \right) + L_{g,u} \left(\ell_{\pi,x} \|x_\tau - x'_\tau\| + \ell_{\pi,\theta} \|\theta_\tau - \theta'_\tau\| \right) \quad (5.33b) \\ &\le \left(L_{\pi,\theta}(\ell_{g,x} + \ell_{g,u} L_{\pi,x}) + L_{g,u} \ell_{\pi,x} \right) \|x_\tau - x'_\tau\| + (L_{\pi,\theta}^2 \ell_{g,u} + L_{g,u} \ell_{\pi,\theta}) \|\theta_\tau - \theta'_\tau\|, \quad (5.33c)\end{aligned}$$

where we use the notations $u_\tau = \pi_\tau(x_\tau, \theta_\tau), u'_\tau = \pi_\tau(x_\tau, \theta'_\tau)$. We use the triangle inequality in (5.33a); we use Assumption 5.2.1 in both (5.33b) and (5.33c).

When $t > \tau + 1$, we see that

$$\begin{aligned}&\left\| \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\ &\le \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} - \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \cdot \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x'_\tau,\theta'_\tau} \right\| \quad (5.34a) \\ &\le \left\| \left(\frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x_{\tau+1},\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\Big|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right) \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau} \right\|\end{aligned}$$

$$
\begin{aligned}
&+\left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\bigg|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \cdot \left(\frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\bigg|_{x'_\tau,\theta'_\tau} \right) \right\| && (5.34b)\\
&\le \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\bigg|_{x_{\tau+1},\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\bigg|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_\tau} \right\| \\
&\quad + \left\| \frac{\partial g_{t|\tau+1}}{\partial x_{\tau+1}}\bigg|_{x'_{\tau+1},\theta_{\tau+1:t-1}} \right\| \cdot \left\| \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_\tau} - \frac{\partial g_{\tau+1|\tau}}{\partial \theta_\tau}\bigg|_{x'_\tau,\theta'_\tau} \right\| \\
&\le \frac{\left((1+L_{\pi,x})\left(\ell_{g,x}+\ell_{g,u}\cdot L_{\pi,x}\right) + L_{g,x}\cdot \ell_{\pi,x}\right) C_0^3}{\rho(1-\rho)} \cdot \rho^{t-\tau-1} \cdot \left\|x_{\tau+1} - x'_{\tau+1}\right\| \cdot L_{g,u}L_{\pi,\theta} \\
&\quad + C_0 \cdot \rho^{t-\tau-1} \cdot \left(L_{\pi,\theta}(\ell_{g,x}+\ell_{g,u}L_{\pi,x}) + L_{g,u}\ell_{\pi,x}\right) \left\|x_\tau - x'_\tau\right\| \\
&\quad + C_0 \cdot \rho^{t-\tau-1} \cdot (L_{\pi,\theta}^2 \ell_{g,u} + L_{g,u}\ell_{\pi,\theta}) \left\|\theta_\tau - \theta'_\tau\right\| && (5.34c)\\
&\le C_{\ell,g,(\theta,x)}\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\| + C_{\ell,g,(\theta,\theta)}\rho^{t-\tau}\left\|\theta_\tau - \theta'_\tau\right\|, && (5.34d)
\end{aligned}
$$

where we use the notations $x_{\tau+1} = g_{\tau+1|\tau}(x_\tau, \theta_\tau), x'_{\tau+1} = g_{\tau+1|\tau}(x_\tau, \theta'_\tau)$. We use the chain rule decomposition in (5.34a); we use the triangle inequality in (5.34b); we use the first and the third inequality of Lemma 5.B.3 as well as (5.33) in (5.34c); we use the first two inequalities of Lemma 5.B.3 in (5.34d).

Proof of Corollary 5.B.4

To show the first inequality, note that

$$
\frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t}} = \left(\frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t} + \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}}, \qquad (5.35)
$$

where $x_t = g_{t|\tau}(x_\tau, \theta_{\tau:t-1}), u_t = \pi_t(x_t, \theta_t)$. Thus, by ε-time-varying contractive perturbation, we see that

$$
\begin{aligned}
\left\| \frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t}} \right\| &\le \left(\left\| \frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t} \right\| + \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \right\| \cdot \left\| \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right\| \right) \cdot \left\| \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}} \right\| \\
&\le L_h(1+L_{\pi,x}) \cdot C\rho^{t-\tau}.
\end{aligned}
$$

For the second inequality, when $\tau = t$, since $x_t \in B_n(0, R'_x)$ and $u_t \in B_m(0, R'_u)$, we see that

$$
\left\| \frac{\partial h_{t|t}}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| = \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| \le \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \right\| \cdot \left\| \frac{\partial \pi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| \le L_h L_{\pi,\theta}.
$$

When $\tau < t$, the second inequality can be shown similarly with the first inequality in Corollary 5.B.4 because we have the chain-rule decomposition

$$
\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_{\tau:t}} = \left(\frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t} + \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau,\theta_{\tau:t-1}}. \qquad (5.36)
$$

Applying Lemma 5.B.3 gives that $C_{L,h,\theta} = L_h C_{L,g,\theta}(1+L_{\pi,x})$. Thus, we have proved the first two inequalities.

For the third inequality, using (5.35), we see that

$$\begin{aligned}
&\left\| \frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \frac{\partial h_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t}} \right\| \\
\le{}& \left\| \frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\| \\
&+ \left\| \left(\frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t} - \frac{\partial h_t}{\partial x_t}\bigg|_{x'_t,u'_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&+ \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \cdot \left(\frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right) \right\| \\
&+ \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} \cdot \left(\frac{\partial \pi_t}{\partial x_t}\bigg|_{x_t,\theta_t} - \frac{\partial \pi_t}{\partial x_t}\bigg|_{x'_t,\theta_t} \right) \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\| \\
&+ \left\| \left(\frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x'_t,u'_t} \right) \cdot \frac{\partial \pi_t}{\partial x_t}\bigg|_{x'_t,\theta_t} \cdot \frac{\partial g_{t|\tau}}{\partial x_\tau}\bigg|_{x'_\tau,\theta'_\tau,\theta_{\tau+1:t-1}} \right\|
\end{aligned} \tag{5.37a}$$

$$\begin{aligned}
\le{}& L_h \rho^{t-\tau} \left(C_{\ell,g,(x,x)} \|x_\tau - x'_\tau\| + C_{\ell,g,(x,\theta)} \|\theta_\tau - \theta'_\tau\| \right) \\
&+ \left(\ell_{f,x} \|x_t - x'_t\| + \ell_{f,u} \|u_t - u'_t\| \right) \cdot C_{L,g,x} \rho^{t-\tau} \\
&+ L_h L_{\pi,x} \rho^{t-\tau} \left(C_{\ell,g,(x,x)} \|x_\tau - x'_\tau\| + C_{\ell,g,(x,\theta)} \|\theta_\tau - \theta'_\tau\| \right) \\
&+ L_h \ell_{\pi,x} \|x_t - x'_t\| \cdot C_{L,g,x} \rho^{t-\tau} + \left(\ell_{f,x} \|x_t - x'_t\| + \ell_{f,u} \|u_t - u'_t\| \right) \cdot L_{\pi,x} C_{L,g,x} \rho^{t-\tau},
\end{aligned} \tag{5.37b}$$

where we use the notations $x_t = g_{t|\tau}(x_\tau, \theta_{\tau:t-1})$, $x'_t = g_{t|\tau}(x'_\tau, \theta_{\tau:t-1})$, $u_t = \pi_t(x_t, \theta_t)$, and $u'_t = \pi_t(x'_t, \theta_t)$. We use (5.35) and the triangle inequality in (5.37a); we use Lemma 5.B.3 in (5.37b). Note that by the first two inequalities in Lemma 5.B.3, we have

$$\begin{aligned}
\|x_t - x'_t\| &\le \rho^{t-\tau} \left(C_{L,g,x} \|x_\tau - x'_\tau\| + C_{L,g,\theta} \|\theta_\tau - \theta'_\tau\| \right), \\
\|u_t - u'_t\| &\le L_{\pi,x} \rho^{t-\tau} \left(C_{L,g,x} \|x_\tau - x'_\tau\| + C_{L,g,\theta} \|\theta_\tau - \theta'_\tau\| \right).
\end{aligned}$$

Substituting these two inequalities into (5.37) finishes the proof of the third inequality.

For the last inequality, when $\tau = t$, we have that

$$\left\| \frac{\partial h_{t,t}}{\partial \theta_t}\bigg|_{x_t,\theta_t} - \frac{\partial h_{t,t}}{\partial \theta_t}\bigg|_{x'_t,\theta'_t} \right\|$$

$$
\begin{aligned}
&= \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t} \cdot \left.\frac{\partial \pi_t}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t'} \cdot \left.\frac{\partial \pi_t}{\partial \theta_t}\right|_{x_t',\theta_t'} \right\| && \text{(5.38a)}\\
&\le \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t} \cdot \left(\left.\frac{\partial \pi_t}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \pi_t}{\partial \theta_t}\right|_{x_t',\theta_t'} \right) \right\| + \left\| \left(\left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t'} \right) \cdot \left.\frac{\partial \pi_t}{\partial \theta_t}\right|_{x_t',\theta_t'} \right\| && \text{(5.38b)}\\
&\le L_h \left(\ell_{\pi,\theta}\left\|\theta_t - \theta_t'\right\| + \ell_{\pi,x}\left\|x_t - x_t'\right\|\right) + \left(\ell_{f,x}\left\|x_t - x_t'\right\| + \ell_{f,u}\left\|u_t - u_t'\right\|\right) \cdot L_{\pi,\theta} && \text{(5.38c)}\\
&\le \left(L_h\ell_{\pi,\theta} + (\ell_{f,x} + \ell_{f,u}L_{\pi,x})L_{\pi,\theta}\right)\left\|x_t - x_t'\right\| + \left(L_h\ell_{\pi,\theta} + \ell_{f,u}L_{\pi,\theta}^2\right)\left\|\theta_t - \theta_t'\right\|, && \text{(5.38d)}
\end{aligned}
$$

where we use the chain rule decomposition in (5.38a); we use the triangle inequality in (5.38b); we use Assumption 5.2.1 in both (5.38c) and (5.38d).

When $\tau < t$, by (5.36), we have that

$$
\begin{aligned}
&\left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t}} - \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau',\theta_\tau',\theta_{\tau+1:t}} \right\|\\
&\le \left\| \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t} \cdot \left(\left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau',\theta_\tau',\theta_{\tau+1:t-1}} \right) \right\|\\
&\quad + \left\| \left(\left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t} - \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t',u_t'} \right) \cdot \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau',\theta_\tau',\theta_{\tau+1:t-1}} \right\|\\
&\quad + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t} \cdot \left.\frac{\partial \pi_t}{\partial x_t}\right|_{x_t,\theta_t} \cdot \left(\left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,\theta_{\tau+1:t-1}} - \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau',\theta_\tau',\theta_{\tau+1:t-1}} \right) \right\|\\
&\quad + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t} \cdot \left(\left.\frac{\partial \pi_t}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \pi_t}{\partial x_t}\right|_{x_t',\theta_t} \right) \cdot \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau',\theta_\tau',\theta_{\tau+1:t-1}} \right\|\\
&\quad + \left\| \left(\left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t',u_t'} \right) \cdot \left.\frac{\partial \pi_t}{\partial x_t}\right|_{x_t',\theta_t} \cdot \left.\frac{\partial g_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau',\theta_\tau',\theta_{\tau+1:t-1}} \right\| && \text{(5.39a)}\\
&\le L_h\rho^{t-\tau}\left(C_{\ell,g,(\theta,x)}\left\|x_\tau - x_\tau'\right\| + C_{\ell,g,(\theta,\theta)}\left\|\theta_\tau - \theta_\tau'\right\|\right)\\
&\quad + \left(\ell_{f,x}\left\|x_t - x_t'\right\| + \ell_{f,u}\left\|u_t - u_t'\right\|\right) \cdot C_{L,g,\theta}\rho^{t-\tau}\\
&\quad + L_hL_{\pi,x}\rho^{t-\tau}\left(C_{\ell,g,(\theta,x)}\left\|x_\tau - x_\tau'\right\| + C_{\ell,g,(\theta,\theta)}\left\|\theta_\tau - \theta_\tau'\right\|\right)\\
&\quad + L_h\ell_{\pi,x}\left\|x_t - x_t'\right\| \cdot C_{L,g,\theta}\rho^{t-\tau} + \left(\ell_{f,x}\left\|x_t - x_t'\right\| + \ell_{f,u}\left\|u_t - u_t'\right\|\right) \cdot L_{\pi,x}C_{L,g,\theta}\rho^{t-\tau}, && \text{(5.39b)}
\end{aligned}
$$

where we use (5.36) and the triangle inequality in (5.39a); we use Assumption 5.2.1 in (5.39b). Note that by the first two inequalities in Lemma 5.B.3, we have

$$
\left\|x_t - x_t'\right\| \le \rho^{t-\tau}\left(C_{L,g,x}\left\|x_\tau - x_\tau'\right\| + C_{L,g,\theta}\left\|\theta_\tau - \theta_\tau'\right\|\right),
$$

$$\left\|u_t - u_t'\right\| \leq L_{\pi,x}\rho^{t-\tau}\left(C_{L,g,x}\left\|x_\tau - x_\tau'\right\| + C_{L,g,\theta}\left\|\theta_\tau - \theta_\tau'\right\|\right).$$

Substituting these into (5.39) finishes the proof of the fourth inequality.

Proof of Theorem 5.B.5

To simplify the notation, we use the shorthand $\hat{x}_\tau(\theta) \coloneqq g_{\tau|0}(x_0, \theta_{\times\tau})$ and $\hat{u}_\tau(\theta) = \pi_\tau(\hat{x}_\tau(\theta), \theta)$ for any time τ and policy parameter θ.

We first derive an upper bound of G_t in order to bound the difference between θ_t and θ_{t+1}. Recall that

$$G_t \coloneqq \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,\theta_{0:t}} = \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\right|_{x_{t-\tau},\theta_{t-\tau:t}}. \tag{5.40}$$

Now we use induction to show that for all time step $t \in \mathcal{T}$,

$$\|G_t\| \leq \frac{C_{L,h,\theta}}{1-\rho}, x_t \in B_n(0, R_S + C\|x_0\|), u_t \in \mathcal{U}, \text{ and } \|\theta_{t+1} - \theta_t\| \leq \varepsilon. \tag{5.41}$$

Note that $\|G_0\| \leq C_{L,h,\theta} \leq \frac{C_{L,h,\theta}}{1-\rho}$ by Corollary 5.B.4. We also have $x_0 \in B_n(0, R_S + C\|x_0\|)$ and $u_0 \in \mathcal{U}$.

Suppose $\|G_{t-1}\| \leq \frac{C_{L,h,\theta}}{1-\rho}$ for some $t \geq 1$. Then, since $\eta \leq \frac{(1-\rho)\varepsilon}{C_{L,h,\theta}}$ and the projection onto Θ is a contraction (see Theorem 1.2.1 in Schneider, 2014), we see that

$$\|\theta_t - \theta_{t-1}\| \leq \|\eta G_{t-1}\| \leq \varepsilon.$$

Suppose $\|\theta_\tau - \theta_{\tau-1}\| \leq \varepsilon$ holds for all $\tau \leq t$, i.e., $\theta_{0:t} \in S_\varepsilon(0:t)$. By Lemma 5.B.2, we see that

$$x_t \in B_n(0, R_S + C\|x_0\|), \text{ and } u_t \in \mathcal{U}.$$

Taking norm on both sides of (5.40), we see that

$$\begin{aligned}
\|G_t\| &= \left\|\sum_{\tau=0}^{t} \left.\frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\right|_{x_{t-\tau},\theta_{t-\tau:t}}\right\| \\
&\leq \sum_{\tau=0}^{t} \left\|\left.\frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\right|_{x_{t-\tau},\theta_{t-\tau:t}}\right\| && (5.42a) \\
&\leq \sum_{\tau=0}^{t} C_{L,h,\theta}\rho^\tau && (5.42b) \\
&\leq \frac{C_{L,h,\theta}}{1-\rho},
\end{aligned}$$

where we use the triangle inequality in (5.42a) and Corollary 5.B.4 in (5.42b). Note that we can apply Corollary 5.B.4 because $x_t \in B_n(0, R_S + C\|x_0\|)$. Therefore, we have shown (5.41) by induction. One can use the same technique as (5.42) to show $\|\nabla F_t(\theta_t)\| \leq \frac{C_{L,h,\theta}}{1-\rho}$.

Since the projection onto the set Θ is a contraction, we obtain that for any $t > \tau$,

$$\|\theta_t - \theta_\tau\| \leq \frac{C_{L,h,\theta}\eta(t-\tau)}{1-\rho}. \tag{5.43}$$

Now we bound the distance between x_τ and $\hat{x}_\tau(\theta_t)$ for $\tau \leq t$. We see that

$$\begin{aligned}
\|x_\tau - \hat{x}_\tau(\theta_t)\| &= \left\|g_{\tau|0}(x_0, \theta_{0:\tau-1}) - g_{\tau|0}(x_0, (\theta_t)_{\times\tau})\right\| \\
&\leq \sum_{\tau'=0}^{\tau-1} \left\|g_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')})\right\| && \text{(5.44a)} \\
&\leq \sum_{\tau'=0}^{\tau-1} \left\|g_{\tau|\tau'}(x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g_{\tau|\tau'}(x_{\tau'}, (\theta_t)_{\times(\tau-\tau')})\right\| && \text{(5.44b)} \\
&\leq \sum_{\tau'=0}^{\tau-1} C_{L,g,\theta}\rho^{\tau-\tau'}\|\theta_t - \theta_{\tau'}\| && \text{(5.44c)} \\
&\leq \frac{C_{L,h,\theta}C_{L,g,\theta}\eta}{1-\rho}\sum_{\tau'=0}^{\tau-1}(t-\tau')\rho^{\tau-\tau'} && \text{(5.44d)} \\
&\leq \frac{C_{L,h,\theta}C_{L,g,\theta}\rho}{(1-\rho)^2}\left((t-\tau) + \frac{1}{1-\rho}\right)\cdot\eta,
\end{aligned}$$

where we use the triangle inequality in (5.44a); we use the definition of multi-step dynamics in (5.44b); we use Lemma 5.B.3 in (5.44c); we use (5.43) in (5.44d).

Similarly, since (5.40) guarantees that $x_t \in B_n(0, R_S + C\|x_0\|)$ and we also see that $\hat{x}_t(\theta_t) \in B_n(0, R_S + C\|x_0\|)$, we obtain that

$$\begin{aligned}
|h_t(x_t, u_t) - F_t(\theta_t)| &= |h_t(x_t, u_t) - h_t(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t))| \\
&\leq L_h\left(\|x_t - \hat{x}_t(\theta_t)\| + \|u_t - \hat{u}_t(\theta_t)\|\right) && \text{(5.45a)} \\
&= L_h\left(\|x_t - \hat{x}_t(\theta_t)\| + \|\pi_t(x_t, \theta_t) - \pi_t(\hat{x}_t(\theta_t), \theta_t)\|\right) \\
&\leq L_h(1 + L_{\pi,x})\|x_t - \hat{x}_t(\theta_t)\| && \text{(5.45b)} \\
&\leq \frac{C_{L,h,\theta}C_{L,g,\theta}L_h(1 + L_{\pi,x})\rho}{(1-\rho)^3}\cdot\eta, && \text{(5.45c)}
\end{aligned}$$

where we use Assumption 5.2.1 in (5.45a) and (5.45b); we use (5.44) in (5.45c).

To show the last inequality in Theorem 5.B.5, note that we have

$$
\begin{aligned}
\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| \le{}& \sum_{\tau=0}^{t-1} \left\| g_{t|0}\left(x_0, \tilde{\theta}_{0:\tau-1}, (\tilde{\theta}_t)_{\times(t-\tau)}\right) - g_{t|0}\left(x_0, \tilde{\theta}_{0:\tau}, (\tilde{\theta}_t)_{\times(t-\tau-1)}\right)\right\| && (5.46a)\\
\le{}& C L_{\pi,\theta} L_{g,x} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \left\|\tilde{\theta}_t - \tilde{\theta}_\tau\right\| && (5.46b)\\
\le{}& C L_{\pi,\theta} L_{g,x} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \sum_{\tau'=\tau}^{t-1} \left\|\tilde{\theta}_{\tau'+1} - \tilde{\theta}_{\tau'}\right\| && (5.46c)\\
\le{}& \frac{C L_{\pi,\theta} L_{g,x}}{1-\rho} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \left\|\tilde{\theta}_{\tau+1} - \tilde{\theta}_\tau\right\|, && (5.46d)
\end{aligned}
$$

where we use the triangle inequality in (5.46a) and (5.46c); we use the assumption that $\|\tilde{x}_t\| \le \min\{R_C, R_x\}$ and the time-invariant contractive perturbation property in (5.46b); we rearrange the terms and use $\sum_{\tau=0}^{\infty} \rho^\tau \le \frac{1}{1-\rho}$ in (5.46d).

Therefore, since $\tilde{x}_t, \hat{x}_t(\tilde{\theta}_t) \in \mathcal{X}$ and $\tilde{u}_t, \hat{u}_t(\tilde{\theta}_t) \in \mathcal{U}$, we see that

$$
\begin{aligned}
\left|h_t(\tilde{x}_t, \tilde{u}_t) - F_t(\tilde{\theta}_t)\right| &= \left|h_t(\tilde{x}_t, \tilde{u}_t) - h_t(\hat{x}_t(\tilde{\theta}_t), \hat{u}_t(\tilde{\theta}_t))\right| \\
&\le L_h \left(\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| + \left\|\tilde{u}_t - \hat{u}_t(\tilde{\theta}_t)\right\|\right) && (5.47a)\\
&= L_h \left(\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| + \left\|\pi_t(\tilde{x}_t, \tilde{\theta}_t) - \pi_t(\hat{x}_t(\tilde{\theta}_t), \tilde{\theta}_t)\right\|\right)\\
&\le L_h(1 + L_{\pi,x})\left\|\tilde{x}_t - \hat{x}_t(\tilde{\theta}_t)\right\| && (5.47b)\\
&\le \frac{C L_{\pi,\theta} L_{g,x} L_h (1 + L_{\pi,x})}{1-\rho} \sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \left\|\tilde{\theta}_{\tau+1} - \tilde{\theta}_\tau\right\|, && (5.47c)
\end{aligned}
$$

where we use Assumption 5.2.1 in (5.47a) and (5.47b); we use (5.46) in (5.47c).

Proof of Theorem 5.B.6

Theorem 5.B.9 (Gradient Bias)**.** *Suppose Assumptions 5.2.1 and 5.2.2 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of M-GAPS (Algorithm 6) with learning rate $\eta \le \frac{(1-\rho)\varepsilon}{C_{L,h,\theta}}$. Then, the following holds for all $\tau \le t$:*

$$
\left\| \left.\frac{\partial h_{t|0}}{\partial \theta_\tau}\right|_{x_0, \theta_{0:t}} - \left.\frac{\partial h_{t|0}}{\partial \theta_\tau}\right|_{x_0, (\theta_t)_{\times(t+1)}} \right\| \le \left(\hat{C}_0 + \hat{C}_1 (t-\tau) + \hat{C}_2 (t-\tau)^2\right) \rho^{t-\tau} \cdot \eta,
$$

for

$$
\hat{C}_0 = \frac{\rho C_{L,h,\theta} C_{L,g,\theta} C_{\ell,h,(\theta,x)}}{(1-\rho)^3}, \hat{C}_1 = \frac{(1-\rho) C_{L,h,\theta} C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta} C_{L,g,\theta} C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2},
$$

$$\hat{C}_2 = \frac{C_{\ell,h,(x,\theta)} C_{L,g,\theta} C_{L,h,\theta}}{1-\rho}.$$

Next,

$$\|G_t - \nabla F_t(\theta_t)\| \le \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta.$$

Proof of Theorem 5.B.9. To simplify the notation, we adopt the shorthand notations $\hat{x}_\tau(\theta) \coloneqq g_{\tau|0}(x_0, \theta_{\times\tau})$ and $\hat{u}_\tau(\theta) \coloneqq \pi_\tau(\hat{x}_\tau(\theta), \theta)$ throughout the proof.

As we discussed below Theorem 5.B.6 in the proof outline, we use the triangle inequality to do the decomposition

$$\begin{aligned}
&\left\| \frac{\partial h_{t|0}}{\partial \theta_\tau}\Big|_{x_0,\theta_{0:t}} - \frac{\partial h_{t|0}}{\partial \theta_\tau}\Big|_{x_0,(\theta_t)_{\times(t+1)}} \right\| \\
={}& \left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:t}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
\le{}& \left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,(\theta_t)_{\times(t-\tau)}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
&+ \sum_{\tau'=\tau+1}^{t-1} \left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:\tau'},(\theta_t)_{\times(t-\tau')}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:\tau'-1},(\theta_t)_{\times(t-\tau'+1)}} \right\|.
\end{aligned} \tag{5.48}$$

Note that we can apply Corollary 5.B.4 to bound each term in (5.48). For the first term in (5.48), since $x_\tau, \hat{x}_\tau(\theta_t), x_{\tau+1} \in B_n(0, R_S + C\|x_0\|)$, we see that

$$\begin{aligned}
&\left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_\tau,(\theta_t)_{\times(t-\tau)}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
\le{}& \rho^{t-\tau}\left(C_{\ell,h,(\theta,x)}\|x_\tau - \hat{x}_\tau(\theta_t)\| + C_{\ell,h,(\theta,\theta)}\|\theta_t - \theta_\tau\|\right) && \text{(5.49a)} \\
\le{}& \frac{(1-\rho)C_{L,h,\theta}C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta}C_{L,g,\theta}C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2} \cdot (t-\tau)\rho^{t-\tau}\cdot\eta \\
&+ \frac{\rho C_{L,h,\theta}C_{L,g,\theta}C_{\ell,h,(\theta,x)}}{(1-\rho)^3}\cdot\rho^{t-\tau}\cdot\eta, && \text{(5.49b)}
\end{aligned}$$

where we use Corollary 5.B.4 in (5.49a) and Theorem 5.B.5 in (5.49b).

For any $\tau' \in [\tau+1 : t-1]$, since $x_{\tau'}, x_{\tau'+1} \in B_n(0, R_S + C\|x_0\|)$, we see that

$$\begin{aligned}
&\left\| \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:\tau'},(\theta_t)_{\times(t-\tau')}} - \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:\tau'-1},(\theta_t)_{\times(t-\tau'+1)}} \right\| \\
={}& \left\| \left(\frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\Big|_{x_{\tau'},\theta_{\tau'},(\theta_t)_{\times(t-\tau')}} - \frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\Big|_{x_{\tau'},(\theta_t)_{\times(t-\tau'+1)}} \right) \frac{\partial g_{\tau'|\tau}}{\partial \theta_\tau}\Big|_{x_\tau,\theta_{\tau:\tau'-1}} \right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \left.\frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'},\theta_{\tau'},(\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'},(\theta_t)_{\times(t-\tau'+1)}} \right\| \cdot \left\| \left.\frac{\partial g_{\tau'|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'-1}} \right\| \\
&\leq C_{\ell,h,(x,\theta)} \rho^{t-\tau'} \|\theta_t - \theta_{\tau'}\| \cdot C_{L,g,\theta} \rho^{\tau'-\tau} && (5.50a) \\
&\leq \frac{C_{\ell,h,(x,\theta)} C_{L,g,\theta} C_{L,h,\theta}}{1-\rho} \cdot (t-\tau)\rho^{t-\tau} \cdot \eta, && (5.50b)
\end{aligned}
$$

where we use Lemma 5.B.3 and Corollary 5.B.4 in (5.50a); we use Theorem 5.B.5 in (5.50b). Substituting (5.49) and (5.50) into (5.48) finishes the proof of the first inequality.

For the second inequality, recall that G_t and $\nabla \ell_t(\theta_t)$ are given by

$$
G_t \coloneqq \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,\theta_{0:t}}, \nabla \ell_t(\theta_t) = \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,(\theta_t)_{\times(t+1)}}.
$$

Therefore, we see that

$$
\begin{aligned}
\|G_t - \nabla F_t(\theta_t)\| &= \left\| \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,\theta_{0:t}} - \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,(\theta_t)_{\times(t+1)}} \right\| \\
&\leq \sum_{\tau=0}^{t} \left\| \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,\theta_{0:t}} - \left.\frac{\partial h_{t|0}}{\partial \theta_{t-\tau}}\right|_{x_0,(\theta_t)_{\times(t+1)}} \right\| && (5.51a) \\
&\leq \sum_{\tau=0}^{t} \left(\hat{C}_0 + \hat{C}_1\tau + \hat{C}_2\tau^2\right)\rho^\tau \eta && (5.51b) \\
&\leq \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta,
\end{aligned}
$$

where we use the triangle inequality in (5.51a); we use the first inequality in Theorem 5.B.6 that we have shown and Corollary 5.B.4 in (5.51b). □

Regret of M-GAPS under Nonconvex Surrogate Costs

To derive the local regret bound for GAPS in online policy selection, we first bound the local regret for OGD (with biased gradients) in online nonconvex optimization.

Theorem 5.B.10. *Suppose $\Theta = \mathbb{R}^d$. Consider the biased OGD update rule $\theta_{t+1} = \theta_t - \eta G_t$, where G_t satisfies $\|G_t - \nabla F_t(\theta_t)\| \leq \beta$. Suppose at every time t, F_t is L_F-Lipschitz and ℓ_F-smooth. If the learning rate $\eta < \frac{1}{\ell_F}$, we have that*

$$
\sum_{t=0}^{T-1} \|\nabla F_t(\theta_t)\|^2 \leq \frac{2}{\eta}\left(F_0(\theta_0) + \sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1})\right) + \left(2(1-\ell_F\eta)L_F\beta + \ell_F\eta \cdot \beta^2\right)T,
$$

where $\mathsf{dist}_s(F, F') \coloneqq \sup_{\theta\in\Theta} |F(\theta) - F'(\theta)|$.

We provide the proof of Theorem 5.B.10 later in this section. Our proof is inspired by the analysis for (stochastic) gradient descent in offline nonconvex optimization (see, e.g., Ghadimi and Lan, 2013) with the additional step to handle the time-varying function sequence $F_{0:T-1}$ via the measure of variation $\mathsf{dist}_s(F_t, F_{t-1})$.

Our approximation error bound (Theorem 5.3.1) guarantees that the bias $\beta = O(1/\sqrt{T})$ if we set the learning rate $\eta = O(1/\sqrt{T})$. Therefore, the remaining task is to bound the measure of variation in online nonconvex optimization $\sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1})$ by the variation intensity in online policy selection V (Definition 5.3.2). To derive this bound, we need to show a convergence result on applying a fixed policy parameter in a time-invariant system. We begin with a definition that characterize this (imaginary) dynamical process.

Definition 5.B.1. *For fixed dynamics function g, policy function π, and policy parameter θ, we define $\tilde{x}_\tau^{(g,\pi)}(\theta)$ recursively by the equation*

$$\tilde{x}_{\tau+1}^{(g,\pi)}(\theta) = g\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \pi\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \theta\right)\right), \forall \tau \geq 0, \textit{ where } \tilde{x}_0^{(g,\pi)}(\theta) = x_0.$$

Compared with $\hat{x}_t(\theta)$ we defined before Definition 5.1.2, the state $\tilde{x}_{\tau+1}^{(g,\pi)}(\theta)$ is produced by a time-invariant dynamical system induced by g and π, while $\hat{x}_t(\theta)$ is produced by the actual time-varying dynamics induced by $g_{0:t-1}$ and $\pi_{0:t-1}$.

We show the time-invariant evolution $\tilde{x}_\tau^{(g,\pi)}$ has a unique limitation point as τ tends to infinity. This limit is also a fixed point, and the states will converge to the limit exponentially fast with respect to τ. We state this result formally in Lemma 5.B.11.

Lemma 5.B.11. *Suppose Assumptions 5.2.1 and 5.2.2 hold, and $(g, \pi) \in \mathcal{G}$. The limit $\lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$ exists. Let $\tilde{x}_\infty^{(g,\pi)}(\theta) \coloneqq \lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$. Further, we also have that*

$$\left\|\tilde{x}_\tau^{(g,\pi)}(\theta) - \tilde{x}_\infty^{(g,\pi)}(\theta)\right\| \leq C\rho^\tau \left\|x_0 - \tilde{x}_\infty^{(g,\pi)}(\theta)\right\| \leq C\rho^\tau \cdot \mathsf{diam}(\mathcal{X}),$$

where $\mathsf{diam}(\mathcal{X}) = 2C(R_S + C\|x_0\|) + 2R_S$.

We provide the proof of Lemma 5.B.11 later in this section. With the fixed point and convergence result in Lemma 5.B.11, we bound the measure of variation based on $F_{0:T-1}$ by the variation intensity V based on $g_{0:T-1}, \pi_{0:T-1}$, and $h_{0:T-1}$ in Lemma 5.B.12.

Lemma 5.B.12. *Suppose Assumptions 5.2.1 and 5.2.2 hold. Then, we have*

$$\sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1}) \le \frac{2CL_h(1+L_{\pi,x})(1+L_{g,u})}{(1-\rho)^2\rho} \cdot V + \frac{2CL_h(1+L_{\pi,x})}{1-\rho} \cdot \mathsf{diam}(\mathcal{X}),$$

where $\mathsf{diam}(\mathcal{X}) = 2C(R_S + C\|x_0\|) + 2R_S$.

With these auxiliary results, we restate Theorem 5.3.4 with complete expressions and present the proof.

Theorem 5.B.13. *Under the same assumptions as Theorem 5.3.1, if we additionally assume that* $\Theta = \mathbb{R}^d$ *for some integer* d*, then GAPS satisfies local regret*

$$\begin{aligned} R^L(T) \le{} & \frac{2}{\eta}\left(c_0 + \frac{2CL_F L_h(1+L_{\pi,x})}{1-\rho}\left(\frac{(1+L_{g,u})V}{(1-\rho)\rho} + 2C(R_S + C\|x_0\|) + 2R_S\right)\right) \\ & + 2(1-\ell_F\eta)L_F\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T \\ & + 2\ell_F\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)^2 \eta^3 T, \end{aligned} \tag{5.52}$$

where $\hat{C}_0, \hat{C}_1, \hat{C}_2$ *are defined in Theorem 5.B.9,* $c_0 = f_0(x_0, \pi_0(x_0, \theta_0))$*, and*

$$L_F = \frac{C_{L,h,\theta}}{1-\rho},\ \ell_F = \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,h,(\theta,x)} + \rho C_{\ell,h,(x,\theta)}C_{L,g,\theta}}{(1-\rho)^2}.$$

Proof of Theorem 5.B.13. By Theorem 5.B.10 and Theorem 5.3.1, we know the parameter sequence of GAPS satisfies that

$$\sum_{t=0}^{T-1} \|\nabla F_t(\theta_t)\|^2 \le \frac{2}{\eta}\left(F_0(\theta_0) + \sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1})\right) + \left(2(1-\ell_F\eta)L_F\beta + \ell_F\eta \cdot \beta^2\right)T, \tag{5.53}$$

where $\beta = \left(\hat{C}_0(1-\rho)^{-1} + (\hat{C}_1 + \hat{C}_2)(1-\rho)^{-2} + \hat{C}_2(1-\rho)^{-3}\right)\eta + C_{L,h,\theta}(1-\rho)^{-1} \cdot \rho^B$. Here, $\hat{C}_0, \hat{C}_1, \hat{C}_2$ are defined in Theorem 5.B.9.

Note that $L_F = \frac{C_{L,h,\theta}}{1-\rho}$ by Theorem 5.B.5. Now we show that we can set

$$\ell_F = \frac{CL_{g,u}L_{\pi,\theta}C_{\ell,h,(\theta,x)} + \rho C_{\ell,h,(x,\theta)}C_{L,g,\theta}}{(1-\rho)^2}.$$

To see this, by Lemma 5.B.3, we obtain that the following inequality holds for every time step t,

$$\|\nabla F_t(\theta) - \nabla F_t(\theta')\| = \left\| \sum_{\tau=0}^{t} \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\bigg|_{\hat{x}_\tau(\theta), \theta_{\times(t-\tau+1)}} - \sum_{\tau=0}^{t} \frac{\partial h_{t|\tau}}{\partial \theta_\tau}\bigg|_{\hat{x}_\tau(\theta'), \theta'_{\times(t-\tau+1)}} \right\| \tag{5.54a}$$

$$\leq \sum_{\tau=0}^{t} \left\| \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta), \theta_{\times(t-\tau+1)}} - \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(t-\tau+1)}} \right\|, \tag{5.54b}$$

where we use the definition of surrogate cost functions in (5.54a); we use the triangle inequality in (5.54b).

Note that for each term in (5.54), we can decompose it as

$$\begin{aligned}
&\left\| \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta), \theta_{\times(t-\tau+1)}} - \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(t-\tau+1)}} \right\| \\
\leq &\left\| \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta), \theta_{\times(t-\tau+1)}} - \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta_{\times(t-\tau+1)}} \right\| \\
&+ \sum_{j=\tau}^{t} \left\| \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(j-\tau)}, \theta_{\times(t-j+1)}} - \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(j-\tau+1)}, \theta_{\times(t-j)}} \right\|.
\end{aligned} \tag{5.55}$$

Note that for any time step τ, by the triangle inequality, we have

$$\begin{aligned}
\|\hat{x}_\tau(\theta) - \hat{x}_\tau(\theta')\| &\leq \sum_{j=0}^{\tau-1} \left\| g_{\tau|0}(x_0, \theta_{\times j}, \theta'_{\times(\tau-j)}) - g_{\tau|0}(x_0, \theta_{\times(j+1)}, \theta'_{\times(\tau-j-1)}) \right\| \\
&\leq \sum_{j=0}^{\tau-1} C\rho^{\tau-j-1} \cdot L_{g,u} L_{\pi,\theta} \|\theta - \theta'\| \leq \frac{C L_{g,u} L_{\pi,\theta}}{1-\rho} \cdot \|\theta - \theta'\|,
\end{aligned}$$

where we apply the time-invariant contractive perturbation in the last inequality. Therefore, by Corollary 5.B.4, we obtain that

$$\begin{aligned}
&\left\| \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta), \theta_{\times(t-\tau+1)}} - \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta_{\times(t-\tau+1)}} \right\| \\
\leq{}& C_{\ell,h,(\theta,x)} \rho^{t-\tau} \|\hat{x}_\tau(\theta) - \hat{x}_\tau(\theta')\| \leq \frac{C L_{g,u} L_{\pi,\theta} C_{\ell,h,(\theta,x)}}{1-\rho} \cdot \rho^{t-\tau} \cdot \|\theta - \theta'\|.
\end{aligned} \tag{5.56}$$

We also see that

$$\begin{aligned}
&\left\| \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(j-\tau)}, \theta_{\times(t-j+1)}} - \left. \frac{\partial h_{t|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(j-\tau+1)}, \theta_{\times(t-j)}} \right\| \\
={}& \left\| \left(\left. \frac{\partial h_{t|j}}{\partial x_j} \right|_{\hat{x}_j(\theta'), \theta_{\times(t-j+1)}} - \left. \frac{\partial h_{t|j}}{\partial \theta_j} \right|_{\hat{x}_j(\theta'), \theta', \theta_{\times(t-j)}} \right) \left. \frac{\partial g_{j|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(j-\tau)}} \right\| \\
\leq{}& \left\| \left. \frac{\partial h_{t|j}}{\partial x_j} \right|_{\hat{x}_j(\theta'), \theta_{\times(t-j+1)}} - \left. \frac{\partial h_{t|j}}{\partial \theta_j} \right|_{\hat{x}_j(\theta'), \theta', \theta_{\times(t-j)}} \right\| \cdot \left\| \left. \frac{\partial g_{j|\tau}}{\partial \theta_\tau} \right|_{\hat{x}_\tau(\theta'), \theta'_{\times(j-\tau)}} \right\| \\
\leq{}& C_{\ell,h,(x,\theta)} \rho^{t-j} \|\theta - \theta'\| \cdot C_{L,g,\theta} \rho^{j-\tau} = C_{\ell,h,(x,\theta)} C_{L,g,\theta} \rho^{t-\tau} \|\theta - \theta'\|.
\end{aligned} \tag{5.57}$$

Substituting (5.56) and (5.57) into (5.55) gives that

$$\begin{aligned}&\left\| \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta), \theta_{\times(t-\tau+1)}} - \left.\frac{\partial h_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta'), \theta'_{\times(t-\tau+1)}} \right\| \\ \le\ &\left(\frac{C L_{g,u} L_{\pi,\theta} C_{\ell,h,(\theta,x)}}{1-\rho} + (t-\tau) C_{\ell,h,(x,\theta)} C_{L,g,\theta}\right) \cdot \rho^{t-\tau} \|\theta - \theta'\|.\end{aligned}$$

Substituting this inequality into (5.54) gives that

$$\begin{aligned}&\|\nabla F_t(\theta) - \nabla F_t(\theta')\| \\ \le\ &\sum_{\tau=0}^{t} \left(\frac{C L_{g,u} L_{\pi,\theta} C_{\ell,h,(\theta,x)}}{1-\rho} + (t-\tau) C_{\ell,h,(x,\theta)} C_{L,g,\theta}\right) \cdot \rho^{t-\tau} \|\theta - \theta'\| \\ \le\ &\frac{C L_{g,u} L_{\pi,\theta} C_{\ell,h,(\theta,x)} + \rho C_{\ell,h,(x,\theta)} C_{L,g,\theta}}{(1-\rho)^2} \cdot \|\theta - \theta'\|.\end{aligned}$$

Therefore, we can set $\ell_F = \frac{C L_{g,u} L_{\pi,\theta} C_{\ell,h,(\theta,x)} + \rho C_{\ell,h,(x,\theta)} C_{L,g,\theta}}{(1-\rho)^2}$.

Recall that the notation dist_s is defined in Theorem 5.B.10. By Lemma 5.B.12, we know that

$$\begin{aligned}&\sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1}) \\ \le\ &\frac{2 C L_h (1 + L_{\pi,x})(1 + L_{g,u})}{(1-\rho)^2 \rho} \cdot V + \frac{4 C L_h (1 + L_{\pi,x}) \left(C(R_S + C\|x_0\|) + R_S\right)}{1-\rho}.\end{aligned}$$

Substituting this inequality and the expressions of L_F, ℓ_F into (5.53) finishes the proof. □

Proof of Theorem 5.B.10

By the smoothness of $F_t(\cdot)$, we see that

$$\begin{aligned} F_t(\theta_{t+1}) &\le F_t(\theta_t) + \langle \nabla F_t(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\ell_F}{2} \|\theta_{t+1} - \theta_t\|^2 \\ &= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), G_t \rangle + \frac{\ell_F \eta^2}{2} \|G_t\|^2 \\ &= F_t(\theta_t) - \eta \langle \nabla F_t(\theta_t), \nabla F_t(\theta_t) \rangle + \frac{\ell_F \eta^2}{2} \|\nabla F_t(\theta_t)\|^2 \\ &\quad - \eta \langle \nabla F_t(\theta_t), G_t - \nabla F_t(\theta_t) \rangle + \ell_F \eta^2 \langle \nabla F_t(\theta_t), G_t - \nabla F(\theta_t) \rangle \\ &\quad + \frac{\ell_F \eta^2}{2} \|\nabla F_t(\theta_t) - G_t\|^2 \\ &\le F_t(\theta_t) - \eta \left(1 - \frac{\ell_F \eta}{2}\right) \|\nabla F_t(\theta_t)\|^2 + \eta (1 - \ell_F \eta) L_F \beta + \frac{\ell_F \eta^2}{2} \cdot \beta^2. \end{aligned} \tag{5.58}$$

Summing (5.58) over $t = 0, 1, \ldots, T-1$ and rearranging the terms gives the following inequalities:

$$\begin{aligned}
&\eta\left(1 - \frac{\ell_F \eta}{2}\right)\sum_{t=0}^{T-1} \|\nabla F_t(\theta_t)\|^2 \\
&\leq \sum_{t=0}^{T-1} (F_t(\theta_t) - F_t(\theta_{t+1})) + \left(\eta(1-\ell_F\eta)L_F\beta + \frac{\ell_F\eta^2}{2}\cdot\beta^2\right)T \\
&\leq F_0(\theta_0) + \sum_{t=1}^{T-1} (F_t(\theta_t) - F_{t-1}(\theta_t)) + \left(\eta(1-\ell_F\eta)L_F\beta + \frac{\ell_F\eta^2}{2}\cdot\beta^2\right)T \\
&\leq F_0(\theta_0) + \sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1}) + \left(\eta(1-\ell_F\eta)L_F\beta + \frac{\ell_F\eta^2}{2}\cdot\beta^2\right)T. \qquad (5.59)
\end{aligned}$$

Proof of Lemma 5.B.11

We first show the limit $\lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$ exists. It suffices to show that $\{\tilde{x}_\tau^{(g,\pi)}(\theta)\}$ is a Cauchy sequence. Note that for $\tau' > \tau \geq 0$, we have

$$\begin{aligned}
\left\|\tilde{x}_{\tau'}^{(g,\pi)}(\theta) - \tilde{x}_\tau^{(g,\pi)}(\theta)\right\| &\leq \sum_{j=\tau}^{\tau'-1} \left\|\tilde{x}_{j+1}^{(g,\pi)}(\theta) - \tilde{x}_j^{(g,\pi)}(\theta)\right\| && (5.60a)\\
&\leq \sum_{j=\tau}^{\tau'-1} C\rho^j \left\|\tilde{x}_1^{(g,\pi)}(\theta) - x_0\right\| && (5.60b)\\
&\leq \frac{C\rho^\tau}{1-\rho}\cdot\left\|\tilde{x}_1^{(g,\pi)}(\theta) - x_0\right\|,
\end{aligned}$$

where we use the triangle inequality in (5.60a) and Assumption 5.2.2 in (5.60b). Therefore, we see the limit $\lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$ exists because $\{\tilde{x}_\tau^{(g,\pi)}(\theta)\}$ is a Cauchy sequence and we denote $\tilde{x}_\infty^{(g,\pi)}(\theta) \coloneqq \lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta)$.

Now, we show that $\tilde{x}_\infty^{(g,\pi)}(\theta)$ is a fixed point of the time-invariant closed-loop dynamics induced by (g, π, θ). To see this, note that

$$\begin{aligned}
g\left(\tilde{x}_\infty^{(g,\pi)}(\theta), \pi(\tilde{x}_\infty^{(g,\pi)}(\theta), \theta)\right) &= g\left(\lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta), \pi(\lim_{\tau\to\infty} \tilde{x}_\tau^{(g,\pi)}(\theta), \theta)\right)\\
&= \lim_{\tau\to\infty} g\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \pi(\tilde{x}_\tau^{(g,\pi)}(\theta), \theta)\right)\\
&= \lim_{\tau\to\infty} \tilde{x}_{\tau+1}^{(g,\pi)}(\theta) = \tilde{x}_\infty^{(g,\pi)}(\theta),
\end{aligned}$$

where we can pull out $\lim_{\tau\to\infty}$ in the second equation because the right hand side is a continuous function of $\tilde{x}_\tau^{(g,\pi)}(\theta)$ at the point $\tilde{x}_\infty^{(g,\pi)}(\theta)$ by Assumption 5.2.1.

Therefore, applying the contractive perturbation property in Assumption 5.2.2 gives that

$$\left\|\tilde{x}_\tau^{(g,\pi)}(\theta) - \tilde{x}_\infty^{(g,\pi)}(\theta)\right\| \leq C\rho^\tau\left\|x_0 - \tilde{x}_\infty^{(g,\pi)}(\theta)\right\| \leq C\rho^\tau \mathsf{diam}(\mathcal{X}).$$

Proof of Lemma 5.B.12

To simplify the notation, we introduce the notations

$$\begin{aligned}\mathsf{dist}_d(g, g') &\coloneqq \sup_{x\in\mathcal{X}, u\in\mathcal{U}} \|g(x,u) - g'(x,u)\|,\\ \mathsf{dist}_p(\pi, \pi') &\coloneqq \sup_{x\in\mathcal{X}, \theta\in\Theta} \|\pi(x,\theta) - \pi'(x,\theta)\|,\\ \mathsf{dist}_c(h, h') &\coloneqq \sup_{x\in\mathcal{X}, u\in\mathcal{U}} |h(x,u) - h'(x,u)|.\end{aligned}$$

For any $\theta \in \Theta$, we see that

$$\begin{aligned}&|F_t(\theta) - F_{t-1}(\theta)|\\ &= |h_t\left(\hat{x}_t(\theta), \pi_t(\hat{x}_t(\theta), \theta)\right) - h_{t-1}\left(\hat{x}_{t-1}(\theta), \pi_{t-1}(\hat{x}_{t-1}(\theta), \theta)\right)| && (5.61a)\\ &\le |h_t\left(\hat{x}_t(\theta), \pi_t(\hat{x}_t(\theta), \theta)\right) - h_t\left(\hat{x}_{t-1}(\theta), \pi_{t-1}(\hat{x}_{t-1}(\theta), \theta)\right)| + \mathsf{dist}_c(h_t, h_{t-1}) && (5.61b)\\ &\le L_h\left(\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| + \|\pi_t(\hat{x}_t(\theta), \theta) - \pi_{t-1}(\hat{x}_{t-1}(\theta), \theta)\|\right) + \mathsf{dist}_c(h_t, h_{t-1}) && (5.61c)\\ &\le L_h\left(\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| + \|\pi_t(\hat{x}_t(\theta), \theta) - \pi_t(\hat{x}_{t-1}(\theta), \theta)\|\right)\\ &\quad + \left(L_h \mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(h_t, h_{t-1})\right) && (5.61d)\\ &\le L_h(1 + L_{\pi,x})\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| + \left(L_h \mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(h_t, h_{t-1})\right), && (5.61e)\end{aligned}$$

where we use the definition of the surrogate cost in (5.61a); we use the definition of dist_c in (5.61b); we use the assumption that h_t is Lipschitz in (5.61c); we use the definition of dist_π on (5.61d); we use the assumption that π_t is Lipschitz in (5.61e).

To bound $\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\|$, we first bound the difference between $\hat{x}_t(\theta)$ and $\tilde{x}_t^{(g,\pi)}(\theta)$ for arbitrary $(g, \pi) \in \mathcal{G}$. Note that $\hat{x}_t(\theta) = g_{t|0}\left(\tilde{x}_0^{(g,\pi)}, \theta_{\times t}\right)$, thus

$$\begin{aligned}\left\|\hat{x}_t(\theta) - \tilde{x}_t^{(g,\pi)}(\theta)\right\| &\le \sum_{\tau=0}^{t-1} \left\|g_{t|\tau+1}\left(\tilde{x}_{\tau+1}^{(g,\pi)}(\theta), \theta_{\times(t-\tau-1)}\right) - g_{t|\tau}\left(\tilde{x}_\tau^{(g,\pi)}(\theta), \theta_{\times(t-\tau)}\right)\right\| && (5.62a)\\ &\le C\sum_{\tau=0}^{t-1} \rho^{t-\tau-1} \left\|\tilde{x}_{\tau+1}^{(g,\pi)}(\theta) - g_{\tau+1|\tau}\left(\tilde{x}_{\tau+1}^{(g,\pi)}(\theta), \theta\right)\right\| && (5.62b)\\ &\le C\sum_{\tau=0}^{t-1} \rho^{t-\tau-1}\left(\mathsf{dist}_d(g_\tau, g) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi)\right), && (5.62c)\end{aligned}$$

where we use the triangle inequality in (5.62a); we use the contractive perturbation property in (5.62b); we use the definitions of dist_d and dist_p in (5.62c).

Therefore, we can decompose the distance between $\hat{x}_t(\theta)$ and $\hat{x}_{t-1}(\theta)$, and bound it by the total variation of dynamics and policies:

$$
\begin{aligned}
\|\hat{x}_t(\theta) - \hat{x}_{t-1}(\theta)\| \le{}& \left\|\hat{x}_t(\theta) - \tilde{x}_t^{(g_t,\pi_t)}(\theta)\right\| + \left\|\hat{x}_{t-1}(\theta) - \tilde{x}_{t-1}^{(g_t,\pi_t)}(\theta)\right\| \\
&+ \left\|\tilde{x}_t^{(g_t,\pi_t)}(\theta) - \tilde{x}_\infty^{(g_t,\pi_t)}(\theta)\right\| + \left\|\tilde{x}_{t-1}^{(g_t,\pi_t)}(\theta) - \tilde{x}_\infty^{(g_t,\pi_t)}(\theta)\right\| && \text{(5.63a)} \\
\le{}& C\sum_{\tau=0}^{t-1} \rho^{t-\tau-1}\left(\mathsf{dist}_d(g_\tau, g_t) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_t)\right) \\
&+ C\sum_{\tau=0}^{t-2} \rho^{t-\tau-2}\left(\mathsf{dist}_d(g_\tau, g_t) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_t)\right) \\
&+ C\rho^t\mathsf{diam}(\mathcal{X}) + C\rho^{t-1}\mathsf{diam}(\mathcal{X}) && \text{(5.63b)} \\
\le{}& \frac{2C}{1-\rho}\cdot\sum_{\tau=0}^{t-1} \rho^{t-\tau-2}\left(\mathsf{dist}_d(g_\tau, g_{\tau+1}) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_{\tau+1})\right) \\
&+ 2C\rho^{t-1}\mathsf{diam}(\mathcal{X}), && \text{(5.63c)}
\end{aligned}
$$

where we use the triangle inequality in (5.63a); in (5.63b), we use the bound we derived in (5.62); in (5.63c), we use the triangle inequality decomposition

$$
\mathsf{dist}_d(g_\tau, g_t) \le \sum_{j=\tau}^{t-1}\mathsf{dist}_d(g_{j+1}, g_j), \text{ and } \mathsf{dist}_p(\pi_\tau, \pi_t) \le \sum_{j=\tau}^{t-1}\mathsf{dist}_p(\pi_{j+1}, \pi_j).
$$

Substituting this into (5.61) gives that

$$
\begin{aligned}
&\mathsf{dist}_s(F_t, F_{t-1}) \\
\le{}& \frac{2CL_h(1+L_{\pi,x})}{1-\rho}\cdot\sum_{\tau=0}^{t-1}\rho^{t-\tau-2}\left(\mathsf{dist}_d(g_\tau, g_{\tau+1}) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_{\tau+1})\right) \\
&+ 2CL_h(1+L_{\pi,x})\rho^{t-1}\mathsf{diam}(\mathcal{X}) + (L_h\mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(h_t, h_{t-1})) \quad \text{(5.64)}
\end{aligned}
$$

because (5.63) holds for arbitrary $\theta \in \Theta$.

Summing (5.64) over $t = 1, \ldots, T-1$ and rearranging the terms give that

$$
\begin{aligned}
&\sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1}) \\
\le{}& \frac{2CL_h(1+L_{\pi,x})}{1-\rho}\cdot\sum_{t=1}^{T-1}\sum_{\tau=0}^{t-1}\rho^{t-\tau-2}\left(\mathsf{dist}_d(g_\tau, g_{\tau+1}) + L_{g,u}\mathsf{dist}_p(\pi_\tau, \pi_{\tau+1})\right) \\
&+ 2CL_h(1+L_{\pi,x})\sum_{t=1}^{T-1}\rho^{t-1}\mathsf{diam}(\mathcal{X}) + \sum_{t=1}^{T-1}(L_h\mathsf{dist}_p(\pi_t, \pi_{t-1}) + \mathsf{dist}_c(h_t, h_{t-1}))
\end{aligned}
$$

Table 5.1: Important notations in Section 5.4.

Notation	Meaning
$t_1 : t_2$	The integer sequence $\{t_1, \dots, t_2\}$;
$a_{t_1:t_2}$	A sequence of variables $\{a_t\}_{t=t_1,\dots,t_2}$;
$\Vert \cdot \Vert$	ℓ_2 (Euclidean) norm;
$\Vert \cdot \Vert_F$	Frobenius norm;
$\Vert \cdot \Vert_P$	Norm induced by matrix P;
$\mathbb{Z}_{\geq 0}$	The set of non-negative integers;
$\mathbb{R}_{\geq 0}$	The set of non-negative reals;
$\sigma(z_{1:t}, z'_{1:t})$	Product sigma-algebra generated by sequences $z_{1:t}$ and $z'_{1:t}$;
x_t	$x_t \in \mathbb{R}^n$ is the system state;
u_t	$u_t \in \mathbb{R}^m$ is the control input;
w_t	$w_t \in \mathcal{W} \subseteq \mathbb{R}^n$ is a disturbance term;
$f_t(x_t, a_t^*)$	f_t is a nonlinear residual term where the online agent makes (noisy) observations;
a_t^*	$a_t^* \in \mathcal{A} \subseteq \mathbb{R}^p$ is the unknown parameter in f_t
$f_t(\cdot, \hat{a}_t)$	An estimation of the true nonlinear residual function $f_t(\cdot, a_t^*)$;
$q_t(x_t, y_t, \theta_t, a_t^*)$	The joint dynamics of the system at time t;
$\Pi_\Theta(y)$	Euclidean projection of y to set Θ;

$$\leq \frac{2CL_h(1 + L_{\pi,x})(1 + L_{g,u})}{(1-\rho)^2\rho} \cdot P_{\mathcal{T}} + \frac{2CL_h(1 + L_{\pi,x})}{1-\rho} \cdot \mathsf{diam}(\mathcal{X}).$$

5.C Notations and Definitions of Meta-Framework for Unknown Dynamics

We provide a notation table (Table 5.1) that summarizes the important notations in Section 5.4.

A key concept that we explore in this paper is how to compare the actual trajectory of our meta-framework with an "ideal" trajectory that the agent could achieve with exact knowledge of the true model parameters $a^*_{0:T-1}$, we introduce the important notations of multi-step dynamics/cost that characterize how the system would evolve under a sequence of policy parameters $\theta_{0:T-1}$ when $a^*_{0:T-1}$ is known. The concepts of multi-step dynamics/cost are first introduced in Lin, Preiss, Anand, et al., 2023, which studies online policy selection with known dynamical systems. In this work, we replace all estimated $\hat{a}_t$ in the policy classes with true a_t^* to reproduce the same definition as Lin, Preiss, Anand, et al., 2023.

Definition 5.C.1 (Multi-Step Dynamics and Cost)**.** *The multi-step dynamics $g^*_{t|\tau}$ between two time steps $\tau \leq t$ specifies the state x_t as a function of the previous state x_τ and previous policy parameters $\theta_{\tau:t-1}$ under exact predictions $\{a_t^*\}$. It is defined recursively, with the base case $g^*_{\tau|\tau}(x_\tau) \coloneqq x_\tau$ and the recursive case*

$$g^*_{t+1|\tau}(x_\tau, \theta_{\tau:t}) = g_t\left(z_t, \pi_t\left(z_t, \theta_t, f_t(z_t, a_t^*)\right), f_t(z_t, a_t^*)\right) + w_t, \ \forall t \geq \tau,$$

in which $z_t \coloneqq g^*_{t|\tau}(x_\tau, \theta_{\tau:t-1})$.[6] *The multi-step cost* $h^*_{t|\tau}$ *specifies the cost* c_t *as function of* x_τ *and* $\theta_{\tau:t}$*. It is defined as*

$$h^*_{t|\tau}(x_\tau, \theta_{\tau:t}) \coloneqq h_t\left(z_t, \pi_t\left(z_t, \theta_t, f_t(z_t, a_t^*)\right), \theta_t\right).$$

It is worth emphasizing that, in our work, the concepts of multi-step dynamics/cost are only used for the theoretical analysis, because their definitions involve true model parameters that are unknown to the online agent. When doing online policy optimization, the online agent may use the estimations $\hat{g}_{t+1|t}$ and $\hat{h}_{t|t}$ (see (5.14)) as the estimations of $g^*_{t+1|t}$ and $h^*_{t|t}$, respectively. Note that this is different than the case when true dynamics are known (Lin, Preiss, Anand, et al., 2023), where the online agent can directly construct multi-step dynamics/cost ($g^*_{t+1|t}$ and $h^*_{t|t}$) or compute the exact Jacobian matrices.

Another important definition that we require is the *projected gradient*, which is introduced in Hazan, Singh, and Zhang, 2017 to accommodate the challenge of converging to stationary points on a constrained set.

Definition 5.C.2 (Projected gradient)**.** *Let* $F : \Theta \to \mathbb{R}$ *be a differentiable function on a closed convex set* $\Theta \subseteq \mathbb{R}^d$*. For* $\eta > 0$*, the* (Θ, η)*-projected gradient of* F *is defined as*

$$\nabla_{\Theta,\eta} F(\theta) \coloneqq \frac{1}{\eta}\left(\theta - \Pi_\Theta(\theta - \eta \nabla F(\theta))\right).$$

When Θ is equal to the whole Euclidean space $\mathbb{R}^d$ (unconstrained), the project gradient in Definition 5.C.2 will be identical with the normal gradient $\nabla F(\theta)$. This concept of projected gradient is used to define the local regret in Theorem 5.4.8 and Appendix 5.E to study online gradient descent for online nonconvex optimization with constraints.

5.D Assumptions of Meta-Framework for Unknown Dynamics

We state our key assumptions below: Assumption 5.2.1 is about the Lipschitzness/smoothness properties of the dynamics, policy, nonlinear residual, and the cost functions.

[6] z_t is an auxiliary variable to denote the state at t under initial state x_τ and parameters $\theta_{\tau:t}$.

Assumption 5.D.1. *The dynamics $\phi_{0:T-1}$, policies $\psi_{0:T-1}$, residuals $f_{0:T-1}$, and costs $h_{0:T-1}$ are differentiable at every time step and satisfy that, for any convex compact sets $\mathcal{X} \subseteq \mathbb{R}^n, \mathcal{U} \subseteq \mathcal{R}^m$, one can find Lipschitzness/smoothness constants (can depend on $\mathcal{X}$ and $\mathcal{U}$) such that:*

1. *$\phi_t(x, u)$ is $(L_{\phi,x}, L_{\phi,u})$-Lipschitz and $(\ell_{\phi,x}, \ell_{\phi,u})$-smooth in (x, u) on $\mathcal{X} \times \mathcal{U}$.*
2. *$\psi_t(x, \theta)$ is $(L_{\psi,x}, L_{\psi,\theta})$-Lipschitz and $(\ell_{\psi,x}, \ell_{\psi,\theta})$-smooth in (x, θ) on $\mathcal{X} \times \Theta$.*
3. *$f_t(x, a)$ is $(L_{f,x}, L_{f,a})$-Lipschitz and $(\ell_{f,x}, \ell_{f,a})$-smooth in (x, a) on $\mathcal{X} \times \mathcal{A}$.*
4. *$h_t(x, u, \theta)$ is $(L_{h,x}, L_{h,u}, L_{h,\theta})$-Lipschitz and $(\ell_{h,x}, \ell_{h,u}, \ell_{h,\theta})$-smooth in (x, u, θ) on $\mathcal{X} \times \mathcal{U} \times \Theta$.*

Compared with Assumption 2.1 in Lin, Preiss, Anand, et al., 2023, our Assumption 5.2.1 additionally assumes the Lipschitzness and smoothness of the nonlinear residual function f_t, which is part of our dynamics and policy classes. The second assumption (Assumption 5.D.2) is on the contractive perturbation and the stability of the multi-step dynamics $g^*_{t|\tau}$.

Assumption 5.D.2. *Let $\mathcal{G}$ denote the set of all possible sequences $\{\phi_t, f_t, w_t, \psi_t\}_{t \in \mathcal{T}}$ the environment may provide. For a fixed $\epsilon_\theta \in \mathbb{R}_{\geq 0}$, the ϵ_θ-time-varying contractive perturbation holds with $(R_C, \bar{C}, \rho)$ for any sequence in $\mathcal{G}$. The ϵ_θ-time-varying stability holds with $R_S < R_C$ for any sequence in $\mathcal{G}$. We assume that the initial state satisfies $\|x_0\| < (R_C - R_S)/\bar{C}$. Further, we assume that if $\{\phi, f, w, \psi\}$ is the dynamics/residual/disturbance/policy at an intermediate time step of a sequence in $\mathcal{G}$, then the time-invariant sequence $\{\phi, f, w, \psi\}_{\times T}$ is also in $\mathcal{G}$.*[7]

Compared with Assumption 2.2 in Lin, Preiss, Anand, et al., 2023, our Assumption 5.D.2 also includes the disturbance w_t as a part of the system configuration. This is because for every time t, $g^*_{t+1|t}$ is formed by ϕ_t, π_t, and w_t together. While w_t can also be viewed as a part of the dynamics ϕ_t, we choose to represent it separately because we will leverage the randomness of w_t to bound the first-order model mismatches of EST. Like Lin, Preiss, Anand, et al., 2023, in Assumption 5.D.2, we assume there exists a positive real number $\bar{R}_C$ such that $R_C > \bar{R}_C > R_S + \bar{C}\|x_0\|$. Here, we introduce the real constant $\bar{R}_C$ because R_C can be $+\infty$ when time-varying contractive perturbation (Definition 5.2.2) holds globally. Similarly, to leverage the

[7] For $\{\phi, f, w, \psi\}_{\times T}$ to be in $\mathcal{G}$, it must satisfy other assumptions about contractive perturbation and stability that we impose on $\mathcal{G}$ but does not need to occur in real problem instances. This assumption can be made without the loss of generality for time-invariant dynamics and policy classes.

Lipschitzness/smoothness property, we require $\mathcal{X} \supseteq B(0, R_x)$ where $R_x \geq \bar{C}\bar{R}_C + R_S$ and $\mathcal{U} = \{-f(x, a) + \pi(x, \theta) \mid x \in \mathcal{X}, \theta \in \Theta, a \in \mathcal{A}, \pi, f \in \mathcal{G}\}$. Since the coefficients in Assumption 5.2.1 depend on $\mathcal{X}$ and $\mathcal{U}$, we will set $\mathcal{X} = B_n(0, R_x)$ and $R_x = \bar{C}\bar{R}_C + R_S$ by default when presenting these constants. We also set $\mathcal{Y} = B_p(0, R_y)$ with $R_y = \frac{\bar{C} L_{\phi,u} L_{\psi,\theta}}{\rho(1-\rho)}$, so that the internal state y_t will stay in $\mathcal{Y}$.

It is straightforward to verify that the joint dynamics of M-GAPS satisfy the three properties required by the meta-framework. We state this result in Lemma 5.D.1.

Lemma 5.D.1. *Under Assumptions 5.2.1 and 5.D.2, M-GAPS (Algorithm 6) satisfy Properties 5.4.1, 5.4.2, and 5.4.3 when applied to dynamics* (5.12) *and policy class* (5.13).

We present the specific constants and the formal proof of Lemma 5.D.1 in Section 5.E.

For the part of EST that is instantiated with the gradient estimator (Algorithm 9), we first introduce an assumption about the magnitude of the nonlinear residual to guarantee that (several) bad estimations of the unknown model parameters will not destabilize the system or violate the constraints of the contractive perturbation property.

Assumption 5.D.3. *The set of all possible model parameter $\mathcal{A}$ is a convex compact subset of $\mathbb{R}^k$. For any fixed $x \in B_n(0, R_x)$, $f_t(x, \cdot) : \mathcal{A} \to \mathbb{R}^m$ is an affine function whose gradient is uniformly bounded, i.e., for some positive constant D'_f, $\|\nabla_x f_t(x, a)\| \leq D'_f$ hold for all $a \in \mathcal{A}$. It also satisfies that for any $a, a' \in \mathcal{A}$,*

$$\|f_t(x, a) - f_t(x, a')\| \leq C_f, \ \|\nabla_x f_t(x, a) - \nabla_x f_t(x, a')\|_F \leq \beta \leq C'_f, \text{ and}$$
$$\left\|\nabla_x^2 f_t(x, a)_i - \nabla_x^2 f_t(x, a')_i\right\|_F \leq \gamma, \text{ for any dimension } i \in [1 : m]$$

hold with some positive constants β, γ, and the upper bounds C_f and C'_f are given by

$$C_f = \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C\alpha_x}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C\alpha_x}, \frac{\bar{\zeta}}{2\alpha_\theta}\right\},$$
$$C'_f = \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C\beta_x}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C\beta_x}, \frac{\bar{\zeta}}{2\beta_\theta}\right\}.$$

The expressions of $\alpha_x, \beta_x, \alpha_\theta, \beta_\theta, L_{\theta,x}, L_{\theta,y}, R_x^*, R_y^*$, and $\bar{\zeta}$ are given in Section 5.E.

Note that we need Assumption 5.D.3 to bound the prediction errors uniformly because even if an online parameter estimator performs well in the long term (e.g., achieving a regret bound on the total prediction errors), it may incur a large error at a single time step that can potentially destabilize the system especially when a_t^* changes abruptly. Our simulation (see Appendix A in Lin, Preiss, Xie, et al., 2024) provides a good illustration of this intuition: The model prediction error may increase dramatically right after the system switches to a different true model parameter a_t^*; Then, the error converges back to near zero as the gradient estimator learns the model. Addressing this challenge with other assumptions like slowly time-varying a_t^* is an interesting future direction.

The second assumption we need is about the randomness provided by the environment:

Assumption 5.D.4. *The total path length of the true model parameters satisfies*

$$1 + \sum_{t=1}^{T-1} \left\| a_t^* - a_{t-1}^* \right\| \leq C_p$$

for some positive constant C_p. At every time step t, the noisy observation $\tilde{f}_t$ satisfies that

$$\left\| \tilde{f}_t - f_t(x_t, a_t^*) \right\| \leq e_f, \text{ and } \mathbb{E}[\tilde{f}_t \mid \mathcal{F}_t] = f_t(x_t, a_t^*).$$

Further, the random disturbance w_t in the dynamical system (5.12) *satisfies that*

$$\|w_t\| \leq \bar{\epsilon},\ \mathbb{E}[w_t \mid \mathcal{F}_t'] = 0, \text{ and } Cov(w_t \mid \mathcal{F}_t') \succeq c\bar{\epsilon}^2 I.$$

Here, $\bar{\epsilon}$ satisfies that

$$(C_f + e_f)\left(2D_f'\sqrt{3C_p/T}\right)^{\frac{1}{3}} \leq \bar{\epsilon} \leq \min\{\frac{1}{4}, \frac{1}{2\gamma}, \frac{1}{4\beta\gamma}\},$$

where β and γ are defined in Assumption 5.D.3.

Intuitively, Assumption 5.D.4 put requirements on both the lower and upper bounds of the level of randomness in the system. The lower bound $\bar{\epsilon} = \Omega(T^{-1/6})$ is required due to the condition $R_0^\ell(T) \leq \bar{\epsilon}^3 T$ in Theorem 5.4.5. This guarantees that the zeroth-order regret is sufficiently small to be used for bounding the first-order gradients in Taylor's expansion, which are multiplied by $\bar{\epsilon}^2$ when we take the square. The upper bound of $\bar{\epsilon}$ is required to ignore the higher-order terms in Taylor's expansion. With these assumptions, we show the following guarantee on the total prediction error achieved by the gradient estimator (Algorithm 9).

Lemma 5.D.2. *Under Assumptions 5.D.3 and 5.D.4, the total squared zeroth-order prediction errors of the gradient estimator can be bounded by* $\mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t^2\right] \leq 2\sqrt{3}(C_f + e_f)^3 D_f' \sqrt{C_p T}$.

We defer the proof of Lemma 5.D.2 to Section 5.E. Note that our meta-framework only requires us to bound the total squared zeroth-order prediction error incurred by an instantiation of EST. Under Assumptions 5.D.3 and 5.D.4, we can apply Theorem 5.4.5 in the meta-framework to bound the total squared first-order prediction error of the gradient estimator by $\mathbb{E}\left[\sum_{t=0}^{T-1} (\varepsilon_t')^2\right] = O(m\bar{\epsilon}T)$. Recall that m is the dimension of the unknown component, which is identical with the control input in this application.

In Lemmas 5.D.1 and 5.D.2, we have shown that M-GAPS and the gradient estimator satisfy all the required properties for ALG and EST, respectively, in our meta-framework. Therefore, we can obtain the local regret guarantees for instantiating our meta-framework with M-GAPS and the gradient estimator in the application with matched-disturbance dynamics (Theorem 5.4.8).

5.E Proof of Meta-Framework for Unknown Dynamics

Proof of Theorem 5.4.3

To simplify the notation, we define

$$\bar{\varepsilon} := \min\left\{\frac{\sqrt{2}\bar{\zeta}}{4(L_{\theta,x} + L_{\theta,y})C}, \frac{\min\{R_x - R_x^*, R_y - R_y^*\}}{C}\right\}.$$

By the assumption, we know that the following inequality holds for all time step t:

$$(\alpha_x + \alpha_y)\varepsilon_t + (\beta_x + \beta_y)\varepsilon_t' \leq \bar{\varepsilon}. \tag{5.65}$$

Now we show that $\|x_t\| \leq R_x$, $\|y_t\| \leq R_y$, and $\|\theta_t - \theta_{t-1}\| \leq \epsilon_\theta$ by induction. These inequalities hold for time step 0. Suppose they hold for all time steps $\tau \leq t$. Then, for time step $t + 1$, by Property 5.4.2, we see that

$$\|\tilde{x}_{t+1}\| \leq R_x^*, \text{ and } \|\tilde{y}_{t+1}\| \leq R_y^*. \tag{5.66}$$

By Property 5.4.2 about the contraction of states x_t and y_t under policy parameters $\theta_{0:t}$, we see that

$$\begin{aligned}&\|(x_{t+1}, y_{t+1}) - (\tilde{x}_{t+1}, \tilde{y}_{t+1})\| \\ &\leq \sum_{\tau=0}^{t} \left\| q_{t+1|t+1-\tau}^{(x,y)*}(x_{t+1-\tau}, y_{t+1-\tau}, \theta_{t+1-\tau:t}) - q_{t+1|t-\tau}^{(x,y)*}(x_{t-\tau}, y_{t-\tau}, \theta_{t-\tau:t}) \right\| \quad (5.67a)\end{aligned}$$

$$\leq \sum_{\tau=0}^{t} \gamma(\tau) \left\| (x_{t+1-\tau}, y_{t+1-\tau}) - q_{t+1-\tau|t-\tau}^{(x,y)*}(x_{t-\tau}, y_{t-\tau}, \theta_{t-\tau}) \right\| \tag{5.67b}$$

$$\leq \sum_{\tau=0}^{t} \gamma(\tau) \left((\alpha_x + \alpha_y)\varepsilon_{t-\tau} + (\beta_x + \beta_y)\varepsilon'_{t-\tau} \right) \tag{5.67c}$$

$$\leq \sum_{\tau=0}^{t} \gamma(\tau)\bar{\varepsilon} \leq C\bar{\varepsilon} \tag{5.67d}$$

$$\leq \min\{R_x - R_x^*, R_y - R_y^*\}, \tag{5.67e}$$

where we use the triangle inequality in (5.67a); we use the contractive perturbation property in Property 5.4.2 in (5.67b); we use the induction assumption and Property 5.4.1 in (5.67c); we use (5.65) in (5.67d) and the definition of $\bar{\varepsilon}$ in (5.67e). By (5.66) and (5.67), we see that

$$\begin{aligned} \|x_{t+1}\| &\leq \|\tilde{x}_{t+1}\| + \|\tilde{x}_{t+1} - x_{t+1}\| \leq R_x, \text{ and} \\ \|y_{t+1}\| &\leq \|\tilde{y}_{t+1}\| + \|\tilde{y}_{t+1} - y_{t+1}\| \leq R_y. \end{aligned} \tag{5.68}$$

Note that we can construct the disturbance sequence $\{\zeta_t\}$ in Property 5.4.3 such that the dynamics

$$\begin{pmatrix} \tilde{x}_{t+1} \\ \tilde{y}_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*) + \begin{pmatrix} 0 \\ 0 \\ \zeta_t \end{pmatrix}$$

produce the same policy parameter sequence $\{\theta_t\}$ as the dynamics

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \\ \theta_{t+1} \end{pmatrix} = q_t(x_t, y_t, \theta_t, \hat{a}_t).$$

Therefore, under this construction, we see that

$$\begin{aligned} \|\zeta_t\| &\leq \left\|\theta_{t+1} - q_t^\theta(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*)\right\| \\ &= \left\|q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) - q_t^\theta(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*)\right\| \\ &\leq \left\|q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) - q_t^\theta(x_t, y_t, \theta_t, a_t^*)\right\| \\ &\quad + \left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(\tilde{x}_t, \tilde{y}_t, \theta_t, a_t^*)\right\| && (5.69a) \\ &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + L_{\theta,x}\|x_t - \tilde{x}_t\| + L_{\theta,y}\|y_t - \tilde{y}_t\| && (5.69b) \\ &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}(L_{\theta,x} + L_{\theta,y})\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| && (5.69c) \\ &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}C\bar{\varepsilon}(L_{\theta,x} + L_{\theta,y}) \leq \bar{\zeta}, && (5.69d) \end{aligned}$$

where we use the triangle inequality in (5.69a); we use Property 5.4.1 in (5.69b); we use the inequality

$$\|x_t - \tilde{x}_t\| + \|y_t - \tilde{y}_t\| \leq \sqrt{2}\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\|$$

in (5.69c) and (5.67d) in (5.69d). Thus, by Property 5.4.3, we see that $\|\theta_{t+1} - \theta_t\| \leq \epsilon_\theta$. Therefore, we have shown that

$$\|x_t\| \leq R_x, \|y_t\| \leq R_y, \text{ and } \|\theta_t - \theta_{t-1}\| \leq \epsilon_\theta$$

hold for all time step t by induction.

By (5.69c) and (5.67c), we also see that

$$\begin{aligned}\|\zeta_t\| &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t + \sqrt{2}(L_{\theta,x} + L_{\theta,y})\|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \\ &\leq \alpha_\theta \varepsilon_t + \beta_\theta \varepsilon'_t \\ &\quad + \sqrt{2}(L_{\theta,x} + L_{\theta,y}) \sum_{\tau=0}^{t-1} \gamma(\tau) \left((\alpha_x + \alpha_y)\varepsilon_{t-1-\tau} + (\beta_x + \beta_y)\varepsilon'_{t-1-\tau}\right). \quad (5.70)\end{aligned}$$

Summing (5.70) over $t = 0, 1, \ldots, T-1$ gives that

$$\begin{aligned}\sum_{t=0}^{T-1} \|\zeta_t\| \leq &\left(\alpha_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\alpha_x + \alpha_y)\right) \sum_{t=0}^{T-1} \varepsilon_t \\ &+ \left(\beta_\theta + \sqrt{2}C(L_{\theta,x} + L_{\theta,y})(\beta_x + \beta_y)\right) \sum_{t=0}^{T-1} \varepsilon'_t.\end{aligned}$$

Summing (5.67c) over $t = 0, 1, \ldots, T-1$ gives that

$$\sum_{t=1}^{T} \|(x_t, y_t) - (\tilde{x}_t, \tilde{y}_t)\| \leq C\left((\alpha_x + \alpha_y)\sum_{t=0}^{T-1} \varepsilon_t + (\beta_x + \beta_y)\sum_{t=0}^{T-1} \varepsilon'_t\right).$$

Proof of Theorem 5.4.5

To simplify the notation, we let $\check{x}_{t+1} \coloneqq \mathbb{E}[x_{t+1} \mid \mathcal{G}_t]$ and let $\iota_{t+1} \coloneqq x_{t+1} - \check{x}_{t+1}$.

We first focus on one dimension i of the model mismatch. By Taylor's expansion, we see that

$$e_t(x_t, \hat{a}_t)_i = e_t(\check{x}_t, \hat{a}_t)_i + \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \iota_t + \frac{1}{2}\iota_t^\top \nabla_x^2 e_t(\tilde{x}_t, \hat{a}_t)_i \iota_t, \quad (5.71)$$

where $\tilde{x}_t = \omega x_t + (1-\omega)\check{x}_t$ for some $\omega \in [0, 1]$. Note that we have

$$\mathbb{E}\left[e_t(\check{x}_t, \hat{a}_t)_i \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \iota_t \mid \mathcal{G}_{t-1}\right] = e_t(\check{x}_t, \hat{a}_t)_i \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \mathbb{E}\left[\iota_t \mid \mathcal{G}_{t-1}\right] = 0. \quad (5.72)$$

Therefore, we see that the conditional expectation of the squared estimation error of one dimension i can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[e_t(x_t, \hat{a}_t)_i^2 \mid \mathcal{G}_{t-1}\right] \\
&\geq e_t(\check{x}_t, \hat{a}_t)_i^2 + \nabla_x e_t(\check{x}_t, \hat{a}_t)_i^\top \mathrm{Cov}(\iota_t \mid \mathcal{G}_{t-1}) \nabla_x e_t(\check{x}_t, \hat{a}_t)_i \\
&\quad - \bar{\epsilon}^2 \gamma_e |e_t(\check{x}_t, \hat{a}_t)_i| - \bar{\epsilon}^3 \beta_e \gamma_e \qquad (5.73a) \\
&\geq e_t(\check{x}_t, \hat{a}_t)_i^2 + \underline{\sigma} \|\nabla_x e_t(\check{x}_t, \hat{a}_t)_i\|^2 - \bar{\epsilon}^2 \gamma_e |e_t(\check{x}_t, \hat{a}_t)_i| - \bar{\epsilon}^3 \beta_e \gamma_e. \qquad (5.73b)
\end{aligned}$$

Summing over $t = 1, \ldots, T$ and taking expectation on both sides gives that

$$\begin{aligned}
R_0^\ell(T) \geq{}& \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] + \underline{\sigma}\mathbb{E}\left[\sum_{t=1}^T \|\nabla_x e_t(\check{x}_t, \hat{a}_t)\|_i^2\right] - \bar{\epsilon}^2 \gamma_e \mathbb{E}\left[\sum_{t=1}^T |e_t(\check{x}_t, \hat{a}_t)_i|\right] \\
&- \bar{\epsilon}^3 \beta_e \gamma_e T. \qquad (5.74)
\end{aligned}$$

Now we show that $\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] \leq \bar{\epsilon}^2 T$. For the sake of contradiction, suppose

$$\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] > \bar{\epsilon}^2 T.$$

By (5.74), we see that

$$\begin{aligned}
R_0^\ell(T) &\geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] - \bar{\epsilon}^2 \gamma_e \mathbb{E}\left[\sum_{t=1}^T |e_t(\check{x}_t, \hat{a}_t)_i|\right] - \bar{\epsilon}^3 \beta_e \gamma_e T \\
&\geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] - \bar{\epsilon}^2 \gamma_e \sqrt{\mathbb{E}\left[\left(\sum_{t=1}^T |e_t(\check{x}_t, \hat{a}_t)_i|\right)^2\right]} - \bar{\epsilon}^3 \beta_e \gamma_e T \qquad (5.75a) \\
&\geq \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] - \bar{\epsilon}^2 \gamma_e \sqrt{T \cdot \mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right]} - \bar{\epsilon}^3 \beta_e \gamma_e T \qquad (5.75b) \\
&= \sqrt{\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right]} \cdot \left(\sqrt{\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right]} - \bar{\epsilon}^2 \gamma_e \sqrt{T}\right) - \bar{\epsilon}^3 \beta_e \gamma_e T \\
&> \frac{1}{4} \bar{\epsilon}^2 T, \qquad (5.75c)
\end{aligned}$$

where we use Jensen's inequality in (5.75a); we use Cauchy-Schwarz inequality in (5.75b); we use the assumptions that $\bar{\epsilon}\gamma_e \leq \frac{1}{2}$ and $\bar{\epsilon}\beta_e\gamma_e \leq \frac{1}{4}$ in (5.75c). (5.75) contradicts with our assumption that $R(T) \leq \bar{\epsilon}^3 T$. Thus, we have shown that $\mathbb{E}\left[\sum_{t=1}^T e_t(\check{x}_t, \hat{a}_t)_i^2\right] \leq \bar{\epsilon}^2 T$.

Using the same argument as (5.75a) and (5.75b), we see that the expectation of the total estimation error can be upper bounded by

$$\mathbb{E}\left[\sum_{t=1}^{T} |e_t(\check{x}_t, \hat{a}_t)_i|\right] \leq \sqrt{T \cdot \mathbb{E}\left[\sum_{t=1}^{T} e_t(\check{x}_t, \hat{a}_t)_i^2\right]} \leq \bar{\epsilon}T. \tag{5.76}$$

Substituting (5.76) into (5.74) gives that

$$\begin{aligned}\underline{\sigma}\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla_x e_t(\check{x}_t, \hat{a}_t)_i\|^2\right] &\leq R(T) + \bar{\epsilon}^2\gamma_e\mathbb{E}\left[\sum_{t=1}^{T} |e_t(\check{x}_t, \hat{a}_t)_i|\right] + \bar{\epsilon}^3\beta_e\gamma_e T\\ &\leq (1 + \gamma_e + \beta_e\gamma_e)\bar{\epsilon}^3 T.\end{aligned}$$

Therefore, we see that

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla_x e_t(x_t, \hat{a}_t)_i\|^2\right]\\ &\leq 2\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla_x e_t(\check{x}_t, \hat{a}_t)_i\|^2\right] + 2\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla_x e_t(\check{x}_t, \hat{a}_t)_i - \nabla_x e_t(x_t, \hat{a}_t)_i\|^2\right]\\ &\leq \frac{2}{\underline{\sigma}}(1 + \gamma_e + \beta_e\gamma_e)\bar{\epsilon}^3 T + 2\gamma_e^2\bar{\epsilon}^2 T.\end{aligned} \tag{5.77}$$

Summing (5.77) over dimensions $i \in [1 : k]$ finishes the proof of Theorem 5.4.5.

Proof of Lemma 5.D.1

When applied to the dynamical system (5.12) and the policy class (5.13), the joint dynamics induced by applying M-GAPS with exact model parameters $a^*_{0:T-1}$ are given by

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \tag{5.78a}$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g^*_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \tag{5.78b}$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right). \tag{5.78c}$$

The joint dynamics induced by applying M-GAPS with inexact parameters $\hat{a}_{0:T-1}$ are given by

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, \hat{a}_t) = \phi_t(x_t, \psi_t(x_t, \theta_t) + f_t(x_t, a_t^*) - f_t(x_t, \hat{a}_t)) + w_t, \tag{5.79a}$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t, \hat{a}_t) = \left.\frac{\partial g^*_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \tag{5.79b}$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial \hat{h}_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right), \tag{5.79c}$$

where recall that we view $\hat{a}_{0:T-1}$ as external inputs as discussed in Section 5.4.

Since Lemma 5.D.1 consists three properties, we show them separately in Lemmas 5.E.1-5.E.3.

Lemma 5.E.1. *Consider the dynamical system*

$$\begin{aligned}
x_{t+1} &= q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t,\\
y_{t+1} &= q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g^*_{t+1|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1|t}}{\partial \theta_t}\right|_{x_t,\theta_t},\\
\theta_{t+1} &= q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right).
\end{aligned}$$

For any $x_t, y_t, \theta_t, \hat{a}_t$ that satisfies

$$\|x_t\| \le R_x, \|y_t\| \le R_y, \theta_t \in \Theta, \hat{a}_t \in \mathcal{A},$$

the following Lipschitzness conditions hold:

$$\begin{aligned}
\left\|q_t^x(x_t, y_t, \theta_t, a_t^*) - q_t^x(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\le \alpha_x \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_x \varepsilon_t'(x_t, \hat{a}_t, a_t^*),\\
\left\|q_t^y(x_t, y_t, \theta_t, a_t^*) - q_t^y(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\le \alpha_y \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_y \varepsilon_t'(x_t, \hat{a}_t, a_t^*),\\
\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t)\right\| &\le \alpha_\theta \varepsilon_t(x_t, \hat{a}_t, a_t^*) + \beta_\theta \varepsilon_t'(x_t, \hat{a}_t, a_t^*),
\end{aligned}$$

where

$$\begin{aligned}
&\alpha_x = \ell_{h,u} L_{\psi,\theta},\ \beta_x = \alpha_y = \beta_y = 0,\\
&\alpha_\theta = \eta\left(R_y(\ell_{h,x} + \ell_{h,u} L_{f,x} + \ell_{h,u} L_{\psi,x}) + \ell_{h,u} L_{\psi,\theta}\right),\ \beta_\theta = \eta R_y L_{h,u}.
\end{aligned}$$

Further, $q_t^\theta(x, y, \theta, a_t^*)$ is $(L_{\theta,x}, L_{\theta,y})$-Lipschitz in (x, y), where

$$\begin{aligned}
L_{\theta,x} &= \eta R_y\left((\ell_{h,x} + \ell_{h,u}(L_{f,x} + L_{\psi,x}))(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x})\right)\\
&\quad + \eta\left(\ell_{h,x} L_{\psi,\theta} + L_{h,u} \ell_{\psi,x} + \ell_{h,u} L_{\psi,\theta}(L_{f,x} + L_{\psi,x})\right),\\
L_{\theta,y} &= \eta(L_{h,x} + L_{h,u}(L_{f,x} + L_{\psi,x})).
\end{aligned}$$

We provide the proof of Lemma 5.E.1 later in this section.

Lemma 5.E.2. *Suppose the sequence $\theta_{0:T-1}$ is given and it satisfies the constraint that $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$ for all time step t. Consider the dynamical system*

$$\begin{aligned} x_{t+1} &= q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \\ y_{t+1} &= q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g_{t+1|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}. \end{aligned}$$

We have that $\|x_t\| \le R_x^* < R_x$, $\|y_t\| \le R_y^* < R_y$ always hold if the system starts from $(x_\tau, y_\tau) = (0, 0)$. Here,

$$R_x^* = R_S, \text{ and } R_y^* = \frac{C_{L,g,\theta}}{1-\rho},$$

where recall that ρ is the decay factor defined in Assumption 5.D.2. Further, from any initial states $(x_\tau, y_\tau), (x'_\tau, y'_\tau)$ that satisfy $\|x_\tau\|, \|x'_\tau\| \le R_x$ and $\|y_\tau\|, \|y'_\tau\| \le R_y$, the trajectory satisfies

$$\left\|(x_t, y_t) - (x'_t, y'_t)\right\| \le \gamma(t-\tau) \cdot \left\|(x_\tau, y_\tau) - (x'_\tau, y'_\tau)\right\|,$$

where

$$\gamma(\tau) = \left(\bar{C} + C_{\ell,g,(x,x)} R_y + C_{\ell,g,(\theta,x)} \bar{C} \tau\right) \rho^\tau.$$

Note that γ satisfies

$$\sum_{\tau=0}^{\infty} \gamma(\tau) \le C, \text{ where } C = \frac{\bar{C} + C_{\ell,g,(x,x)} R_y}{1-\rho} + \frac{C_{\ell,g,(\theta,x)} \bar{C}}{(1-\rho)^2}.$$

The definitions of the coefficients $C_{L,g,\theta}, C_{\ell,g,(x,x)}, C_{\ell,g,(\theta,x)}$ can be found in Lemma 5.B.3 in Section 5.B. And the proof of Lemma 5.E.2 can be found in Appendix 5.E.

Lemma 5.E.3. *Consider the dynamical system*

$$\begin{aligned} x_{t+1} &= q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \\ y_{t+1} &= q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g_{t+1|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g_{t+1|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}, \\ \theta_{t+1} &= q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h_{t|t}^*}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h_{t|t}^*}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right) + \zeta_t. \quad (5.80) \end{aligned}$$

Suppose the learning rate η satisfies $\eta < \min\left\{\frac{(1-\rho)\epsilon_\theta}{C_{L,h,\theta}}, \frac{1-\rho}{2C_{\ell,h,(\theta,\theta)}}\right\}$. When $\|\zeta_t\| \le \bar{\zeta} := \min\{1, \epsilon_\theta - \frac{C_{L,h,\theta}\eta}{1-\rho}\}$ holds for all t, the resulting $\{\theta_t\}$ satisfies the slowly-time-varying constraint $\|\theta_t - \theta_{t-1}\| \le \epsilon_\theta$. Further, the trajectory $\{\theta_t\}$ achieves the local regret guarantee

$$R_\eta^L(T, \{\|\zeta_t\|\}_{0\le t\le T-1}) \le \frac{2}{\eta}(F_0(\theta_0) + S_0) + \frac{2}{1-\rho}(C_{L,h,\theta} S_1 + C_{\ell,h,(\theta,\theta)} \eta S_2),\ \textit{where}$$

$$\begin{aligned}
S_0 &\coloneqq \frac{2\bar{C} L_h (1 + L_{\psi,x} + L_{f,x})(1 + L_{\phi,u})}{(1-\rho)^2 \rho} \cdot (V_{sys} + V_w) \\
&\quad + \frac{2\bar{C} L_h (1 + L_{\psi,x} + L_{f,x})}{1-\rho} \cdot \left(2\bar{C}\bar{R}_C + 2R_S\right), \\
S_1 &\coloneqq \left(\frac{1}{\eta} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right) \sum_{t=0}^{T-1} \|\zeta_t\| + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right) \eta T, \\
S_2 &\coloneqq \left(1 + \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_2 + \hat{C}_4}{(1-\rho)^3}\right) \cdot \\
&\quad \left[\left(\frac{1}{\eta^2} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right) \sum_{t=0}^{T-1} \|\zeta_t\|^2 + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right) \eta^2 T\right].
\end{aligned}$$

Here, the variation intensity is defined as

$$\begin{aligned}
V_{sys} &= \sum_{t=1}^{T-1} \Big(\sup_{x\in\mathcal{X}, u\in\mathcal{U}} \|\phi_t(x,u) - \phi_{t-1}(x,u)\| + \sup_{x\in\mathcal{X}, \theta\in\Theta} \|\psi_t(x,\theta) - \psi_{t-1}(x,\theta)\| \\
&\quad + \sup_{x\in\mathcal{X}, u\in\mathcal{U}, \theta\in\Theta} |h_t(x,u,\theta) - h_{t-1}(x,u,\theta)|\Big),\ \textit{and} \\
V_w &= \sum_{t=1}^{T-1} \|w_t - w_{t-1}\|.
\end{aligned}$$

The bound can be simplified to

$$R_\eta^L(T, \{\|\zeta_t\|\}_{0\le t\le T-1}) = O\left(\frac{1}{\eta}(1 + V_{sys} + V_w) + \eta T + \eta^3 T + \frac{1}{\eta}\sum_{t=1}^{T-1} \|\zeta_t\|\right),$$

where the $O(\cdot)$ notation hides dependence on $\frac{1}{1-\rho}, R_x, R_y, \bar{C}$, and the Lipchitzness/smoothness coefficients defined in Assumption 5.D.1.

The definition of the coefficient $C_{L,h,\theta}$ can be found in Corollary 5.B.4 in Section 5.B. We provide the proof of Lemma 5.E.3 later in this section.

Proof of Lemma 5.D.2

By Assumptions 5.D.3 and 5.D.4, we see that for any $a \in \mathcal{A}$,

$$\begin{aligned}
\tilde{\ell}_t(x_t, a, \tilde{f}_t) &\le \left(f_t(x_t, a) - \tilde{f}_t\right)^2 \\
&= \left((f_t(x_t, a) - f_t(x_t, a_t^*)) - (\tilde{f}_t - f_t(x_t, a_t^*))\right)^2 \le (C_f + e_f)^2.
\end{aligned}$$

We also see that the gradient of the loss $\tilde{\ell}_t$ with respect to a can be bounded above by

$$\left\|\nabla_a \tilde{\ell}_t(x_t, a, \tilde{f}_t)\right\| \leq 2\|\nabla_a f_t(x_t, a)\| \cdot \left|f_t(x_t, a) - \tilde{f}_t\right| \leq 2D'_f(C_f + e_f).$$

By Theorem 10.1 in Hazan, 2016, we know that Algorithm 9 with the learning rate $\iota = \frac{C_f + e_f}{D'_f} \cdot \sqrt{\frac{C_p}{T}}$ always achieves the guarantee that

$$\sum_{t=0}^{T-1} \tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \sum_{t=0}^{T-1} \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t) \leq R_0^\ell(T) \coloneqq 2\sqrt{3}(C_f + e_f)^3 D'_f \sqrt{C_p T}. \quad (5.81)$$

Let $v_t \coloneqq \tilde{f}_t - f_t(x_t, a_t^*)$. We see that

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\left(f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)\right) - v_t\right\|^2 - \|v_t\|^2 \mid \mathcal{F}_t\right] \\
&= \left\|f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)\right\|^2 - 2\left(f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)\right)^\top \mathbb{E}\left[v_t \mid \mathcal{F}_t\right] \\
&= \left\|f_t(x_t, \hat{a}_t) - f_t(x_t, a_t^*)\right\|^2 = \varepsilon_t^2.
\end{aligned}$$

Therefore, we obtain that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \sum_{t=0}^{T-1} \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\mathbb{E}\left[\tilde{\ell}_t(x_t, \hat{a}_t, \tilde{f}_t) - \tilde{\ell}_t(x_t, a_t^*, \tilde{f}_t) \mid \mathcal{F}_t\right]\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\varepsilon_t^2\right] = \mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t^2\right].
\end{aligned}$$

Combining this with (5.81) gives that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \varepsilon_t^2\right] \leq 2\sqrt{3}(C_f + e_f)^3 D'_f \sqrt{C_p T}.$$

Then, we can apply Theorem 5.4.5 to conclude that

$$\mathbb{E}\left[\sum_{t=0}^{T-1} (\varepsilon'_t)^2\right] \leq \frac{2m}{c}(1 + \gamma + \beta\gamma)\bar{\epsilon}T + 2m\gamma^2\bar{\epsilon}^2 T.$$

Proof of Theorem 5.4.8

By Lemma 5.D.2 and Theorem 5.4.5, we know that the expected total prediction errors achieved by the gradient estimator satisfy that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t^2\right] &\leq 2\sqrt{3}(C_f+e_f)^3 D'_f\sqrt{C_p T}, \text{ and}\\ \mathbb{E}\left[\sum_{t=0}^{T-1}(\varepsilon'_t)^2\right] &\leq \frac{2m}{c}(1+\gamma+\beta\gamma)\bar{\epsilon}T + 2m\gamma^2\bar{\epsilon}^2 T.\end{aligned} \tag{5.82}$$

By Hölder's inequality, we see that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t\right] &\leq \sqrt[4]{12}(C_f+e_f)^{\frac{3}{2}}(D'_f)^{\frac{1}{2}}C_p^{\frac{1}{4}}T^{\frac{3}{4}}, \text{ and}\\ \mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon'_t\right] &\leq \sqrt{\frac{2}{c}(1+\gamma+\beta\gamma)+2\gamma^2\bar{\epsilon}}\cdot\sqrt{m\bar{\epsilon}}\cdot T.\end{aligned} \tag{5.83}$$

By Lemma 5.D.1, we know that trajectory $\tilde{\xi}$ achieves the local regret

$$R_L(T, \{\|\zeta_t\|\}_{0\leq t\leq T-1}) = O\left(\frac{1}{\eta}(1+V_{\text{sys}}+V_w)+\eta T+\eta^3 T+\frac{1}{\eta}\sum_{t=1}^{T-1}\|\zeta_t\|\right). \tag{5.84}$$

By Theorem 5.4.3 and Lemma 5.E.1, we know that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\zeta_t\|\right] \leq{}& \left(\alpha_\theta+\sqrt{2}C(L_{\theta,x}+L_{\theta,y})(\alpha_x+\alpha_y)\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t\right]\\ &+\left(\beta_\theta+\sqrt{2}C(L_{\theta,x}+L_{\theta,y})(\beta_x+\beta_y)\right)\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon'_t\right]\\ \leq{}& C_0\eta\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon_t\right]+R_y L_{h,u}\cdot\eta\mathbb{E}\left[\sum_{t=0}^{T-1}\varepsilon'_t\right],\end{aligned} \tag{5.85}$$

where $C = \frac{\bar{C}+C_{\ell,g,(x,x)}R_y}{1-\rho}+\frac{C_{\ell,g,(\theta,x)}\bar{C}}{(1-\rho)^2}$ by Lemma 5.E.2 and

$$\begin{aligned}C_0 ={}& R_y(\ell_{h,x}+\ell_{h,u}L_{f,x}+\ell_{h,u}L_{\psi,x})+\ell_{h,u}L_{\psi,\theta}\\ &+\sqrt{2}C\ell_{h,u}L_{\psi,\theta}\Bigg(R_y\big((\ell_{h,x}+\ell_{h,u}(L_{f,x}+L_{\psi,x}))(1+L_{f,x}+L_{\psi,x})\\ &+L_{h,u}(\ell_{f,x}+\ell_{\psi,x})\big)\\ &+\ell_{h,x}L_{\psi,\theta}+L_{h,u}\ell_{\psi,x}+\ell_{h,u}L_{\psi,\theta}(L_{f,x}+L_{\psi,x})+L_{h,x}+L_{h,u}(L_{f,x}+L_{\psi,x})\Bigg).\end{aligned}$$

Substituting (5.85) into (5.84) and applying (5.82) and (5.83) give us the following bound:

$$R_L(T, \{\|\zeta_t\|\}_{0\leq t\leq T-1}) = O\left(\frac{1}{\eta}(1 + V_{sys} + \bar{\epsilon}\cdot T) + \eta T + (\sqrt{m\bar{\epsilon}} + m\bar{\epsilon})\cdot T\right).$$

Further, by the last statement of Theorem 5.4.3, we obtain that

$$\mathbb{E}\left[\sum_{t=0}^{T-1}(\|x_t - \tilde{x}_t\| + \|y_t - \tilde{y}_t\|)\right] = O\left(T^{3/4} + \sqrt{m\bar{\epsilon}}\cdot T\right).$$

Proof of Lemma 5.E.1

By Assumption 5.D.1, we see that

$$\left\|q_t^x(x_t, y_t, \theta_t, \hat{a}_t) - q_t^x(x_t, y_t, \theta_t, a_t^*)\right\| \leq L_{\phi,u}\left\|f_t(x_t, a_t^*) - f_t(x_t, \hat{a}_t)\right\| = L_{\phi,u}\varepsilon_t. \tag{5.86}$$

We also have that

$$q_t^y(x_t, y_t, \theta_t, \hat{a}_t) = q_t^y(x_t, y_t, \theta_t, a_t^*). \tag{5.87}$$

Note that

$$\begin{aligned} h_{t|t}^*(x_t, \theta_t) &= h_t(x_t, u_t^1, \theta_t), \text{ where } u_t^1 = -f_t(x_t, a_t^*) + \psi_t(x_t, \theta_t),\\ \hat{h}_{t|t}(x_t, \theta_t) &= h_t(x_t, u_t^2, \theta_t), \text{ where } u_t^2 = -f_t(x_t, \hat{a}_t) + \psi_t(x_t, \theta_t). \end{aligned}$$

Therefore, we see that

$$\begin{aligned} &\left\|\frac{\partial h_{t|t}^*}{\partial x_t}\bigg|_{x_t,\theta_t} - \frac{\partial \hat{h}_{t|t}}{\partial x_t}\bigg|_{x_t,\theta_t}\right\|\\ &\leq \left\|\frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t^1,\theta_t} - \frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t^2,\theta_t}\right\| + \left\|\frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t^1,\theta_t}\cdot\frac{\partial f_t}{\partial x_t}\bigg|_{x_t,a_t^*} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t^2,\theta_t}\cdot\frac{\partial f_t}{\partial x_t}\bigg|_{x_t,\hat{a}_t}\right\|\\ &\quad + \left\|\frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t^1,\theta_t}\cdot\frac{\partial \psi_t}{\partial x_t}\bigg|_{x_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t^2,\theta_t}\cdot\frac{\partial \psi_t}{\partial x_t}\bigg|_{x_t,\theta_t}\right\| &\text{(5.88a)}\\ &\leq \ell_{h,x}\varepsilon_t + (\ell_{h,u}L_{f,x}\varepsilon_t + L_{h,u}\varepsilon_t') + \ell_{h,u}L_{\psi,x}\varepsilon_t &\text{(5.88b)}\\ &= (\ell_{h,x} + \ell_{h,u}L_{f,x} + \ell_{h,u}L_{\psi,x})\varepsilon_t + L_{h,u}\varepsilon_t', \end{aligned}$$

where we use the chain rule and the triangle inequality in (5.88a); we use Assumption 5.D.1 in (5.88b). Similarly, we also see that

$$\left\|\frac{\partial h_{t|t}^*}{\partial \theta_t}\bigg|_{x_t,\theta_t} - \frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\bigg|_{x_t,\theta_t}\right\| = \left\|\frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t^1,\theta_t}\cdot\frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t^2,\theta_t}\cdot\frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t}\right\| \tag{5.89a}$$

$$\le \ell_{h,u} L_{\psi,\theta} \varepsilon_t, \tag{5.89b}$$

where we use the chain rule in (5.89a) and Assumption 5.D.1 in (5.89b).

For q_t^θ, we see that

$$\begin{align}
&\left\|q_t^\theta(x_t, y_t, \theta_t, \hat{a}_t) - q_t^\theta(x_t, y_t, \theta_t, a_t^*)\right\| \nonumber\\
&\le \eta \left\| \left(\left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \hat{h}_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \right) \cdot y_t + \left(\left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial \hat{h}_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t} \right) \right\| \tag{5.90a}\\
&\le \eta \left(R_y (\ell_{h,x} + \ell_{h,u} L_{f,x} + \ell_{h,u} L_{\psi,x}) + \ell_{h,u} L_{\psi,\theta} \right) \varepsilon_t + \eta R_y L_{h,u} \varepsilon'_t, \tag{5.90b}
\end{align}$$

where we use the property that projection onto Θ is contractive in (5.90a); we use (5.88) and (5.89) in (5.90b). We also see that

$$\begin{align}
&\left\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x'_t, y'_t, \theta_t, a_t^*)\right\| \nonumber\\
&\le \eta \left\| \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t - \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x'_t,\theta_t} \cdot y'_t + \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x'_t,\theta_t} \right\| \tag{5.91a}\\
&\le \eta \left\| \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x'_t,\theta_t} \right\| \cdot \|y_t\| + \eta \left\| \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x'_t,\theta_t} \right\| \cdot \left\|y_t - y'_t\right\| \nonumber\\
&\quad + \eta \left\| \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h^*_{t|t}}{\partial \theta_t}\right|_{x'_t,\theta_t} \right\|, \tag{5.91b}
\end{align}$$

where we use the property that projection onto Θ is contractive in (5.91a), and apply the triangle inequality in (5.91b). Note that

$$\begin{align*}
h^*_{t|t}(x_t, \theta_t) &= h_t(x_t, u_t, \theta_t), \text{ where } u_t = -f_t(x_t, a_t^*) + \psi_t(x_t, \theta_t),\\
h^*_{t|t}(x'_t, \theta_t) &= h_t(x'_t, u'_t, \theta_t), \text{ where } u'_t = -f_t(x'_t, a_t^*) + \psi_t(x'_t, \theta_t).
\end{align*}$$

Therefore, we see that

$$\begin{align}
&\left\| \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h^*_{t|t}}{\partial x_t}\right|_{x'_t,\theta_t} \right\| \nonumber\\
&\le \left\| \left.\frac{\partial h_t}{\partial x_t}\right|_{x_t,u_t,\theta_t} - \left.\frac{\partial h_t}{\partial x_t}\right|_{x'_t,u'_t,\theta_t} \right\| + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} \cdot \left.\frac{\partial f_t}{\partial x_t}\right|_{x_t,a_t^*} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x'_t,u'_t,\theta_t} \cdot \left.\frac{\partial f_t}{\partial x_t}\right|_{x'_t,a_t^*} \right\| \nonumber\\
&\quad + \left\| \left.\frac{\partial h_t}{\partial u_t}\right|_{x_t,u_t,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x_t,\theta_t} - \left.\frac{\partial h_t}{\partial u_t}\right|_{x'_t,u'_t,\theta_t} \cdot \left.\frac{\partial \psi_t}{\partial x_t}\right|_{x'_t,\theta_t} \right\| \tag{5.92a}
\end{align}$$

$$
\begin{aligned}
&\leq \left\| \frac{\partial h_t}{\partial x_t}\bigg|_{x_t,u_t,\theta_t} - \frac{\partial h_t}{\partial x_t}\bigg|_{x_t',u_t',\theta_t} \right\| + \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \frac{\partial f_t}{\partial x_t}\bigg|_{x_t,a_t^*} \right\| \\
&\quad + \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \frac{\partial f_t}{\partial x_t}\bigg|_{x_t,a_t^*} - \frac{\partial f_t}{\partial x_t}\bigg|_{x_t',a_t^*} \right\| + \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \frac{\partial \psi_t}{\partial x_t}\bigg|_{x_t,\theta_t} \right\| \\
&\quad + \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \frac{\partial \psi_t}{\partial x_t}\bigg|_{x_t,\theta_t} - \frac{\partial \psi_t}{\partial x_t}\bigg|_{x_t',\theta_t} \right\| \qquad (5.92b)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \ell_{h,x}\|x_t - x_t'\| + \ell_{h,u}\|u_t - u_t'\| + L_{f,x}\left(\ell_{h,x}\|x_t - x_t'\| + \ell_{h,u}\|u_t - u_t'\|\right) \\
&\quad + L_{h,u}\ell_{f,x}\|x_t - x_t'\| + L_{\psi,x}\left(\ell_{h,x}\|x_t - x_t'\| + \ell_{h,u}\|u_t - u_t'\|\right) + L_{h,u}\ell_{\psi,x}\|x_t - x_t'\| \qquad (5.92c)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\ell_{h,x}(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x})\right)\|x_t - x_t'\| \\
&\quad + \ell_{h,u}(1 + L_{f,x} + L_{\psi,x})\|u_t - u_t'\| \\
&\leq \left((\ell_{h,x} + \ell_{h,u}(L_{f,x} + L_{\psi,x}))(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x})\right)\|x_t - x_t'\|, \qquad (5.92d)
\end{aligned}
$$

where we use the triangle inequality in (5.92a) and (5.92b); we use Assumption 5.D.1 in (5.92c) and (5.92d). Similarly, we also see that

$$
\begin{aligned}
&\left\| \frac{\partial h^*_{t|t}}{\partial \theta_t}\bigg|_{x_t,\theta_t} - \frac{\partial h^*_{t|t}}{\partial \theta_t}\bigg|_{x_t',\theta_t} \right\| \\
&= \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t,\theta_t} \cdot \frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \cdot \frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t',\theta_t} \right\| \qquad (5.93a) \\
&\leq \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t,u_t,\theta_t} - \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} \right\| + \left\| \frac{\partial h_t}{\partial u_t}\bigg|_{x_t',u_t',\theta_t} \right\| \cdot \left\| \frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t,\theta_t} - \frac{\partial \psi_t}{\partial \theta_t}\bigg|_{x_t',\theta_t} \right\| \qquad (5.93b) \\
&\leq \left(\ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x}\right)\|x_t - x_t'\| + \ell_{h,u}L_{\psi,\theta}\|u_t - u_t'\|, \qquad (5.93c) \\
&\leq \left(\ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x} + \ell_{h,u}L_{\psi,\theta}(L_{f,x} + L_{\psi,x})\right)\|x_t - x_t'\|, \qquad (5.93d)
\end{aligned}
$$

where we use the chain rule in (5.93a); we use the triangle inequality in (5.93b); we use Assumption 5.D.1 in (5.93c). Substituting (5.92) and (5.93) into (5.91) gives that

$$
\begin{aligned}
&\|q_t^\theta(x_t, y_t, \theta_t, a_t^*) - q_t^\theta(x_t', y_t', \theta_t, a_t^*)\| \\
&\leq \eta R_y\left((\ell_{h,x} + \ell_{h,u}(L_{f,x} + L_{\psi,x}))(1 + L_{f,x} + L_{\psi,x}) + L_{h,u}(\ell_{f,x} + \ell_{\psi,x})\right)\|x_t - x_t'\| \\
&\quad + \eta(L_{h,x} + L_{h,u}(L_{f,x} + L_{\psi,x}))\|y_t - y_t'\| \\
&\quad + \eta\left(\ell_{h,x}L_{\psi,\theta} + L_{h,u}\ell_{\psi,x} + \ell_{h,u}L_{\psi,\theta}(L_{f,x} + L_{\psi,x})\right)\|x_t - x_t'\|
\end{aligned}
$$

$$\leq L_{\theta,x}\left\|x_t - x'_t\right\| + L_{\theta,y}\left\|y_t - y'_t\right\|. \tag{5.94}$$

Proof of Lemma 5.E.2

Consider two trajectories $\{x_{t_1:t_2}, y_{t_1:t_2}\}$ and $\{x'_{t_1:t_2}, y'_{t_1:t_2}\}$ given by

$$\begin{aligned} x_{\tau+1} &= \phi_\tau(x_\tau, \psi_t(x_\tau, \theta_\tau)) + w_\tau, \\ y_{\tau+1} &= \left.\frac{\partial g^*_{\tau+1|\tau}}{\partial x_\tau}\right|_{x_\tau,\theta_\tau} \cdot y_\tau + \left.\frac{\partial g^*_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau}, \end{aligned}$$

and

$$\begin{aligned} x'_{\tau+1} &= \phi_\tau(x'_\tau, \psi_t(x'_\tau, \theta_\tau)) + w_\tau, \\ y'_{\tau+1} &= \left.\frac{\partial g^*_{\tau+1|\tau}}{\partial x_\tau}\right|_{x'_\tau,\theta_\tau} \cdot y'_\tau + \left.\frac{\partial g^*_{\tau+1|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_\tau}, \end{aligned}$$

where $\tau = t_1, t_1+1, \ldots, t_2$. Note that by Assumption 5.D.2, we have that $\left\|x_{t_2}\right\| \leq R_S$ and for any x_{t_1}, x'_{t_1} whose norms are upper bounded by R_C

$$\left\|x_{t_2} - x'_{t_2}\right\| \leq \bar{C}\rho^{t_2-t_1}\left\|x_{t_1} - x'_{t_1}\right\|, \tag{5.95}$$

where ρ is the decay factor of the contractive perturbation property defined in Assumption 5.D.2. For the y sequence, note that y_{t_2} and y'_{t_2} can be expressed equivalently as

$$y_{t_2} = \left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x_{t_1},\theta_{t_1:t_2-1}} \cdot y_{t_1} + \sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}, \tag{5.96a}$$

$$y'_{t_2} = \left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}} \cdot y'_{t_1} + \sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x'_\tau,\theta_{\tau:t_2-1}}. \tag{5.96b}$$

By Lemma 5.B.3, we see that if $y_{t_1} = 0$, then

$$\left\|y_{t_2}\right\| = \left\|\sum_{\tau=t_1}^{t_2-1} \left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}\right\| \leq \sum_{\tau=t_1}^{t_2-1} \left\|\left.\frac{\partial g^*_{t_2|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:t_2-1}}\right\| \leq \sum_{\tau=t_1}^{t_2-1} C_{L,g,\theta}\rho^{t_2-\tau} = \frac{C_{L,g,\theta}}{1-\rho}. \tag{5.97}$$

We also see that

$$\begin{aligned} &\left\|y_{t_2} - y'_{t_2}\right\| \\ &= \left\|\left(\left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x_{t_1},\theta_{t_1:t_2-1}} - \left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}}\right) \cdot y_{t_1}\right\| + \left\|\left.\frac{\partial g^*_{t_2|t_1}}{\partial x_{t_1}}\right|_{x'_{t_1},\theta_{t_1:t_2-1}} \cdot (y_{t_1} - y'_{t_1})\right\| \end{aligned}$$

$$+\sum_{\tau=t_1}^{t_2-1}\left\|\frac{\partial g^*_{t_2|\tau}}{\partial\theta_\tau}\bigg|_{x_\tau,\theta_{\tau:t_2-1}}-\frac{\partial g^*_{t_2|\tau}}{\partial\theta_\tau}\bigg|_{x'_\tau,\theta_{\tau:t_2-1}}\right\| \tag{5.98a}$$

$$\leq C_{\ell,g,(x,x)}\rho^{t_2-t_1}\left\|x_{t_1}-x'_{t_1}\right\|\cdot R_y+C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1}-y'_{t_1}\right\|$$

$$+C_{\ell,g,(\theta,x)}\sum_{\tau=t_1}^{t_2-1}\rho^{t_2-\tau}\left\|x_\tau-x'_\tau\right\| \tag{5.98b}$$

$$\leq C_{\ell,g,(x,x)}\rho^{t_2-t_1}\left\|x_{t_1}-x'_{t_1}\right\|\cdot R_y+C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1}-y'_{t_1}\right\|$$

$$+C_{\ell,g,(\theta,x)}\sum_{\tau=t_1}^{t_2-1}\rho^{t_2-\tau}\cdot\bar{C}\rho^{\tau-t_1}\left\|x_{t_1}-x'_{t_1}\right\| \tag{5.98c}$$

$$\leq \left(C_{\ell,g,(x,x)}R_y+C_{\ell,g,(\theta,x)}\bar{C}(t_2-t_1)\right)\rho^{t_2-t_1}\left\|x_{t_1}-x'_{t_1}\right\|+C_{L,g,x}\rho^{t_2-t_1}\left\|y_{t_1}-y'_{t_1}\right\|. \tag{5.98d}$$

Therefore, we see that

$$\left\|(x_{t_2},y_{t_2})-(x'_{t_2},y'_{t_2})\right\|$$

$$\leq \left\|x_{t_2}-x'_{t_2}\right\|+\left\|y_{t_2}-y'_{t_2}\right\| \tag{5.99a}$$

$$\leq \bar{C}\rho^{t_2-t_1}\left\|x_{t_1}-x'_{t_1}\right\|+\left(C_{\ell,g,(x,x)}R_y+C_{\ell,g,(\theta,x)}\bar{C}(t_2-t_1)\right)\rho^{t_2-t_1}\left\|x_{t_1}-x'_{t_1}\right\|$$

$$+\bar{C}\rho^{t_2-t_1}\left\|y_{t_1}-y'_{t_1}\right\| \tag{5.99b}$$

$$\leq \gamma(t_2-t_1)\left\|(x_{t_1},y_{t_1})-(x'_{t_1},y'_{t_1})\right\|, \tag{5.99c}$$

where we use the triangle inequality in (5.99a); we use (5.95) and (5.98) and $\bar{C}=C_{L,g,x}$ in (5.99b); we use the inequality that

$$\left\|x_{t_1}-x'_{t_1}\right\|+\left\|y_{t_1}-y'_{t_1}\right\|\leq\sqrt{2}\left\|(x_{t_1},y_{t_1})-(x'_{t_1},y'_{t_1})\right\|$$

and the definition of $\gamma(\cdot)$ in (5.99c).

Proof of Lemma 5.E.3

We compare the dynamical system (5.80) with the Ideal OGD update rule:

$$\theta_{t+1}=\Pi_\Theta(\theta_t-\eta\nabla F_t(\theta_t)). \tag{5.100}$$

Note that the update on θ_t that the dynamical system (5.80) performs can be written equivalently as

$$\theta_{t+1}=\Pi_\Theta(\theta_t-\eta G_t)+\zeta_t, \tag{5.101}$$

where

$$G_t\coloneqq\sum_{\tau=0}^{t}\frac{\partial h^*_{t|0}}{\partial\theta_{t-\tau}}\bigg|_{x_0,\theta_{0:t}}. \tag{5.102}$$

By Theorem 5.E.7, we know that the bias of the gradient estimator G_t compared with the surrogate cost's gradient ∇F_t can be bounded by

$$\begin{aligned}\|G_t - \nabla F_t(\theta_t)\| \leq{}& \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta \\ &+ \sum_{\tau=0}^{t-1}\left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5(t-\tau)\right)\rho^{t-\tau}\|\zeta_\tau\|,\end{aligned}$$

where the constants $\hat{C}_{0:5}$ are given in Theorem 5.E.7. Let θ_{t+1} be the actual next policy parameter (following the update rule (5.101)). By Lemma 5.E.5, we see that

$$\begin{aligned}\|\theta_{t+1} - \Pi_\Theta(\theta_t - \eta\nabla F_t(\theta_t))\| &\leq \eta\|G_t - \nabla F_t(\theta_t)\| + \|\zeta_t\| \\ &\leq \|\zeta_t\| + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2 \\ &\quad + \eta\sum_{\tau=0}^{t-1}\left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5(t-\tau)\right)\rho^{t-\tau}\|\zeta_\tau\|.\end{aligned}$$

Then, we can apply Theorem 5.E.4 to obtain that

$$\sum_{t=0}^{T-1}\left\|\nabla_{\Theta,\eta}F_t(\theta_t)\right\|^2 \leq \frac{1}{\eta(1-\eta\ell_F)}\left(F_0(\theta_0) + \sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1})\right) + \frac{L_F S_1 + \ell_F\eta S_2}{1-\eta\ell_F}, \tag{5.103}$$

where dist_S is a metric that measures the distance between two surrogate cost functions (see Theorem 5.E.4 for definition), and S_1 and S_2 are given by

$$\begin{aligned}S_1 &:= \left(\frac{1}{\eta} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right)\sum_{t=0}^{T-1}\|\zeta_t\| + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta T, \\ S_2 &:= \left(1 + \frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2 + \hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_2 + \hat{C}_4}{(1-\rho)^3}\right)\cdot \\ &\quad \left[\left(\frac{1}{\eta^2} + \frac{\hat{C}_3 + \hat{C}_5}{(1-\rho)^2} + \frac{\hat{C}_4}{(1-\rho)^3}\right)\sum_{t=0}^{T-1}\|\zeta_t\|^2 + \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta^2 T\right].\end{aligned}$$

By applying Lemma F.4 in Lin, Preiss, Anand, et al., 2023, we can bound the total variational intensity on the surrogate costs by

$$\sum_{t=1}^{T-1}\mathsf{dist}_s(F_t, F_{t-1}) \leq \frac{2\bar{C}L_h(1 + L_{\psi,x} + L_{f,x})(1 + L_{\phi,u})}{(1-\rho)^2\rho}\cdot(V_{sys} + V_w)$$

$$+\frac{2\bar{C}L_h(1+L_{\psi,x}+L_{f,x})}{1-\rho}\cdot\left(2\bar{C}\bar{R}_C+2R_S\right).$$

Substituting the above inequality and $L_F=\frac{C_{L,f,\theta}}{1-\rho}$, $\ell_F=\frac{C_{\ell,h,(\theta,\theta)}}{1-\rho}$ into (5.103) finishes the proof.

Local Regret of Online Gradient Descent

Theorem 5.E.4. *Consider the parameter sequence $\{\theta_t\}$ that satisfies*

$$\left\|\theta_{t+1}-(\theta_t-\eta\nabla_{\Theta,\eta}F_t(\theta_t))\right\|\leq\eta\beta_t,\ \textit{for all } t\geq 0.$$

Suppose at every time t, F_t is ℓ_F-smooth and L_F-Lipschitz in Θ. If the learning rate $\eta\leq\frac{1}{\ell_F}$, then the local regret $\sum_{t=0}^{T-1}\left\|\nabla_{\Theta,\eta}F_t(\theta_t)\right\|^2$ is upper bounded by

$$\frac{1}{\eta(1-\eta\ell_F)}\left(F_0(\theta_0)+\sum_{t=1}^{T-1}\mathsf{dist}_s(F_t,F_{t-1})\right)+\frac{L_F\sum_{t=0}^{T-1}\beta_t+\ell_F\eta\sum_{t=0}^{T-1}\beta_t^2}{1-\eta\ell_F},$$

where $\mathsf{dist}_s(F,F'):=\sup_{\theta\in\Theta}|F(\theta)-F'(\theta)|$.

Next, we state a property of projection onto the compact convex set $\Theta\in\mathbb{R}^d$ in Lemma 5.E.5. This is a classic result in convex optimization (see, for example, Theorem 1.2.1 in Schneider, 2014).

Lemma 5.E.5. *Let q and q' be arbitrary points in $\mathbb{R}^d$. Let $p=\Pi_\Theta(q)$ and $p'=\Pi_\Theta(q')$. Then, the following inequality holds:*

$$\|p-p'\|\leq\|q-q'\|.$$

Now we come back to the proof of Theorem 5.E.4.

Define the quantity

$$\epsilon_t:=\frac{1}{\eta}\left(\theta_{t+1}-(\theta_t-\eta\nabla_{\Theta,\eta}F_t(\theta_t))\right).$$

We see that

$$\theta_{t+1}-\theta_t=-\eta\nabla_{\Theta,\eta}F_t(\theta_t)+\eta\epsilon_t. \tag{5.104}$$

By the smoothness of $F_t(\cdot)$, we see that

$$F_t(\theta_{t+1})\leq F_t(\theta_t)+\langle\nabla F_t(\theta_t),\theta_{t+1}-\theta_t\rangle+\frac{\ell_F}{2}\|\theta_{t+1}-\theta_t\|^2$$

$$= F_t(\theta_t) - \eta\langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) - \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\left\|\nabla_{\Theta,\eta} F_t(\theta_t) - \epsilon_t\right\|^2 \tag{5.105a}$$

$$\begin{aligned}
&= F_t(\theta_t) - \eta\langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) \rangle + \frac{\ell_F \eta^2}{2}\left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 \\
&\quad + \eta\langle \nabla F_t(\theta_t), \epsilon_t \rangle - \ell_F \eta^2 \langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\epsilon_t\|^2,
\end{aligned} \tag{5.105b}$$

where we use (5.104) in (5.105a). Recall that Θ is a closed convex subset of $\mathbb{R}^d$. Since $\theta_t - \eta\nabla_{\Theta,\eta} F_t(\theta_t)$ is the projection of $\theta_t - \eta\nabla F_t(\theta_t)$ onto Θ and $\theta_t \in \Theta$, we have

$$\langle (\theta_t - \eta\nabla F_t(\theta_t)) - (\theta_t - \eta\nabla_{\Theta,\eta} F_t(\theta_t)), \theta_t - (\theta_t - \eta\nabla_{\Theta,\eta} F_t(\theta_t)) \rangle \leq 0.$$

Rearranging terms gives that

$$\langle \nabla F_t(\theta_t), \nabla_{\Theta,\eta} F_t(\theta_t) \rangle \geq \left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2.$$

Substituting this inequality into (5.105) gives that

$$\begin{aligned}
F_t(\theta_{t+1}) &\leq F_t(\theta_t) - \eta\left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 + \frac{\ell_F \eta^2}{2}\left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 \\
&\quad + \eta\langle \nabla F_t(\theta_t), \epsilon_t \rangle - \ell_F \eta^2 \langle \nabla_{\Theta,\eta} F_t(\theta_t), \epsilon_t \rangle + \frac{\ell_F \eta^2}{2}\|\epsilon_t\|^2 \\
&\leq F_t(\theta_t) - \eta(1 - \ell_F \eta)\left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 + \eta\|\nabla F_t(\theta_t)\| \cdot \|\epsilon_t\| \\
&\quad - \frac{\ell_F \eta^2}{2}\left\|\nabla_{\Theta,\eta} F_t(\theta_t) + \epsilon_t\right\|^2 + \ell_F \eta^2 \|\epsilon_t\|^2 && (5.106a) \\
&\leq F_t(\theta_t) - \eta(1 - \ell_F \eta)\left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 + \eta L_F \beta_t + \ell_F \eta^2 \beta_t^2, && (5.106b)
\end{aligned}$$

where we rearrange the terms and use the Cauchy-Schwarz inequality in (5.106a); In (5.106b), we use the assumption $\|\epsilon_t\| \leq \beta_t$. Summing (5.106) over $t = 0, 1, \ldots, T-1$ gives that

$$\begin{aligned}
&\eta(1 - \ell_F \eta) \sum_{t=0}^{T-1} \left\|\nabla_{\Theta,\eta} F_t(\theta_t)\right\|^2 \\
&\leq \sum_{t=0}^{T-1} (F_t(\theta_t) - F_t(\theta_{t+1})) + \eta L_F \sum_{t=0}^{T-1} \beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1} \beta_t^2 \\
&\leq F_0(\theta_0) + \sum_{t=1}^{T-1} (F_t(\theta_t) - F_{t-1}(\theta_t)) + \sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1}) + \eta L_F \sum_{t=0}^{T-1} \beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1} \beta_t^2 && (5.107a) \\
&\leq F_0(\theta_0) + \sum_{t=1}^{T-1} \mathsf{dist}_s(F_t, F_{t-1}) + \eta L_F \sum_{t=0}^{T-1} \beta_t + \ell_F \eta^2 \sum_{t=0}^{T-1} \beta_t^2, && (5.107b)
\end{aligned}$$

where we rearrange the terms and use $F_{T-1}(\theta_T) \geq 0$ in (5.107a); we use the definition of $\mathsf{dist}_s(\cdot, \cdot)$ in (5.107b).

Useful Lemmas

Theorem 5.E.6 bounds the distances between the trajectory of M-GAPS with the imaginary trajectory achieved by using θ_t repeatedly from time step 0. It can be shown using a similar approach as Theorem D.5 in Lin, Preiss, Anand, et al., 2023, while a difference is that we consider an additional disturbance ζ_t in the update rule of policy parameters. We include the proof of Theorem 5.E.6 in Appendix 5.E for completeness.

Theorem 5.E.6. *Suppose Assumptions 5.D.1 and 5.D.2 hold. Let $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ denote the trajectory of*

$$x_{t+1} = q_t^x(x_t, y_t, \theta_t, a_t^*) = \phi_t(x_t, \psi_t(x_t, \theta_t)) + w_t, \tag{5.108a}$$

$$y_{t+1} = q_t^y(x_t, y_t, \theta_t, a_t^*) = \left.\frac{\partial g^*_{t+1\mid t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial g^*_{t+1\mid t}}{\partial \theta_t}\right|_{x_t,\theta_t}, \tag{5.108b}$$

$$\theta_{t+1} = q_t^\theta(x_t, y_t, \theta_t, a_t^*) = \Pi_\Theta\left(\theta_{t+1} - \eta\left(\left.\frac{\partial h^*_{t\mid t}}{\partial x_t}\right|_{x_t,\theta_t} \cdot y_t + \left.\frac{\partial h^*_{t\mid t}}{\partial \theta_t}\right|_{x_t,\theta_t}\right)\right) + \zeta_t. \tag{5.108c}$$

Suppose η and $\bar{\zeta}$ satisfy the constraint that $\bar{\varepsilon} \coloneqq \frac{C_{L,h,\theta}\eta}{1-\rho} + \bar{\zeta} \le \varepsilon$. Then, both $\|G_t\|$ and $\|\nabla F_t(\theta_t)\|$ are upper bounded by $\frac{C_{L,h,\theta}}{1-\rho}$, and the following inequalities holds for any two time steps τ, t ($\tau \le t$):

$$\|\theta_t - \theta_\tau\| \le \frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau)\eta + \sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\|, \text{ and } \|x_\tau - \hat{x}_\tau(\theta_t)\| \le$$

$$\frac{C_{L,h,\theta}C_{L,\phi,\theta}\rho}{(1-\rho)^2}\left((t-\tau) + \frac{1}{1-\rho}\right) \cdot \eta + \frac{C_{L,\phi,\theta}\rho}{1-\rho} \cdot \left(\sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\| + \sum_{\tau'=0}^{\tau-1} \rho^{\tau-\tau'} \|\zeta_{\tau'}\|\right),$$

where we use the notation $\hat{x}_\tau(\theta) \coloneqq g^*_{\tau\mid 0}(x_0, \theta_{\times(\tau+1)}), \forall\theta \in \Theta$. Further, we have that

$$|h_t(x_t, u_t, \theta_t) - F_t(\theta_t)| \le \frac{C_{L,h,\theta}C_{L,\phi,\theta}L_h(1 + L_{\psi,x} + L_{f,x})\rho}{(1-\rho)^3} \cdot \eta + \frac{C_{L,\phi,\theta}L_h(1 + L_{\psi,x} + L_{f,x})\rho}{1-\rho} \cdot \sum_{\tau=0}^{t-1} \rho^{t-\tau}\|\zeta_\tau\|.$$

Recall that we define the gradient approximation G_t for M-GAPS in Algorithm 8. Using this notation, the update rule of $\theta_{0:T-1}$ in joint dynamics (5.108) can be simplified as

$$\theta_{t+1} = \Pi_\Theta\left(\theta_{t+1} - \eta G_t\right) + \zeta_t.$$

To compare the trajectory of M-GAPS with the trajectory achieved by the online gradient descent trajectory $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta\nabla F_t(\theta_t))$, we bound the difference between G_t and $\nabla F_t(\theta_t)$ in Theorem 5.E.7. We provide its proof in Appendix 5.E for completeness.

Theorem 5.E.7 (Gradient Bias)**.** *Suppose Assumptions 5.D.1 and 5.D.2 hold. Let* $\{x_t, u_t, \theta_t\}_{t\in\mathcal{T}}$ *denote the trajectory of* (5.108)*. Suppose* η *and* $\bar{\zeta}$ *satisfy the constraint that* $\bar{\varepsilon} := \frac{C_{L,h,\theta}\eta}{1-\rho} + \bar{\zeta} \le \varepsilon$*. Then, the following holds for all* $\tau \le t$*:*

$$\begin{aligned}
&\left\| \left.\frac{\partial h^*_{t|0}}{\partial \theta_\tau}\right|_{x_0,\theta_{0:t}} - \left.\frac{\partial h^*_{t|0}}{\partial \theta_\tau}\right|_{x_0,(\theta_t)_{\times(t+1)}} \right\| \\
\le\ &\left(\hat{C}_0 + \hat{C}_1(t-\tau) + \hat{C}_2(t-\tau)^2\right)\rho^{t-\tau}\cdot\eta \\
&+ \left(\hat{C}_3\sum_{\tau'=\tau}^{t-1}\|\zeta_{\tau'}\| + \hat{C}_4\sum_{\tau'=\tau}^{t-1}(\tau'-\tau)\|\zeta_{\tau'}\|\right)\cdot\rho^{t-\tau} \\
&+ \hat{C}_5\sum_{\tau'=0}^{\tau-1}\rho^{t-\tau'}\|\zeta_{\tau'}\|.
\end{aligned}$$

for

$$\begin{aligned}
&\hat{C}_0 = \frac{\rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}}{(1-\rho)^3},\ \hat{C}_1 = \frac{(1-\rho)C_{L,h,\theta}C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2}, \\
&\hat{C}_2 = \frac{C_{L,h,\theta}C_{\ell,h,(x,\theta)}C_{L,\phi,\theta}}{1-\rho},\ \hat{C}_3 = \frac{C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}\rho}{1-\rho}, \\
&\hat{C}_4 = C_{\ell,h,(x,\theta)}C_{L,\phi,\theta},\ \hat{C}_5 = \frac{C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}\rho}{1-\rho}.
\end{aligned}$$

Next,

$$\begin{aligned}
\|G_t - \nabla F_t(\theta_t)\| \le\ &\left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1+\hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right)\eta \\
&+ \sum_{\tau=0}^{t-1}\left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5(t-\tau)\right)\rho^{t-\tau}\|\zeta_\tau\|.
\end{aligned}$$

Proof of Theorem 5.E.6

We first use induction to show that for all time step $t \in \mathcal{T}$,

$$\|G_t\| \le \frac{C_{L,h,\theta}}{1-\rho}, x_t \in B_n(0, R_S + \bar{C}\|x_0\|), u_t \in \mathcal{U}, \text{ and } \|\theta_{t+1} - \theta_t\| \le \epsilon_\theta, \quad (5.109)$$

where $\mathcal{U} = \{\psi(x,\theta) - f(x,a) \mid x \in B_n(0,R_x), \theta \in \Theta, a \in \mathcal{A}, (\psi, f) \in \mathcal{G}\}$.

Note that $\|G_0\| \leq C_{L,h,\theta} \leq \frac{C_{L,h,\theta}}{1-\rho}$ by Corollary 5.B.4. We also have $x_0 \in B_n(0, R_S + C\|x_0\|)$ and $u_0 \in \mathcal{U}$.

Suppose $\|G_{t-1}\| \leq \frac{C_{L,h,\theta}}{1-\rho}$ for some $t \geq 1$. Then, since $\eta \leq \frac{(1-\rho)\epsilon_\theta}{C_{L,h,\theta}}$ and the projection onto Θ is a contraction, we see that

$$\|\theta_t - \theta_{t-1}\| \leq \|\eta G_{t-1}\| + \|\zeta_t\| \leq \epsilon_\theta.$$

Suppose $\|\theta_\tau - \theta_{\tau-1}\| \leq \epsilon_\theta$ holds for all $\tau \leq t$, i.e., $\theta_{0:t} \in S_{\epsilon_\theta}(0:t)$. By Lemma D.2 in Lin, Preiss, Anand, et al., 2023, we see that

$$x_t \in B_n(0, R_S + C\|x_0\|), \text{ and } u_t \in \mathcal{U}.$$

Therefore, by taking norm on both sides of the expression of G_t, we see that

$$\begin{aligned}
\|G_t\| &= \left\| \sum_{\tau=0}^{t} \frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\bigg|_{x_{t-\tau},\theta_{t-\tau:t}} \right\| \\
&\leq \sum_{\tau=0}^{t} \left\| \frac{\partial h_{t|t-\tau}}{\partial \theta_{t-\tau}}\bigg|_{x_{t-\tau},\theta_{t-\tau:t}} \right\| && (5.110a) \\
&\leq \sum_{\tau=0}^{t} C_{L,h,\theta}\rho^\tau && (5.110b) \\
&\leq \frac{C_{L,h,\theta}}{1-\rho},
\end{aligned}$$

where we use the triangle inequality in (5.110a) and Corollary 5.B.4 in (5.110b). Note that we can apply Corollary 5.B.4 because $x_t \in B_n(0, R_S + C\|x_0\|)$. Therefore, we have shown (5.109) by induction. One can use the same technique as (5.110) to show $\|\nabla F_t(\theta_t)\| \leq \frac{C_{L,f,\theta}}{1-\rho}$.

Since the projection onto the set Θ is a contraction, we obtain that for any $t > \tau$,

$$\|\theta_t - \theta_\tau\| \leq \frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau)\eta + \sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\|. \tag{5.111}$$

Now we bound the distance between x_τ and $\hat{x}_\tau(\theta_t)$ for $\tau \leq t$. We see that

$$\begin{aligned}
\|x_\tau - \hat{x}_\tau(\theta_t)\| &= \left\| g^*_{\tau|0}(x_0, \theta_{0:\tau-1}) - g^*_{\tau|0}(x_0, (\theta_t)_{\times\tau}) \right\| \\
&\leq \sum_{\tau'=0}^{\tau-1} \left\| g^*_{\tau|0}(x_0, \theta_{0:\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g^*_{\tau|0}(x_0, \theta_{0:\tau'-1}, (\theta_t)_{\times(\tau-\tau')}) \right\| && (5.112a)
\end{aligned}$$

$$\le \sum_{\tau'=0}^{\tau-1} \left\| g^*_{\tau|\tau'}(x_{\tau'}, \theta_{\tau'}, (\theta_t)_{\times(\tau-\tau'-1)}) - g^*_{\tau|\tau'}(x_{\tau'}, (\theta_t)_{\times(\tau-\tau')}) \right\| \tag{5.112b}$$

$$\le \sum_{\tau'=0}^{\tau-1} C_{L,g,\theta}\rho^{\tau-\tau'} \|\theta_t - \theta_{\tau'}\| \tag{5.112c}$$

$$\le \frac{C_{L,h,\theta}C_{L,g,\theta}\eta}{1-\rho} \sum_{\tau'=0}^{\tau-1} \left(\frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau')\eta + \sum_{\tau''=\tau'}^{t-1} \|\zeta_{\tau''}\| \right) \tag{5.112d}$$

$$\begin{aligned} &\le \frac{C_{L,h,\theta}C_{L,\phi,\theta}\rho}{(1-\rho)^2} \left((t-\tau) + \frac{1}{1-\rho} \right) \cdot \eta \\ &\quad + \frac{C_{L,\phi,\theta}\rho}{1-\rho} \cdot \left(\sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\| + \sum_{\tau'=0}^{\tau-1} \rho^{\tau-\tau'} \|\zeta_{\tau'}\| \right), \end{aligned}$$

where we use the triangle inequality in (5.112a); we use the definition of multi-step dynamics in (5.112b); we use Lemma 5.B.3 in (5.112c); we use (5.111) in (5.112d).

Similarly, since $x_t \in B_n(0, R_S + C\|x_0\|)$ and we also see that $\hat{x}_t(\theta_t) \in B_n(0, R_S + C\|x_0\|)$, we obtain that

$$\begin{aligned} |h_t(x_t, u_t, \theta_t) - F_t(\theta_t)| &= |h_t(x_t, u_t, \theta_t) - h_t(\hat{x}_t(\theta_t), \hat{u}_t(\theta_t), \theta_t)| \\ &\le L_h \left(\|x_t - \hat{x}_t(\theta_t)\| + \|u_t - \hat{u}_t(\theta_t)\| \right) \qquad (5.113a) \\ &\le L_h(1 + L_{\psi,x} + L_{f,x})\|x_t - \hat{x}_t(\theta_t)\| \qquad (5.113b) \\ &\le \frac{C_{L,h,\theta}C_{L,\phi,\theta}L_h(1 + L_{\psi,x} + L_{f,x})\rho}{(1-\rho)^3} \cdot \eta \\ &\quad + \frac{C_{L,\phi,\theta}L_h(1 + L_{\psi,x} + L_{f,x})\rho}{1-\rho} \cdot \sum_{\tau=0}^{t-1} \rho^{t-\tau}\|\zeta_\tau\|, \qquad (5.113c) \end{aligned}$$

where we use Assumption 5.D.1 in (5.113a) and (5.113b); we use (5.112) in (5.113c).

Proof of Theorem 5.E.7

To simplify the notation, we adopt the shorthand notations $\hat{x}_\tau(\theta) := g^*_{\tau|0}(x_0, \theta_{\times\tau})$ and $\hat{u}_\tau(\theta) := \pi_\tau(\hat{x}_\tau(\theta), \theta)$ throughout the proof.

We use the triangle inequality to do the decomposition

$$\begin{aligned} &\left\| \frac{\partial h^*_{t|0}}{\partial \theta_\tau}\bigg|_{x_0, \theta_{0:t}} - \frac{\partial h^*_{t|0}}{\partial \theta_\tau}\bigg|_{x_0, (\theta_t)_{\times(t+1)}} \right\| \\ =& \left\| \frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\bigg|_{x_\tau, \theta_{\tau:t}} - \frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\bigg|_{\hat{x}_\tau(\theta_t), (\theta_t)_{\times(t-\tau+1)}} \right\| \end{aligned}$$

$$\begin{aligned}
&\leq \left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,(\theta_t)_{\times(t-\tau)}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
&\quad + \sum_{\tau'=\tau+1}^{t-1} \left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'},(\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'-1},(\theta_t)_{\times(t-\tau'+1)}} \right\|. \qquad (5.114)
\end{aligned}$$

Note that we can apply Corollary 5.B.4 to bound each term in (5.114). For the first term in (5.114), since $x_\tau, \hat{x}_\tau(\theta_t), x_{\tau+1} \in B_n(0, \bar{R}_C)$, we see that

$$\begin{aligned}
&\left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_\tau,(\theta_t)_{\times(t-\tau)}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{\hat{x}_\tau(\theta_t),(\theta_t)_{\times(t-\tau+1)}} \right\| \\
&\leq \rho^{t-\tau}\left(C_{\ell,h,(\theta,x)}\|x_\tau - \hat{x}_\tau(\theta_t)\| + C_{\ell,h,(\theta,\theta)}\|\theta_t - \theta_\tau\|\right) \qquad (5.115a) \\
&\leq \frac{(1-\rho)C_{L,h,\theta}C_{\ell,h,(\theta,x)} + \rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,\theta)}}{(1-\rho)^2} \cdot (t-\tau)\rho^{t-\tau} \cdot \eta \\
&\quad + \frac{\rho C_{L,h,\theta}C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}}{(1-\rho)^3} \cdot \rho^{t-\tau} \cdot \eta \\
&\quad + \frac{C_{L,\phi,\theta}C_{\ell,h,(\theta,x)}\rho}{1-\rho} \cdot \left(\sum_{\tau'=\tau}^{t-1}\|\zeta_{\tau'}\| + \sum_{\tau'=0}^{\tau-1}\rho^{\tau-\tau'}\|\zeta_{\tau'}\|\right) \cdot \rho^{t-\tau}, \qquad (5.115b)
\end{aligned}$$

where we use Corollary 5.B.4 in (5.115a) and Theorem 5.E.6 in (5.115b).

For any $\tau' \in [\tau+1 : t-1]$, since $x_{\tau'}, x_{\tau'+1} \in B_n(0, \bar{R}_C)$, we see that

$$\begin{aligned}
&\left\| \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'},(\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'-1},(\theta_t)_{\times(t-\tau'+1)}} \right\| \\
&= \left\| \left(\left.\frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'},\theta_{\tau'},(\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'},(\theta_t)_{\times(t-\tau'+1)}} \right) \left.\frac{\partial g^*_{\tau'|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'-1}} \right\| \\
&\leq \left\| \left.\frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'},\theta_{\tau'},(\theta_t)_{\times(t-\tau')}} - \left.\frac{\partial h^*_{t|\tau'}}{\partial x_{\tau'}}\right|_{x_{\tau'},(\theta_t)_{\times(t-\tau'+1)}} \right\| \cdot \left\| \left.\frac{\partial g^*_{\tau'|\tau}}{\partial \theta_\tau}\right|_{x_\tau,\theta_{\tau:\tau'-1}} \right\| \\
&\leq C_{\ell,h,(x,\theta)}\rho^{t-\tau'}\|\theta_t - \theta_{\tau'}\| \cdot C_{L,\phi,\theta}\rho^{\tau'-\tau} \qquad (5.116a) \\
&\leq C_{\ell,h,(x,\theta)}C_{L,\phi,\theta} \cdot \rho^{t-\tau} \cdot \left(\frac{C_{L,h,\theta}}{1-\rho} \cdot (t-\tau')\eta + \sum_{\tau''=\tau'}^{t-1}\|\zeta_{\tau''}\|\right), \qquad (5.116b)
\end{aligned}$$

where we use Lemma 5.B.3 and Corollary 5.B.4 in (5.116a); we use Theorem 5.E.6 in (5.116b). Substituting (5.115) and (5.116) into (5.114) finishes the proof of the first inequality.

For the second inequality, recall that G_t and $\nabla F_t(\theta_t)$ are given by

$$G_t \coloneqq \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}^*}{\partial \theta_{t-\tau}}\right|_{x_0, \theta_{0:t}}, \nabla F_t(\theta_t) = \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}^*}{\partial \theta_{t-\tau}}\right|_{x_0, (\theta_t)_{\times(t+1)}}.$$

Therefore, we see that

$$\begin{aligned}
\|G_t - \nabla F_t(\theta_t)\| &= \left\| \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}^*}{\partial \theta_{t-\tau}}\right|_{x_0, \theta_{0:t}} - \sum_{\tau=0}^{t} \left.\frac{\partial h_{t|0}^*}{\partial \theta_{t-\tau}}\right|_{x_0, (\theta_t)_{\times(t+1)}} \right\| \\
&\le \sum_{\tau=0}^{t} \left\| \left.\frac{\partial h_{t|0}^*}{\partial \theta_{t-\tau}}\right|_{x_0, \theta_{0:t}} - \left.\frac{\partial h_{t|0}^*}{\partial \theta_{t-\tau}}\right|_{x_0, (\theta_t)_{\times(t+1)}} \right\| &\text{(5.117a)} \\
&\le \sum_{\tau=0}^{t} \left(\hat{C}_0 + \hat{C}_1 \tau + \hat{C}_2 \tau^2\right) \rho^\tau \eta &\text{(5.117b)} \\
&\quad + \sum_{\tau=0}^{t-1} \left(\hat{C}_3 \sum_{\tau'=\tau}^{t-1} \|\zeta_{\tau'}\| + \hat{C}_4 \sum_{\tau'=\tau}^{t-1} (\tau' - \tau) \|\zeta_{\tau'}\|\right) \cdot \rho^{t-\tau} \\
&\quad + \hat{C}_5 \sum_{\tau=0}^{t-1} \sum_{\tau'=0}^{\tau-1} \rho^{t-\tau'} \|\zeta_{\tau'}\| &\text{(5.117c)} \\
&\le \left(\frac{\hat{C}_0}{1-\rho} + \frac{\hat{C}_1 + \hat{C}_2}{(1-\rho)^2} + \frac{\hat{C}_2}{(1-\rho)^3}\right) \eta \\
&\quad + \sum_{\tau=0}^{t-1} \left(\frac{\hat{C}_3}{1-\rho} + \frac{\hat{C}_4}{(1-\rho)^2} + \hat{C}_5 (t-\tau)\right) \rho^{t-\tau} \|\zeta_\tau\|,
\end{aligned}$$

where we use the triangle inequality in (5.117a); we use the first inequality in Theorem 5.E.7 that we have shown and Corollary 5.B.4 in (5.117c).

5.F Proof of Application: Using Predictions Adaptively

Proof of Lemma 5.5.1

By the perturbation bound in (5.16), we see that for any $t' > t$,

$$\left\|\psi_t^T(y_t, 0_{\times(T-t)}; \tilde{Q})_{x_{t'}} - \psi_t^T(y_t', 0_{\times(T-t)}; \tilde{Q})_{x_{t'}}\right\| \le C_0 \rho_0^{t'-t} \left\|y_t - y_t'\right\|.$$

Therefore, we obtain that for any $t' > t$,

$$\left\|(A_{t'-1} - B_{t'-1}\bar{K}_{t'-1}^{(T)})(A_{t'-2} - B_{t'-2}\bar{K}_{t'-2}^{(T)}) \cdots (A_t - B_t \bar{K}_t^{(T)})\right\| \le C_0 \rho_0^{t'-t}. \quad (5.118)$$

Now we show that

$$\left\|(A_{t'-1} - B_{t'-1}\bar{K}_{t'-1}^{(k)})(A_{t'-2} - B_{t'-2}\bar{K}_{t'-2}^{(k)}) \cdots (A_t - B_t \bar{K}_t^{(k)})\right\| \le C_0 \rho^{t'-t}. \quad (5.119)$$

To see this, we construct a sequence $\hat{v}_{t:T-1}$ such that $\hat{v}_{t+k} = -A_{t+k}\psi_t^{t+k}(x_t, 0_{\times k}; \tilde{Q})_{x_{t+k}}$ and $\hat{v}_\tau = 0$ for $\tau \neq t + k, \tau \in [t : T-1]$. We observe that

$$\begin{aligned}
&\left\|\psi_t^{t+k}(x_t, 0_{\times k}; \tilde{Q}) - \psi_t^T(x_t, 0_{\times(T-t)}; \tilde{Q})\right\| \\
={}& \left\|\psi_t^T(x_t, \hat{v}_{t:T-1}; \tilde{Q}) - \psi_t^T(x_t, 0_{\times(T-t)}; \tilde{Q})\right\| \\
\leq{}& C_0\rho_0^k\|\hat{v}_{t+k}\| \leq C_0^2 a\rho_0^{2k}\|x_t\|,
\end{aligned}$$

where we use (5.16) in the last line. To simplify the notation, we define $M_\tau^{(p)} := A_\tau - B_\tau\bar{K}_\tau^{(p)}$ and $\alpha := C_0^2\rho_0^{2k}ab$. By the above inequality, we see that

$$\left\|M_\tau^{(k)} - M_\tau^{(T)}\right\| \leq \alpha, \forall \tau \in \mathcal{T}. \tag{5.120}$$

Therefore, we obtain that

$$\begin{aligned}
\left\|M_{t'-1}^{(k)}M_{t'-2}^{(k)}\cdots M_t^{(k)}\right\| &\leq \sum_{j=0}^{t'-t}\binom{t'-t}{j}C_0^{j+1}\rho_0^{t'-t}\alpha^j && (5.121a)\\
&\leq C_0\rho_0^{t'-t}\left(1 + C_0\alpha\right)^{t'-t} \leq C_0\rho^{t'-t}, && (5.121b)
\end{aligned}$$

where we use the decomposition $M_\tau^{(k)} = M_\tau^{(T)} + (M_\tau^{(k)} - M_\tau^{(T)})$, the triangle inequality, and (5.120) in (5.121a); we use the condition that

$$k \geq \frac{1}{2}\log\left(C_0^3 ab\rho_0/(\rho - \rho_0)\right)/\log(1/\rho_0)$$

in (5.121b). This finishes the proof of (5.119).

Now we consider two trajectories that apply the same policy parameter sequence but start from different states x_τ and x'_τ. For arbitrary $\theta_{\tau:t-1} = \Theta^{t-\tau}$, we see that

$$\begin{aligned}
&\left\|g_{t|\tau}(x_\tau, \theta_{\tau:t-1}) - g_{t|\tau}(x'_\tau, \theta_{\tau:t-1})\right\| \\
={}& \left\|(A_{t-1} - B_{t-1}\bar{K}_{t-1}^{(k)})(A_{t-2} - B_{t-2}\bar{K}_{t-2}^{(k)})\cdots(A_\tau - B_\tau\bar{K}_\tau^{(k)})(x_\tau - x'_\tau)\right\| && (5.122a)\\
\leq{}& C_0\rho^{t-\tau}\left\|x_\tau - x'_\tau\right\|, && (5.122b)
\end{aligned}$$

where we use the affine expression of π_t (5.17) and the fact that these two trajectories experience the same sequence of disturbances and predictions. This finishes the proof of ε-time-varying contractive perturbation with $\varepsilon = +\infty$ and $R_C = +\infty$.

By the perturbation bound in (5.16), we also see that

$$\left\|\psi_t^{t+k}(0, v_{t:t+k-1}; \tilde{Q})_{u_t} - \psi_t^{t+k}(0, v'_{t:t+k-1}; \tilde{Q})_{u_t}\right\| \leq C_0\sum_{\tau=t}^{t+k-1}\rho_0^{\tau-t}\left\|v_\tau - v'_\tau\right\|.$$

Combining this inequality with the affine relationship in (5.118), we see that

$$\left\|\bar{K}_t^{(k,\tau)}\right\| \leq C_0 \rho_0^{\tau-t}. \tag{5.123}$$

Therefore, we obtain that

$$\begin{aligned}
&\left\|g_{t|\tau}(0, \theta_{\tau:t-1})\right\| \\
&= \left\|\sum_{i=\tau}^{t-1} (A_{t-1} - B_{t-1}\bar{K}_{t-1}^{(k)}) \cdots (A_{i+1} - B_{i+1}\bar{K}_{i+1}^{(k)}) \left(w_i - \sum_{j=i}^{i+k-1} \lambda_i^{[j-i]} \bar{K}_i^{(k,j)} w_{j|i}\right)\right\| \\
&\leq \sum_{i=\tau}^{t-1} C_0 \rho^{t-1-i} \left\|w_i - \sum_{j=i}^{i+k-1} \lambda_i^{[j-i]} \bar{K}_i^{(k,j)} w_{j|i}\right\| \qquad (5.124\text{a}) \\
&\leq \sum_{i=\tau}^{t-1} C_0 \rho^{t-1-i} \left(\bar{w} + \sum_{j=i}^{i+k-1} C_0 \rho_0^{j-i} \bar{w}\right) \qquad (5.124\text{b}) \\
&\leq \frac{C_0(1 - \rho_0 + C_0)\bar{w}}{(1 - \rho_0)(1 - \rho)},
\end{aligned}$$

where we use the triangle inequality and (5.119) in (5.124a); we use the triangle inequality and (5.123) in (5.124b). This finishes the proof of ε-time-varying stability.

Chapter 6

REINFORCEMENT LEARNING IN NETWORKED SYSTEMS

The policy optimization setting that we study in Chapter 5 only involves a single agent. However, a practical challenge of policy optimization may arise when we apply it to a large-scale network formed by a team of agents: We cannot afford to gather global information for policy evaluation or gradient approximation. In this chapter, we demonstrate how such challenges can be overcome in a class of networked MARL problems with stochastic, nonlocal dependency structures. Key to our approach is the identification of a structural decay property, which says that the local Q-function of each agent i depends mainly on the states of the agents who are near i in the network. We leverage this decay property to design a scalable actor-critic algorithm with provable finite-time error bound. We discuss the core technical innovation underlying our theoretical analysis and test our algorithm in the applications of wireless networks and spreading networks.

The results in this chapter are based on the following paper:

[Lin, Qu, et al., 2021] Lin, Yiheng, Guannan Qu, Longbo Huang, and Adam Wierman. “Multi-agent reinforcement learning in stochastic networked systems.” Advances in Neural Information Processing Systems 34 (2021): 7825-7837.

6.1 Problem Setting

We consider a network of agents that are associated with an underlying undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, 2, \cdots, n\}$ denotes the set of agents and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ denotes the set of edges. The distance $d_{\mathcal{G}}(i, j)$ between two agents i and j is defined as the number of edges on the shortest path that connects them on graph $\mathcal{G}$. Each agent is associated with its local state $s_i \in \mathcal{S}_i$ and local action $a_i \in \mathcal{A}_i$ where $\mathcal{S}_i$ and $\mathcal{A}_i$ are finite sets. The global state/action is defined as the combination of all local states/actions, i.e., $s = (s_1, \cdots, s_n) \in \mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$, and $a = (a_1, \cdots, a_n) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. We use N_i^κ to denote the κ-hop neighborhood of agent i on $\mathcal{G}$, i.e., $N_i^\kappa := \{j \in \mathcal{N} \mid d_{\mathcal{G}}(i, j) \leq \kappa\}$. Let $f(\kappa) := \sup_i \left|N_i^\kappa\right|$. For a subset $M \subseteq \mathcal{N}$, let s_M/a_M denote the states/actions of agents in M.

Before we define the transitions and rewards, we first define the notion of active link sets, which are directed graphs on the agents $\mathcal{N}$ and they characterize the interaction structure among the agents. More specifically, an active link set is a set of directed edges that contains all self-loops, i.e., a subset of $\mathcal{N} \times \mathcal{N}$ and a super set of $\{(i,i) \mid i \in \mathcal{N}\}$. Generally speaking, $(j,i) \in L$ means agent j can affect agent i in the active link set L. Given an active link set L, we also use $N_i(L) := \{j \in \mathcal{N} \mid (j,i) \in L\}$ to denote the set of all agents (include itself) who can affect agent i in the active link set L. In this paper, we consider a pair of active link sets (L_t^s, L_t^r) that is independently drawn from some joint distribution $\mathcal{D}$ at each time step t,[1] where the distribution $\mathcal{D}$ will be defined using the underlying graph $\mathcal{G}$ later in Section 6.2. The role of L_t^s/L_t^r is that they define the dependence structure of state transition/reward at time t, which we detail below.

Transitions. At time t, given the current state, action $s(t), a(t)$ and the active link set L_t^s, the next individual state $s_i(t+1)$ is independently generated and only depends on the state/action of the agents in $N_i(L_t^s)$. In other words, we have,

$$P(s(t+1)|s(t), a(t), L_t^s) = \prod_{i \in \mathcal{N}} P_i(s_i(t+1)|s_{N_i(L_t^s)}(t), a_{N_i(L_t^s)}(t), L_t^s). \tag{6.1}$$

Rewards. Each agent is associated with a local reward function r_i. At time t, it is a function of L_t^r and the state/action of agents in $N_i(L_t^r)$: $r_i(L_t^r, s_{N_i(L_t^r)}(t), a_{N_i(L_t^r)}(t))$. The global reward $r(t)$ is defined to be the summation of the local rewards $r_i(t)$.

Policy. Each agent follows a localized policy that depends on its β-hop neighborhood, where $\beta \geq 0$ is a fixed integer. Specifically, at time step t, given the global state $s(t)$, agent i adopts a local policy ζ_i parameterized by θ_i to decide the distribution of $a_i(t)$ based on the the states of agents in N_i^β.

Our objective is for all the agents to *cooperatively* maximize the discounted global reward, i.e., $J(\theta) = \mathbb{E}_{s \sim \pi_0}\left[\sum_{t=0}^{\infty} \gamma^t r(s(t), a(t)) \mid s(0) = s\right]$, where π_0 is a given distribution on the initial global state, and we recall $r(s(t), a(t))$ is the global stage reward defined as the sum of all local rewards at time t.

Examples. To highlight the applicability of the general model, we include two examples of networked systems that feature the dependence structure captured by our model in Section 6.4: a wireless communication example and an example of controlling a process that spreads over a network.

Note that a limitation of our setting is that the dependence structure we consider is stationary, in the sense that dependencies are sampled i.i.d. from the distribution

[1] Here, correlations between L_t^s and L_t^r are possible.

$\mathcal{D}$. It is important to consider more general time-varying forms (e.g., Markovian) in future research.

Background. Before moving on, we review a few key concepts in RL which will be useful in the rest of the section. We use π_t^θ to denote the distribution of $s(t)$ under policy θ given that $s(0) \sim \pi_0$. A well-known result (Sutton et al., 1999) is that the gradient of the objective $\nabla J(\theta)$ can be computed by

$$\frac{1}{1-\gamma}\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} Q^\theta(s,a)\nabla \log \zeta^\theta(a \mid s), \tag{6.2}$$

where distribution $\pi^\theta(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \pi_t^\theta(s)$ is the *discounted state visitation distribution*. Evaluating the Q-function $Q^\theta(s,a)$ plays a key role in approximating $\nabla J(\theta)$. The local Q-function for agent i is the discounted local reward, i.e., $Q_i^\theta(s,a) = \mathbb{E}_{\zeta^\theta}\left[\sum_{t=0}^{\infty}\gamma^t r_i(t) \mid s(0) = s, a(0) = a\right]$, where we use $r_i(t)$ to denote the local reward of agent i at time step t. Using local Q-functions, we can decompose the global Q-function as $Q^\theta(s,a) = \frac{1}{n}\sum_{i=1}^{n} Q_i^\theta(s,a)$, which allows each node to evaluate its local Q-function separately.

A key challenge in our MARL setting is that directly estimating the Q-functions is not scalable since the size of the Q-functions is exponentially large in the number of agents. Therefore, in Section 6.2, we study structural properties of the Q-functions resulting from the dependence structure in the transition (6.1), which enables us to design a scalable RL algorithm in Section 6.3.

6.2 Decay Properties of Local Q Functions

One of the core challenges for MARL is that the size of the Q function is exponentially large in the number of agents. The key to our algorithm and its analysis is the identification of a novel structural decay property for the Q-function, which says that the local Q-function of each agent i is mainly decided by the states of the agents who are near i. This property is critical for the design of scalable algorithms because it enables the agents to reduce the dimension of the Q-function by truncating its dependence of the states and actions of far away agents. Recently, exponential decay has been shown to hold in networked MARL when the network is static (Qu, Wierman, and Li, 2020; Qu, Lin, et al., 2020), which is exploited to design a scalable RL algorithm. However, in stochastic network settings it is too much to hope for exponential decay in general (Easley, Kleinberg, et al., 2012), and so we introduce the more general notion of μ-decay here, where μ is a function that converges to 0 as κ tends to infinity. The case of exponential decay that has been studied previously corresponds to $\mu(\kappa) = \gamma^\kappa/(1-\gamma)$. The formal definition of μ-decay is given below, where for simplicity, we use $i \xrightarrow{L} j$ to denote $(i,j) \in L$ and denote $N_{-i}^\kappa := \mathcal{N} \setminus N_i^\kappa$.

Definition 6.2.1. *For a function $\mu : \mathbb{N} \to \mathbb{R}^+$ that satisfies $\lim_{\kappa \to +\infty} \mu(\kappa) = 0$, the μ-decay property holds if for any policy θ and any $i \in \mathcal{N}$, the local Q function Q_i^θ satisfies $\left|Q_i^\theta(s,a) - Q_i^\theta(s',a')\right| \leq \mu(\kappa)$ for any $(s,a), (s',a')$ that are identical within N_i^κ, i.e., $s_{N_i^\kappa} = s'_{N_i^\kappa}, a_{N_i^\kappa} = a'_{N_i^\kappa}$.*

Intuitively, if the μ-decay property holds and $\mu(\kappa)$ decays quickly as κ increases, we can approximately decompose the global Q function as $Q^\theta(s,a) = \frac{1}{n}\sum_{i=1}^n Q_i^\theta(s,a) \approx \frac{1}{n}\sum_{i=1}^n \hat{Q}_i^\theta(s_{N_i^\kappa}, a_{N_i^\kappa})$, where $\hat{Q}_i$ only depends on the states and actions within the κ-hop neighborhood of agent i. Before our work, Sunehag et al., 2018 empirically showed that such a value decomposition allows efficient training of MARL. Under the assumption that such decomposition exists, Sunehag et al., 2018 propose an approach to learn this decomposition. In contrast, as we prove in this section, the μ decay property holds provably and therefore, the global Q function can be directly decomposed in the networked MARL model and that the error of such decomposition is provably small.

Our first result is Theorem 6.2.1 which shows the relationship between the random active link sets and the μ-decay property. The proof of Theorem 6.2.1 is deferred to Section 6.A.

Theorem 6.2.1. *Define L^a as the static active link set that contains all pairs (i,j) whose graph distance on $\mathcal{G}$ is less than or equal to β, which is the dependency of local policy. Let random variable $X_i(\kappa)$ denote the smallest $t \in \mathbb{N}$ such that there exists a chain of agents*

$$j_0^a \xrightarrow{L_0^s} j_1^s \xrightarrow{L^a} j_1^a \xrightarrow{L_1^s} \cdots \xrightarrow{L_{t-1}^s} j_t^s \xrightarrow{L^a} j_t^a,$$

that satisfies $j_0^a \in N_{-i}^\kappa$ and $j_t^a \xrightarrow{L_t^r} i$. The μ-decay property holds for $\mu(\kappa) = \frac{1}{1-\gamma}\mathbb{E}\left[\gamma^{X_i(\kappa)}\right]$.

To make the μ-decay result more concrete, we provide several scenarios that yield different upper bounds on the term $\mathbb{E}\left[\gamma^{X_i(\kappa)}\right]$. In the first scenario, we study the case where long range links do not exist in Corollary 6.2.2. In this case, we obtain an exponential decay property that generalizes the result in Qu, Wierman, and Li, 2020. A proof is in Section 6.A.

Corollary 6.2.2 (Exponential Decay)**.** *Consider a distribution $\mathcal{D}$ of active link sets that satisfies*

$$P_{(L^s,L^r)\sim\mathcal{D}}\{(i,j)\in L^s\}=0,\ \textit{for all } i,j\in\mathcal{N} \textit{ s.t. } d_{\mathcal{G}}(i,j)\geq\alpha_1,$$
$$P_{(L^s,L^r)\sim\mathcal{D}}\{(i,j)\in L^r\}=0,\ \textit{for all } i,j\in\mathcal{N} \textit{ s.t. } d_{\mathcal{G}}(i,j)\geq\alpha_2.$$

Then, $\mathbb{E}\left[\gamma^{X_i(\kappa)}\right]\leq C\rho^\kappa$*, where* $\rho=\gamma^{1/(\alpha_1+\beta)}, C=\gamma^{-\alpha_2/(\alpha_1+\beta)}$.

In the second scenario, long range active links can occur, but with exponentially small probability with respect to their distance. In this case, we can obtain a near-exponential decay property where $\mu(\kappa)=O(\rho^{\kappa/\log\kappa}))$ for some $\rho\in(0,1)$. A proof can be found in Section 6.A.

Theorem 6.2.3 (Near-Exponential Decay)**.** *Suppose the distribution $\mathcal{D}$ of active link sets satisfies*

$$P_{(L^s,L^r)\sim\mathcal{D}}\{(i,j)\in L^s\cup L^r\}\leq c\lambda^{d_{\mathcal{G}}(i,j)},\ \textit{for all } i,j\in\mathcal{N},$$

where $c\geq 1, 1>\lambda>0$ *are constants. If the largest size of the* κ *neighborhood in the underlying graph* $\mathcal{G}$ *can be bounded by a polynomial of* κ*, i.e., there exists some constants* $c_0\geq 1, n_0\in\mathbb{N}$ *such that* $\left|\{j\in\mathcal{N}\mid d_{\mathcal{G}}(i,j)=\kappa\}\right|\leq c_0(\kappa+1)^{n_0}$ *holds for all i, then* $\mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]\leq C\rho^{\kappa/(1+\ln(\kappa+1))}$ *for some positive constant C and decay rate* $\rho<1$. [2]

It is interesting to compare the result above with models of the so-called “small world phenomena" in social networks, e.g., Easley, Kleinberg, et al., 2012. In these models, a link (i,j) occurs with probability $1/\text{poly}(d_{\mathcal{G}}(i,j))$, as opposed to the exponential dependence in Lemma 6.2.3. In this case, one can see function $\mu(\kappa)$ is lower bounded by $1/\text{poly}(\kappa)$, which leads us to conjecture that $\mu(\kappa)$ is also upper bounded by $O(1/\text{poly}(\kappa))$. Thus, when information spreads “slowly” it helps a localized algorithm to learn efficiently.

6.3 Learning with Localized Observations

Motivated by the μ-decay property of the Q-functions, we design a novel Scalable Actor Critic algorithm (Algorithm 10) for networked MARL problem, which exploits the μ-decay result in the previous section. The Critic part uses the local trajectory $\{(s_{N_i^\kappa}, a_{N_i^\kappa}, r_i)\}$ to evaluate the local Q-functions under parameter $\theta(m)$. Intuitively, the μ-decay property guarantees that we can achieve good approximation

[2] The explicit expression of C and ρ can be found in Section 6.A.

Algorithm 10: Scalable Actor Critic

```
for $m = 0, 1, 2, \cdots$ do
    Sample initial global state $s(0) \sim \pi_0$.
    Each node $i$ takes action $a_i(0) \sim \zeta_i^{\theta_i(m)}(\cdot \mid s_{N_i^\beta}(0))$ to obtain the global
      state $s(1)$.
    Each node $i$ records $s_{N_i^\kappa}(0), a_{N_i^\kappa}(0), r_i(0)$ and initialize $\hat{Q}_i^0$ to be all zero
      vector.
    for $t = 1, \cdots, T$ do
        Each node $i$ takes action $a_i(t) \sim \zeta_i^{\theta_i(m)}(\cdot \mid s_{N_i^\beta}(t))$ to obtain the global
          state $s(t+1)$.
        Each node $i$ update the local estimation $\hat{Q}_i$ with step size $\alpha_{t-1} = \frac{H}{t-1+t_0}$,
            $\hat{Q}_i^t\left(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)\right) = (1-\alpha_{t-1})\hat{Q}_i^{t-1}\left(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)\right)$
            $\qquad + \alpha_{t-1}\left(r_i(t) + \gamma\hat{Q}_i^{t-1}\left(s_{N_i^\kappa}(t), a_{N_i^\kappa}(t)\right)\right),$
          and for $\left(s_{N_i^\kappa}, a_{N_i^\kappa}\right) \neq \left(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)\right)$, let
            $\hat{Q}_i^t\left(s_{N_i^\kappa}, a_{N_i^\kappa}\right) = \hat{Q}_i^{t-1}\left(s_{N_i^\kappa}, a_{N_i^\kappa}\right).$
    end
    Each node $i$ approximate $\nabla_{\theta_i} J(\theta)$ by
        $\hat{g}_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} \hat{Q}_j^T\left(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t)\right)\nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}\left(a_i(t) \mid s_{N_i^\beta}(t)\right).$
    Each node $i$ conducts gradient ascent by $\theta_i(m+1) = \theta_i(m) + \eta_m \hat{g}_i(m)$.
end
```

error even when κ is not large. The Actor part computes the estimated partial derivative using the estimated local Q-functions, and uses the partial derivative to update local parameter θ_i. Compared with the Scalable Actor Critic algorithm proposed in Qu, Wierman, and Li, 2020, Algorithm 10 extends the policy dependency structure considered. No longer is the dependency completely local; it now extends to all agents within the β-hop neighborhood. Interestingly, the time-varying dependencies do not add complexity into the algorithm (though the analysis is more complex).

Algorithm 10 is highly scalable. Each agent i needs only to query and store the information within its κ-hop neighborhood during the learning process. The parameter κ can be set to balance accuracy and complexity. Specifically, as κ increases, the error bound becomes tighter at the expense of increasing computation, communication, and space complexity.

Error Bounds of the Critic Part

We first describe the assumption needed in our result. It focuses on the Markov chain formed by the global state-action pair (s, a) under a fixed policy parameter θ and is standard for finite-time convergence results in RL, e.g., Srikant and Ying, 2019; Brémaud, 2001; Qu and Wierman, 2020.

Assumption 6.3.1. *Under any fixed policy θ, $\{z(t) := (s(t), a(t))\}$ is an aperiodic and irreducible Markov chain on state space $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$ with a unique stationary distribution $d^\theta = (d^\theta_z, z \in \mathcal{Z})$, which satisfies $d^\theta_z > 0, \forall z \in \mathcal{Z}$. Define $d^\theta(z') = \sum_{z \in \mathcal{Z}: z_{N_i^\kappa} = z'} d^\theta(z)$ and $\sigma'(\kappa) := \inf_{z' \in \mathcal{Z}_{N_i^\kappa}} d^\theta(z')$. There exists positive constants K_1, K_2 such that $K_2 \geq 1$ and $\forall z' \in \mathcal{Z}, \forall t \geq 0$,*

$$\sup_{\mathcal{K} \subseteq \mathcal{Z}} \left| \sum_{z \in \mathcal{K}} d^\theta_z - \sum_{z \in \mathcal{K}} \mathbb{P}(z(t) = z \mid z(0) = z') \right| \leq K_1 e^{-t/K_2}.$$

We next analyze the Critic part of Algorithm 10 within a given outer loop iteration m. Since the policy is fixed in the inner loop, the global state/action pair (s, a) in the original MDP can be viewed as the state of a Markov chain. We observe that each local estimate $\hat{Q}_i^t \left(s_{N_i^\kappa}, a_{N_i^\kappa} \right)$ can be viewed as a form of state aggregation, where the global state (s, a) is "compressed" to $h(s, a) := (s_{N_i^\kappa}, a_{N_i^\kappa})$. Broadly speaking, the technique of state aggregation is one of the easiest-to-deploy schemes for state space compression (Jiang, 2018; Singh, Jaakkola, and Jordan, 1995), while its final performance relies heavily on whether the state aggregation map h only aggregates "similar" states. To have a good approximate equivalence, we need to find a good h, i.e., if two states are mapped to the same abstract state, their value functions are required to be close (to be discussed in Theorem 6.3.3). In the context of networked MARL, the μ decay property (Definition 6.2.1) provides a natural mapping for state aggregation $h(s, a) := (s_{N_i^\kappa}, a_{N_i^\kappa})$ which we defined earlier. This mapping h maps the global state/action to the local states/actions in agent i's κ-hop neighborhood and the μ-decay property guarantees that if $h(s, a) = h(s', a')$, the difference in their Q-functions is upper bounded by $\mu(\kappa)$, which is vanishing as κ increases. This shows that the mapping h we used is "good" in the sense it aggregates very similar global state-action pairs. This idea leads to the following theorem about the Critic part of Scalable Actor Critic (Algorithm 10).

Theorem 6.3.1. *Suppose Assumption 6.3.1 and μ-decay property (Definition 6.2.1) hold. Let the step size be $\alpha_t = \frac{H}{t+t_0}$ with $t_0 = \max(4H, 2K_2 \log T)$, and $H \geq \frac{2}{(1-\gamma)\sigma'(\kappa)}$. Define constant $C_b := 4K_1(1 + 2K_2 + 4H)$. Then, inside outer loop iteration m, for each $i \in \mathcal{N}$, with probability at least $1 - \delta$, we have*

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| Q_i^{\theta(m)}(s,a) - \hat{Q}_i^T(s_{N_i^\kappa}, a_{N_i^\kappa}) \right| \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C'_a}{T+t_0} + \frac{\mu(\kappa)}{1-\gamma},$$

where the constants are given by $C_a = \frac{40H}{(1-\gamma)^2}\sqrt{K_2 \log T \left(\log\left(\frac{4f(\kappa)K_2T}{\delta}\right) + \log\log T\right)}$ and $C'_a = \frac{8}{(1-\gamma)^2}\max\{\frac{144K_2H\log T}{\sigma'(\kappa)} + C_b, 2K_2\log T + t_0\}$.

The proof of Theorem 6.3.1 can be found in Section 6.B. The most related result in the literature to Theorem 6.3.1 is Theorem 7 in Qu, Wierman, and Li, 2020. In comparison, Theorem 6.3.1 applies for more general, potentially non-local, dependencies and, also, improves the constant term by a factor of $1/(1-\gamma)$.

Proof Idea: Stochastic Approximation and State Aggregation

In this section, we present the key technical innovation underlying our results on MARL in Theorem 6.3.1: a new finite-time analysis of a general asynchronous stochastic approximation (SA) scheme. The truncation enabled by μ-decay provides a form of state aggregation, which we analyze via a general SA scheme. Further, this SA scheme is of interest more broadly, e.g., to the settings of TD learning with state aggregation and asynchronous Q-learning with state aggregation (see Section 6.B).

Stochastic Approximation. Consider a finite-state Markov chain whose state space is given by $\mathcal{N} = \{1, 2, \cdots, n\}$. Let $\{i_t\}_{t=0}^{\infty}$ be the sequence of states visited by this Markov chain. Our focus is generalizing the following asynchronous stochastic approximation (SA) scheme, which is studied in Tsitsiklis, 1994; Shah and Xie, 2018; Wainwright, 2019: Let parameter $x \in \mathbb{R}^{\mathcal{N}}$, and $F : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$ be a γ-contraction in the infinity norm. The update rule of the SA scheme is given by

$$\begin{aligned} x_{i_t}(t+1) &= x_{i_t}(t) + \alpha_t \left(F_{i_t}(x(t)) - x_{i_t}(t) + w(t)\right), \\ x_j(t+1) &= x_j(t) \text{ for } j \neq i_t, j \in \mathcal{N}, \end{aligned} \tag{6.3}$$

where $w(t)$ is a noise sequence. It is shown in Qu and Wierman, 2020 that parameter $x(t)$ converges to the unique fixed point of F at the rate of $O\left(1/\sqrt{t}\right)$.

While general, in many cases, including networked MARL, we do not wish to calculate an entry for every state in $\mathcal{N}$ in parameter x, but instead, wish to calculate "aggregated entries." Specifically, at each time step, after i_t is generated, we use a surjection h to decide which dimension of parameter x should be updated. This technique, referred to as state aggregation, is one of the easiest-to-deploy schemes for state space compression in the RL literature (Jiang, 2018; Singh, Jaakkola, and Jordan, 1995). In the generalized SA scheme, our objective is to specify the convergence point as well as obtain a finite-time error bound.

Formally, to define the generalization of (6.3), let $\mathcal{N} = \{1, \cdots, n\}$ be the state space of $\{i_t\}$ and $\mathcal{M} = \{1, \cdots, m\}, (m \leq n)$ be the *abstract* state space. The surjection $h : \mathcal{N} \to \mathcal{M}$ is used to convert every state in $\mathcal{N}$ to its abstraction in $\mathcal{M}$. Given parameter $x \in \mathbb{R}^{\mathcal{M}}$ and function $F : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$, we consider the generalized SA scheme that updates $x(t) \in \mathbb{R}^{\mathcal{M}}$ starting from $x(0) = \mathbf{0}$,

$$\begin{aligned} x_{h(i_t)}(t+1) &= x_{h(i_t)}(t) + \alpha_t \left(F_{i_t}(\Phi x(t)) - x_{h(i_t)}(t) + w(t) \right), \\ x_j(t+1) &= x_j(t) \text{ for } j \neq h(i_t), j \in \mathcal{M}, \end{aligned} \tag{6.4}$$

where the feature matrix $\Phi \in \mathbb{R}^{\mathcal{N} \times \mathcal{M}}$ is defined as

$$\Phi_{ij} = \begin{cases} 1 & \text{if } h(i) = j \\ 0 & \text{otherwise} \end{cases}, \forall i \in \mathcal{N}, j \in \mathcal{M}. \tag{6.5}$$

In order to state our main result characterizing the convergence of (6.4), we must first state a few definitions and assumptions. To begin, we define the weighted infinity norm as in Qu and Wierman, 2020, except that we extend its definition so as to define the contraction of function F. The reason we use the weighted infinity norm as opposed to the standard infinity norm is that its generality can be used in certain settings for undiscounted RL, as shown in Tsitsiklis, 1994; Bertsekas, 2007.

Definition 6.3.1 (Weighted Infinity Norm)**.** *Fix a positive vector $v \in \mathbb{R}^{\mathcal{M}}$. For $x \in \mathbb{R}^{\mathcal{M}}$, we define $\|x\|_v := \sup_{i \in \mathcal{M}} \frac{|x_i|}{v_i}$. For $x \in \mathbb{R}^{\mathcal{N}}$, we define $\|x\|_v := \sup_{i \in \mathcal{N}} \frac{|x_i|}{v_{h(i)}}$.*

Next, we state our assumption on the mixing rate of the Markov chain $\{i_t\}$, which is common in the literature (Tsitsiklis and Van Roy, 1996; Srikant and Ying, 2019). It holds for any finite-state Markov chain which is aperiodic and irreducible (Brémaud, 2001).

Assumption 6.3.2 (Stationary Distribution and Geometric Mixing Rate)**.** *$\{i_t\}$ is an aperiodic and irreducible Markov chain on state space $\mathcal{N}$ with stationary distribution $d = (d_1, d_2, \cdots, d_n)$. Let $d'_j = \sum_{i \in h^{-1}(j)} d_i$ and $\sigma' = \inf_{j \in \mathcal{M}} d'_j$. There exists positive constants K_1, K_2 which satisfy that*

$$\sup_{\mathcal{S}\subseteq\mathcal{N}}\left|\sum_{i\in\mathcal{S}} d_i - \sum_{i\in\mathcal{S}}\mathbb{P}(i_t = i \mid i_0 = j)\right| \leq K_1 \exp(-t/K_2), \forall j \in \mathcal{N},$$

for all $t \geq 0$ and $K_2 \geq 1$.

Our next assumption ensures contraction of F. It is also standard, e.g., Tsitsiklis, 1994; Wainwright, 2019; Qu and Wierman, 2020, and ensures that F has a unique fixed point y^*.

Assumption 6.3.3 (Contraction)**.** *Operator F is a γ contraction in $\|\cdot\|_v$, i.e., for any $x, y \in \mathbb{R}^{\mathcal{N}}$, we have $\|F(x) - F(y)\|_v \leq \gamma\|x - y\|_v$. Further, there exists some constant $C > 0$ such that for any $x \in \mathbb{R}^{\mathcal{N}}$, we have $\|F(x)\|_v \leq \gamma\|x\|_v + C$.*

In Assumption 6.3.3, notice that the first sentence directly implies the second with $C = (1+\gamma)\|y^*\|_v$, where $y^* \in \mathbb{R}^{\mathcal{N}}$ is the unique fixed point of F. Further, while Assumption 6.3.3 implies that F has a unique fixed point y^*, we do not expect our stochastic approximation scheme to converge to it. Instead, we show that the convergence is to the unique x^* that solves

$$\Pi F(\Phi x^*) = x^*, \text{ where } \Pi := \left(\Phi^\top D \Phi\right)^{-1}\Phi^\top D. \tag{6.6}$$

Here $D = diag(d_1, d_2, \cdots, d_n)$ denotes the steady-state probabilities for the process $\{i_t\}$. Note that x^* is well-defined because the operator $\Pi F(\Phi\cdot)$, which defines a mapping from $\mathbb{R}^{\mathcal{M}}$ to $\mathbb{R}^{\mathcal{M}}$, is also a contraction in $\|\cdot\|_v$. We state and prove this as Proposition 6.B.2 in Section 6.B.

Our last assumption is on the noise sequence $w(t)$. It is also standard, e.g., Shah and Xie (2018) and Qu and Wierman (2020).

Assumption 6.3.4 (Martingale Difference Sequence)**.** *w_t is $\mathcal{F}_{t+1}$ measurable and satisfies $\mathbb{E}w(t) \mid \mathcal{F}_t = 0$. Further, $|w(t)| \leq \bar{w}$ almost surely for constant $\bar{w}$.*

We are now ready to state our finite-time convergence result for stochastic approximation.

Theorem 6.3.2. *Suppose Assumptions 6.3.2, 6.3.3, 6.3.4 hold. Further, assume there exists constant $\bar{x} \geq \|x^*\|_v$ such that $\forall t$, $\|x(t)\|_v \leq \bar{x}$ almost surely.*[3] *Let the*

[3] The assumption on $\bar{x}$ follows from Assumptions 6.3.3 and 6.3.4. See Proposition 6.B.1 in Section 6.B.

step size be $\alpha_t = \frac{H}{t+t_0}$ *with* $t_0 = \max(4H, 2K_2 \log T)$*, and* $H \geq \frac{2}{\sigma'(1-\gamma)}$*. Let* x^* *be the unique solution of equation* $\Pi F(\Phi x^*) = x^*$*, and define constants* $C_1 := 2\bar{x} + C + \frac{\bar{w}}{\underline{v}}, C_2 := 4\bar{x} + 2C + \frac{\bar{w}}{\underline{v}}, C_3 := 2K_1(2\bar{x} + C)(1 + 2K_2 + 4H)$*. Then, with probability at least* $1 - \delta$*,*

$$\|x(T) - x^*\|_v \leq \frac{C_a}{\sqrt{T + t_0}} + \frac{C'_a}{T + t_0} = \tilde{O}\left(\frac{1}{\sqrt{T}}\right),$$

where the constants are given by $C_a = \frac{4HC_2}{1-\gamma}\sqrt{K_2 \log T \left(\log\left(\frac{4mK_2T}{\delta}\right) + \log\log T\right)}$ *and* $C'_a = 4\max\{\frac{48K_2C_1H\log T + \sigma' C_3}{(1-\gamma)\sigma'}, \frac{2\bar{x}(2K_2\log T + t_0)}{1-\gamma}\}$.

A proof of Theorem 6.3.2 can be found in Section 6.B. Compared with Theorem 4 in Qu and Wierman, 2020, Theorem 6.3.2 holds for a more general SA scheme where state aggregation is used to reduce the dimension of the parameter x. The proof technique used in Qu and Wierman, 2020 does not apply to our setting because our stationary point x^* has a more complex form (6.5). To do the generalization, we need to use a different error decomposition method compared to Qu and Wierman, 2020 that leverages the stationary distribution D rather than the distribution of i_t condition on $i_{t-\tau}$ (see Section 6.B for details). Because of this generality, Theorem 6.3.2 requires a stronger but standard assumption on the mixing rate of the Markov chain $\{i_t\}$.

State Aggregation. To illustrate the impact of our analysis of SA (Theorem 6.3.2) beyond the network setting, we study a simpler application to the cases of TD-learning and Q-learning with state aggregation in this section. Understanding state aggregation methods is a foundational goal of analysis in the RL literature and it has been studied in many previous works, e.g., Li, Walsh, and Littman, 2006; Jong and Stone, 2005; Jiang, Kulesza, and Singh, 2015; Dann et al., 2018; Singh, Jaakkola, and Jordan, 1995. Further, the result is extremely useful in the analysis in networked MARL that follows since the μ-decay property we introduce (Definition 6.2.1) provides a natural state aggregation in the network setting (see Corollary 6.3.1). Due to space constraints, in this section we only introduce the results on TD-learning; the results on Q-learning are given in Section 6.B.

In TD learning with state aggregation Singh, Jaakkola, and Jordan, 1995; Tsitsiklis and Van Roy, 1997, given the sequence of states visited by the Markov chain is $\{i_t\}$, the update rule of TD(0) is given by

$$\begin{aligned} \theta_{h(i_t)}(t+1) &= \theta_{h(i_t)}(t) + \alpha_t \left(r_t + \gamma \theta_{h(i_{t+1})}(t) - \theta_{h(i_t)}(t)\right), \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(i_t), j \in \mathcal{M}, \end{aligned} \tag{6.7}$$

where $h : \mathcal{N} \to \mathcal{M}$ is a surjection that maps each state in $\mathcal{N}$ to an abstract state in $\mathcal{M}$ and r_t is the reward at time step t such that $\mathbb{E}[r_t] = r(i_t, i_{t+1})$.

Taking F as the Bellman Policy Operator, i.e., the i'th dimension of function F is given by

$$F_i(V) = \mathbb{E}_{i' \sim \mathbb{P}(\cdot|i)} \left[r(i, i') + \gamma V_{i'} \right], \forall i \in \mathcal{N}, V \in \mathbb{R}^{\mathcal{N}}.$$

The value function (vector) V^* is defined as $V_i^* = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(i_t, i_{t+1}) \mid i_0 = i\right], i \in \mathcal{N}$ (Tsitsiklis and Van Roy, 1997). By defining the feature matrix Φ as (6.5) and the noise sequence as

$$w(t) = r_t + \gamma \theta_{h(i_{t+1})}(t) - \mathbb{E}_{i' \sim \mathbb{P}(\cdot|i_t)}[r(i_t, i') + \gamma \theta_{h(i')}(t)],$$

we can rewrite the update rule of TD(0) in (6.7) in the form of an SA scheme (6.4). Therefore, we can apply Theorem 6.3.2 to obtain a finite-time error bound for TD learning with state aggregation. A proof of Theorem 6.3.3 can be found in Section 6.B.

Theorem 6.3.3. *Let Assumption 6.3.2 hold for the Markov chain $\{i_t\}$ and let the stage reward r_t be upper bounded by $\bar{r}$ almost surely. Assume that if $h(i) = h(i')$ for $i, i' \in \mathcal{N}$, we have $\left|V_i^* - V_{i'}^*\right| \le \zeta$ for a constant ζ. Consider TD(0) with the step size $\alpha_t = \frac{H}{t+t_0}$, where $t_0 = \max(4H, 2K_2 \log T)$ and $H \ge \frac{2}{\sigma'(1-\gamma)}$. Define constant $C_4 := 4K_1(1 + 2K_2 + 4H)$. Then, with probability at least $1 - \delta$,*

$$\|\Phi \cdot \theta(T) - V^*\|_\infty \le \frac{C_a}{\sqrt{T + t_0}} + \frac{C_a'}{T + t_0} + \frac{\zeta}{1 - \gamma},$$

where the constants are given by $C_a = \frac{40H\bar{r}}{(1-\gamma)^2}\sqrt{K_2 \log T \left(\log\left(\frac{4mK_2T}{\delta}\right) + \log\log T\right)}$ and $C_a' = \frac{8\bar{r}}{(1-\gamma)^2} \max\{\frac{144K_2H\log T}{\sigma'} + C_4, 2K_2 \log T + t_0\}$.

The most related prior results to Theorem 6.3.3 are Srikant and Ying, 2019; Bhandari, Russo, and Singal, 2018. In contrast to these, Theorem 6.3.3 considers the infinity norm, which is more natural for measuring error when using state aggregation. Further, our analysis is different and extends to the case of Q-learning with state aggregation (see Section 6.B), where we obtain the first finite-time error bound. Moreover, unlike Bhandari, Russo, and Singal, 2018, our TD-learning algorithm does not require a projection step.

6.4 Applications: Wireless and Spreading Networks

Wireless Networks

We consider a wireless network with multiple access points setting shown in Fig. 6.1, where a set of user nodes in a wireless network, denoted by $U = \{u_1, u_2, \cdots, u_n\}$, share a set of access points $Y = \{y_1, y_2, \cdots, y_m\}$ (Zocca, 2019). Each access point y_i is associated with a probability p_i of successful transmission. Each user node u_i only has access to a subset $Y_i \subseteq Y$ of the access points. Typically, this available set is determined by each user node's physical connections to the access points. To apply the networked MARL model, we identify the set of user nodes U as the set of agents $\mathcal{N}$. The underlying graph $G = (\mathcal{N}, \mathcal{E})$ is defined as the conflict graph, i.e., edge $(u_i, u_j) \in \mathcal{E}$ if and only if $Y_i \cap Y_j \neq \emptyset$.

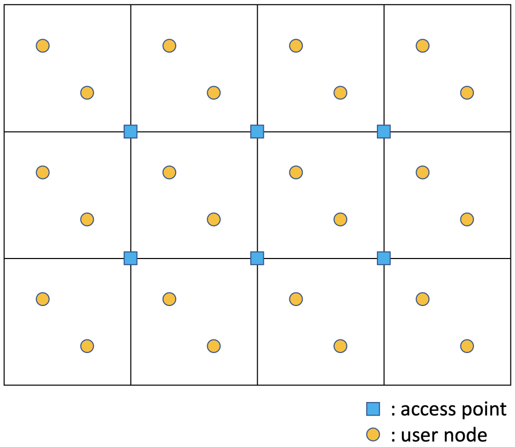

Figure 6.1: An example setup of wireless networks. Each user node can send packets to the access points at the corners of its grid.

At each time step t, each user u_i receives a packet with initial life span d with probability q. Each user maintains a queue to cache the packets it receives. At each time step, if the packet is successfully sent to an access point, it will be removed from the queue. Otherwise, its life span will decrease by 1. A packet is discarded from the queue immediately if its remaining life span is 0. At each time step t, a user node u_i can choose to send one of the packets in its queue to one of the access point $y_{i,t} \in Y_i$. If no other user node sends packets to access point $y_{i,t}$ at time step t, the packet from user i can be delivered successfully with probability p_i. Otherwise, the sending action will fail. A user u_i receives a local reward of $r_{i,t} = 1$ immediately after successfully sending a packet at time step t, and receives $r_{i,t} = 0$ otherwise. Our objective is to find a policy that maximizes the global discounted reward under a discounted factor $0 \leq \gamma < 1$:

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t r_{i,t}\right].$$

To see how this setting fits into our model, we first define the local state/action and specify the parameters. Since each packet has a life span of d, and each user node

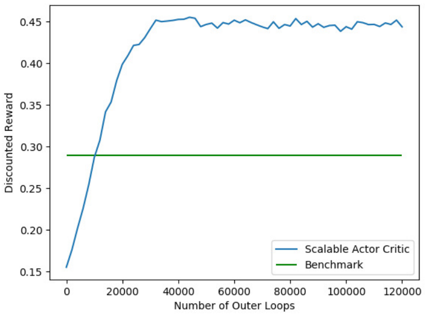

Figure 6.2: Discounted reward in the training process. 5 × 5 grid, 1 user per grid.

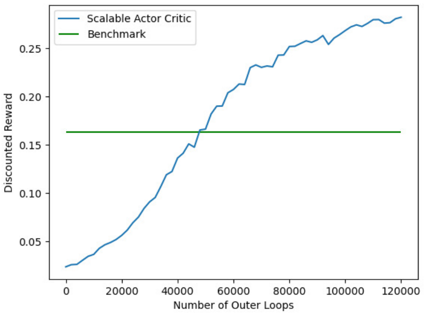

Figure 6.3: Discounted reward in the training process. 3 × 4 grid, 2 users per grid.

receives at most one packet at a time step, we use a d-tuple $s_i = (e_1, e_2, \cdots, e_d) \in \mathcal{S}_i := \{0,1\}^d$ to denote the local state of user node i. Specifically, e_j indicates whether user node u_i has a packet with remaining life span j in its queue. A local action of user node u_i is 2-tuple (l, y), which means sending the packet with remaining life span $l \in \{1, 2, \cdots, d\}$ to an access point $y \in Y_i$. Note that we define an empty action that does nothing at all. If a user node performs an action (l, y) when there is no packet with life span l in its queue, we view this as an empty action. This setting falls into the category we studied in Corollary 6.2.2, where long range links do not exist. Specifically, in this setting, the next local state of user node u_i depends on the current local states/actions in its 1-hop neighborhood ($\alpha_1 = 1$ in Corollary 6.2.2). We assume each user node can choose its action only based on its current local state ($\beta = 0$). Due to potential collisions, the local reward of user u_i also depends on the states/actions in its 1-hop neighborhood ($\alpha_2 = 1$ in Corollary 6.2.2). Though this is a static setting, note that the results of Qu, Wierman, and Li, 2020 do not apply.

The detailed setting we use is as follows. We consider the setting where the user nodes are located in $h \times w$ grids (see Fig. 6.1). There are c user nodes in each grid, and each user can send packets to an access point on the corner of its grid. We set the initial life span $d = 2$, the arrival probability $q = 0.5$, and the discounted factor $\gamma = 0.7$. The successful transmission probability p_i for each access point y_i is sampled uniformly randomly from $[0, 1]$. We run the Scalable Actor Critic algorithm with parameter $\kappa = 1$ to learn a localized stochastic policy in two cases $(h, w, c) = (5, 5, 1)$ (see Fig. 6.2) and $(h, w, c) = (3, 4, 2)$ (see Fig. 6.3). For comparison, we use a benchmark based on the localized ALOHA protocol Roberts,

1975. Specifically, the benchmark policy works as following: At time step t, each user node u_i takes the empty action with a certain probability p'; otherwise, it sends the packet with the minimum remaining life span to a random access point in Y_i, with the probability proportional to the successful transmission probability of this access point and inverse proportional to the number of users sharing this access point. In Fig. 6.2 and Fig. 6.3, we have tuned the parameter p' to find the one with the highest discounted reward.

As shown in Fig. 6.2 and Fig. 6.3, starting from the initial policy that chooses an local action uniformly at random, the Scalable Actor Critic algorithm with parameter $\kappa = 1$ can learn a policy that performs better than the benchmark. As a remark, the benchmark policy requires the set $\{p_i\}_{1\leq i\leq m}$, the probability of successful transmission, as input. Moreover, in the benchmark policy, the probability of performing an empty action also needs to be tuned manually. In contrast, the Scalable Actor Critic algorithm can learn a better policy without these specific inputs by interacting with the system.

Spreading Networks

We consider a spreading network with n agents and an underlying graph $\mathcal{G}$. See Fig. 6.4 for an illustration of $n = wh$ agents on a $w \times h$ grid network. For each agent i, the local state/action space is given by $\mathcal{S}_i = \{0, 1\}$ and $\mathcal{A}_i = \{0, 1\}$. To make the discussion more concrete, in the following we present the spreading network model in the context of SIS epidemic network. This version of the SIS model has been studied in, for example, Ruhi, Thrampoulidis, and Hassibi, 2016. Our setting is more general and can be generalized to other types of spreading networks like opinion networks, social networks, etc. At time step t, the local state $s_i(t) = 0$ means agent i is "susceptible," while the local state $s_i(t) = 1$ means the agent i is "infected." By taking action $a_i(t) = 1$, agent i can suppress its infection probability at the expense of incurring an action cost. In the meantime, agent i will incur an infection cost if $s_i(t) = 1$. The interaction among agents is modeled by a set of undirected links, where two agents can affect each other if they are connected by a link. To model the influence of physical distance on the pattern of social contact, we assume the short range links occur more frequently than long range links. An illustration of the spreading network is shown in Fig. 6.4 (a), where the black nodes denote the agents with state 1; the white nodes denote the agents with state 0; the blue edges denote the set of active links at some time step.

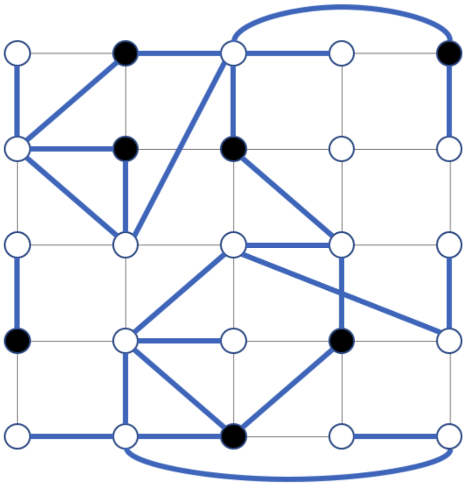

Figure 6.4: An illustration of the spreading network with 25 agents on a 5 × 5 grid network. The black nodes denote "infected" agents; The white nodes denote "susceptible" agents; The blue edges denote the active links at some time step.

Mathematically, the model can be described as follows. At each time step t, each agent i can decide her/his local action $a_i(t)$ based on the information of local states in the 1-hop neighborhood N_i^1, i.e., $\beta = 1$. The local reward $r_i(t)$ is a function of the local state $s_i(t)$ and the local action $a_i(t)$, i.e., L_t^r is static and only contains self loops. Specifically, we define

$$r_i(t) = -c_i^{(a)}\mathbf{1}(a_i(t) = 1) - c_i^{(s)}\mathbf{1}(s_i(t) = 1),$$

where $\left(c_i^{(s)}, c_i^{(a)}\right)$ are parameters associated with agent i and can be different among agents. As mentioned earlier, $c_i^{(s)}$ penalizes the agent for being "infected," while $c_i^{(a)}$ is the cost of taking epidemic control measure. The stage reward is the sum of these two costs.

To describe the state transition rule, we first define the way the active link set L_t^s is generated: independently for each pair of agents $(i, j) \in \mathcal{N} \times \mathcal{N}$ with $i \neq j$, with probability $2^{-d_{\mathcal{G}}(i,j)}$, we include edges (i, j) and (j, i) in the set L_t^s; otherwise, neither edge is included in the set, i.e., $(i, j), (j, i) \notin L_t^s$. Given L_t^s, the next local state $s_i(t+1)$ is sampled from a distribution that depends on the local states in $N_i(L_t^s)$. Specifically, define the quantities

$$n_i(t) = \left|\{j \mid j \in N_i(L_t) \setminus \{i\}, s_j(t) = 1, a_j(t) = 0\}\right|,$$
$$m_i(t) = \left|\{j \mid j \in N_i(L_t) \setminus \{i\}, s_j(t) = 1, a_j(t) = 1\}\right|.$$

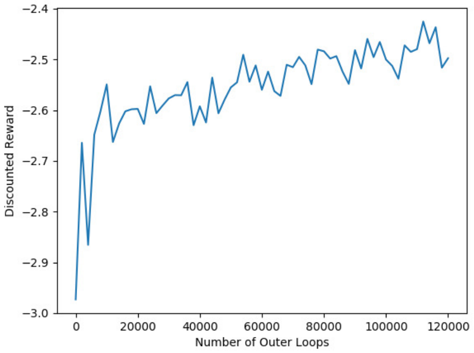

Figure 6.5: Discounted reward in the training process. 5×5 grid.

Then, with different local state and action pairs, the probability that the next local state $s_i(t+1) = 0$ is given by

$$\begin{aligned} & P(s_i(t+1) = 0 \mid s_{N_i(L_t)}, a_{N_i(L_t)}) \\ = & \begin{cases} p_i^{(r)} & \text{if } s_i(t) = 1; \\ \left(1 - p_i^{(h)}\right)^{n_i(t)} \left(1 - p_i^{(m)}\right)^{m_i(t)} & \text{if } s_i(t) = 0, a_i(t) = 1; \\ \left(1 - p_i^{(m)}\right)^{n_i(t)} \left(1 - p_i^{(l)}\right)^{m_i(t)} & \text{if } s_i(t) = 0, a_i(t) = 0, \end{cases} \end{aligned}$$

where $\left(p_i^{(r)}, p_i^{(h)}, p_i^{(m)}, p_i^{(l)}\right)$ are parameters associated with agent i and can be different among agents. Due to control actions, we assume $p_i^{(h)} > p_i^{(m)} > p_i^{(l)}$. This provides the transition rule, and the underlying intuition is that the local state of agent i turns from "infected" ($s_i(t) = 1$) to "susceptible" ($s_i(t+1) = 0$) with a fixed recovering probability $p_i^{(r)}$; the probability that agent i turns from "susceptible" ($s_i(t) = 0$) to "infected" ($s_i(t+1) = 1$) depends on the number of neighboring agents in the active link set that are already infected, and further, whether agent i or the nearby agents j take epidemic control measures ($a_i(t) = 1, a_j(t) = 1$) or not. Roughly speaking, the more nearby infected agents, the more likely agent i will become infected; however, if epidemic control measures are taken by agent i and nearby agents in $N_i(L_t^s)$, the probability of agent i getting infected will be smaller.

We run the Scalable Actor Critic algorithm with parameter $\kappa = 1$ to learn a localized stochastic policy in the case $(h, w) = (5, 5)$ (Fig. 6.5). For each agent i, parameters $\left(c_i^{(s)}, c_i^{(a)}, p_i^{(r)}, p_i^{(h)}\right)$ are sampled independently from the distribution

$$c_i^{(s)} \sim U[1.0, 3.0], c_i^{(a)} \sim U[0.01, 0.20], p_i^{(r)} \sim U[0.1, 0.5], p_i^{(h)} \sim U[0.5, 0.9],$$

and we set $p_i^{(m)} = p_i^{(h)}/4, p_i^{(l)} = p_i^{(m)}/4$. At time step 0, for each $i \in \mathcal{N}$, we initialize local state $s_i(0)$ to be 1 with probability 0.3.

6.A Proof of Decay Properties

Proof of Theorem 6.2.1

For ease of exposition, let A, B be two subsets of the agent set $\mathcal{N}$ and we use $A \xrightarrow{\tau} B$ to denote the event that there exists a chain

$$j_0^a \xrightarrow{L_0^s} j_1^s \xrightarrow{L^a} j_1^a \xrightarrow{L_1^s} \cdots \xrightarrow{L_{\tau-1}^s} j_\tau^s \xrightarrow{L^a} j_\tau^a,$$

whose head and tail satisfies $j_0^a \in A$ and $j_\tau^a \in B$.

Given a sequence of active link sets $\{L_t^s\}_{t=0}^\infty$ and under fixed global policy θ, we say the information at set $A \subseteq \mathcal{N}$ spread to another set $B \subseteq \mathcal{N}$ in τ time steps (denoted by $I(A) \xrightarrow{\tau} I(B)$) if there exists (s, a) and (s', a') such that $(s_{\mathcal{N}\setminus A}, a_{\mathcal{N}\setminus A}) = (s'_{\mathcal{N}\setminus A}, a'_{\mathcal{N}\setminus A})$ and the distribution of $(s_B(\tau), a_B(\tau))$ given $(s(0), a(0)) = (s, a)$ is different with that given $(s(0), a(0)) = (s', a')$.

We show by induction that $I(A) \xrightarrow{\tau} I(B)$ happens only if $A \xrightarrow{\tau} B$ happens.

If $\tau = 0$, since $I(A) \xrightarrow{0} I(B)$, we see that $A \cap B \neq \emptyset$. Therefore, we can let j_0^a be any agent in $A \cap B$. Hence we also have $A \xrightarrow{0} B$.

Suppose the statement holds for $\tau = t$. When $\tau = t+1$, suppose that $I(A) \xrightarrow{t+1} I(B)$. Define sets

$$B' := \{j \in \mathcal{N} \mid \exists k \in B, s.t. j \xrightarrow{L^a} k\}, B'' := \{j \in \mathcal{N} \mid \exists k \in B', s.t. j \xrightarrow{L_t^s} k\}.$$

Notice that $B \subseteq B' \subseteq B''$. By the definition of transition probability and policy dependence, we know that the distribution of $a_B(t+1)$ is decided by $s_{B'}(t+1)$, and the distribution of $s_{B'}(t+1)$ is decided by $(s_{B''}(t), a_{B''}(t))$. Therefore, we must have $I(A) \xrightarrow{t} I(B'')$. By the induction hypothesis, we have $A \xrightarrow{t} B''$, which further implies $A \xrightarrow{t+1} B$. This finishes the induction.

Given a sequence of active link sets $\{(L_t^s, L_t^r)\}$, we use $\pi_{t,i}$ to denote the distribution of

$\left(s_{N_i(L_t^r)}(t), a_{N_i(L_t^r)}(t)\right)$ given that $(s(0), a(0)) = (s, a)$; we use $\pi'_{t,i}$ to denote the distribution of $\left(s_{N_i(L_t^r)}(t), a_{N_i(L_t^r)}(t)\right)$ given that $(s(0), a(0)) = (s', a')$. We notice that $\pi_{t,i} \neq \pi'_{t,i}$ happens only if $I(N_{-i}^\kappa) \xrightarrow{t} I(N_i(L_t^r))$, which is true only if $N_{-i}^\kappa \xrightarrow{t} N_i(L_t^r)$. Recall that $X_i(\kappa)$ is defined as the smallest t such that $N_{-i}^\kappa \xrightarrow{t} N_i(L_t^r)$ holds. Hence, we obtain that

$$\begin{aligned}
&\left|Q_i^\theta(s,a) - Q_i^\theta(s',a')\right| \\
&\le \mathbb{E}_{\{(L_t^s, L_t^r)\}} \sum_{t=0}^{\infty} \left|\gamma^t \mathbb{E}_{\pi_{t,i}} r_i(s_{N_i(L_t^r)}, a_{N_i(L_t^r)}) - \gamma^t \mathbb{E}_{\pi'_{t,i}} r_i(s_{N_i(L_t^r)}, a_{N_i(L_t^r)})\right| \\
&\le \mathbb{E}_{\{(L_t^s, L_t^r)\}} \sum_{t=X_i(\kappa)}^{\infty} \left|\gamma^t \mathbb{E}_{\pi_{t,i}} r_i(s_{N_i(L_t^r)}, a_{N_i(L_t^r)}) - \gamma^t \mathbb{E}_{\pi'_{t,i}} r_i(s_{N_i(L_t^r)}, a_{N_i(L_t^r)})\right| \\
&\le \frac{1}{1-\gamma} \mathbb{E}\left[\gamma^{X_i(\kappa)}\right],
\end{aligned}$$

where we use the definition of $X_i(\kappa)$ in the second step.

Proof of Corollary 6.2.2

Given a sequence of active link sets $\{(L_t^s, L_t^r)\}$, let $t = X_i(\kappa)$. By the definition of $X_i(\kappa)$, we assume that a chain of agents

$$j_0^a \xrightarrow{L_0^s} j_1^s \xrightarrow{L^a} j_1^a \xrightarrow{L_1^s} \cdots \xrightarrow{L_{t-1}^s} j_t^s \xrightarrow{L^a} j_t^a$$

satisfies $j_0^a \in N_{-i}^\kappa$ and $j_t^a \xrightarrow{L_t^r} i$.

By the triangle inequality and the assumptions of Lemma 6.2.2, we obtain that

$$\begin{aligned}
d_{\mathcal{G}}(j_0^a, i) &\le \sum_{\tau=0}^{t-1} \left(d_{\mathcal{G}}(j_\tau^a, j_{\tau+1}^s) + d_{\mathcal{G}}(j_{\tau+1}^s, j_{\tau+1}^a)\right) + d_{\mathcal{G}}(j_t^a, i) \\
&\le t(\beta + \alpha_1) + \alpha_2.
\end{aligned}$$

Therefore, we see that t is lower bounded by $\frac{\kappa-\alpha_2}{\beta+\alpha_1}$, which also gives a lower bound of $X_i(\kappa)$.

Proof of Theorem 6.2.3

To simplify notation, we adopt the same notations as in the proof of Theorem 6.2.1. Specifically, recall that we use $A \xrightarrow{\tau} B$ to denote the event that there exists a chain

$$j_0^a \xrightarrow{L_0^s} j_1^s \xrightarrow{L^a} j_1^a \xrightarrow{L_1^s} \cdots \xrightarrow{L_{\tau-1}^s} j_\tau^s \xrightarrow{L^a} j_\tau^a,$$

whose head and tail satisfies $j_0^a \in A$ and $j_\tau^a \in B$. We will use ∂N_i^κ to denote the set of neighbors whose distance to i is κ, i.e., $\partial N_i^\kappa := \{j \in \mathcal{N} \mid d_{\mathcal{G}}(i,j) = \kappa\} = N_i^\kappa \setminus N_i^{\kappa-1}$. Define $a_\kappa := \mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]$. Define function cat (concatenation) such that for a pair of active link sets (L^s, L^a), $(x,y) \in cat(L^s, L^a)$ if and only if $\exists z \in \mathcal{N}$ such that $x \xrightarrow{L^s} z \xrightarrow{L^a} y$.

Before proving Theorem 6.2.3, we first give an upper bound for the sum of an infinite sequence $\{poly(k+i)\cdot \nu^i\}_{i\in\mathbb{N}}$, where $\nu < 1$ is a positive constant. This result is helpful for showing an upper bound of $P(N^{\kappa}_{-i} \to N^{j}_{i})$.

Lemma 6.A.1. *If $m \in \mathbb{N}^*$ and $0 < \nu < 1$ are constants, for all $k \geq \frac{2m}{\ln(1/\nu)}$, we have*

$$\sum_{i=0}^{\infty}(k+i)^m \nu^i \leq \frac{1}{1-\sqrt{\nu}}\cdot k^m.$$

Proof of Lemma 6.A.1. Define function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+$ as

$$f(t) = (k+t)^m \cdot \nu^{t/2}.$$

The derivative of function f is given by

$$f'(t) = (k+t)^{m-1}\cdot \nu^{t/2}\left(m + \frac{1}{2}\ln\nu\cdot(k+t)\right).$$

Since $k \geq \frac{2m}{\ln(1/\nu)}$, $f'(t) \leq 0$ holds for all $t \geq 0$, hence we have $f(t) \leq f(0) = k^m$. Therefore, we obtain that

$$\begin{aligned}\sum_{i=0}^{\infty}(k+i)^m \nu^i &\leq \sum_{i=0}^{\infty} f(i)\cdot \nu^{i/2}\\ &\leq k^m \sum_{i=0}^{\infty}\nu^{i/2}\\ &\leq \frac{1}{1-\sqrt{\nu}}\cdot k^m.\end{aligned}$$

□

Now we come back to the proof of Theorem 6.2.3.

By union bound, we derive an upper bound of the probability that a link (x,y) is in $cat(L^s, L^a)$. Suppose $d \in \mathbb{N}$ is constant that satisfies $d_{\mathcal{G}}(x,y) \geq d$, and the probability P is taken over $(L^s, L^r) \sim \mathcal{D}$:

$$\begin{aligned}P\left((x,y)\in cat(L^s,L^a)\right) &= P\left(\exists z \in \mathcal{N}, (x,z)\in L^s \wedge (z,y)\in L^a\right)\\ &\leq \sum_{z:d_{\mathcal{G}}(z,y)\leq\beta} P\left((x,z)\in L^s\right)\\ &\leq c_0(\beta+1)^{n_0+1}\cdot c\lambda^{d-\beta}\\ &= c_g\lambda^d, \end{aligned} \tag{6.8}$$

where recall that λ the decay factor defined by the assumption of Theorem 6.2.3, and constant c_g is defined as $c_0 c(\beta+1)^{n_0+1}\lambda^{-\beta}$.

By the assumption on the size of κ-hop neighborhood, we know that for some constant c_0 and $n_0 \in \mathbb{N}^*$, $\left|\partial N_i^{\kappa}\right| \le c_0(\kappa+1)^{n_0}$ holds for all $\kappa \ge 1$. Let $n_1 := 2n_0$. With the help of Lemma 6.A.1, we show that for some constant $c_2 > 0$, $P\left(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\right)$ is upper bounded by $c_2(\kappa+1)^{n_1}\lambda^{\kappa-j}$ for all $j \le \kappa - 1$ when $\kappa \ge \frac{2n_0}{\ln(1/\lambda)}$:

$$P\left(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\right) \le P\left(\exists x \in N_{-i}^{\kappa-1}, y \in \partial N_i^j \text{ s.t. } (x,y) \in cat(L^s, L^a)\right) \tag{6.9a}$$

$$\le \sum_{q=0}^{\infty} P\left(\exists x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j \text{ s.t. } (x,y) \in cat(L^s, L^a)\right) \tag{6.9b}$$

$$\le \sum_{q=0}^{\infty} \sum_{x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j} P\left((x,y) \in cat(L^s, L^a)\right) \tag{6.9c}$$

$$\le \sum_{q=0}^{\infty} \sum_{x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j} c_g \lambda^{(\kappa+q-j)} \tag{6.9d}$$

$$\le c_g \lambda^{\kappa-j} \sum_{q=0}^{\infty} \left|\partial N_i^{\kappa+q}\right| \cdot \left|\partial N_i^j\right| \cdot \lambda^q$$

$$\le c_g c_0^2 (\kappa+1)^{n_0} \lambda^{\kappa-j} \sum_{q=0}^{\infty} (\kappa+q+1)^{n_0}\lambda^q \tag{6.9e}$$

$$\le c_2(\kappa+1)^{n_1}\lambda^{\kappa-j}, \tag{6.9f}$$

where we use the definition of $N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j$ in (6.9a); we use union bound in (6.9b) and (6.9c); we use the fact that $d_{\mathcal{G}}(x,y) \ge \kappa+q-j, \forall x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j$ and (6.8) in (6.9d); we use the bounds $\left|\partial N_i^j\right| \le c_0 j^{n_0} \le c_0 \kappa^{n_0}$ and $\left|\partial N_i^{\kappa+q}\right| \le c_0(\kappa+q)^{n_0}$ in (6.9e); we define $c_2 := \frac{c_g c_0^2}{1-\sqrt{\lambda}}$ and use Lemma 6.A.1 in (6.9f).

Let constants c_3 and q be defined as

$$c_3 := \frac{1}{2}\sqrt[4]{\lambda}(1-\sqrt{\lambda})\left(\frac{1}{\sqrt{\gamma}}-1\right),$$

$$q := \frac{1}{\ln(1/\lambda)} \max\{(\ln c_2 - \ln c_3 - 2\ln(1-\sqrt{\gamma})), (2n_1+4)\},$$

and define function $p(\kappa) := [q(1+\ln(\kappa+1))] + 1$. We can find $\kappa_0 \in \mathbb{Z}^+$ such that $p(\kappa) \ge \kappa$ for all $\kappa \le \kappa_0$, and $p(\kappa) > \kappa$ for all $\kappa > \kappa_0$.

Let ρ be a constant such that $1 > \rho > \max\{\gamma^{1/(2q)}, \sqrt[4]{\lambda}\}$. Let $C := \rho^{-\max\{q+1, \frac{2n_0}{\ln(1/\lambda)}\}}$. Recall that we define $a_\kappa := \mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]$, where $X_i(\kappa - 1)$ denotes the smallest t such that $N_{-i}^{\kappa-1} \xrightarrow{t} N_i(L_t^r)$ holds. Now we show by induction that

$$a_\kappa \leq C\rho^{\kappa/(1+\ln(\kappa+1))}, \forall \kappa \geq 1. \tag{6.10}$$

Since $a_\kappa \leq 1$, (6.10) clearly holds when $\kappa \leq \kappa_0$. To see this, recall that we have $\kappa \leq p(\kappa)$ and $C \geq \rho^{-(q+1)}$ by definition, thus the right hand side of (6.10) can be lower bounded by

$$C\rho^{\kappa/(1+\ln(\kappa+1))} \geq \rho^{-(q+1)} \cdot \rho^{p(\kappa)/(1+\ln(\kappa+1))} \geq \rho^{-(q+1)} \cdot \rho^{q+1} = 1.$$

When $\kappa > \kappa_0$, we have $\kappa > p(\kappa)$. Recall that $a_\kappa := \mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]$. Notice that $X_i(\kappa - 1) = 0$ if and only if $N_{-i}^{\kappa-1} \cap N_i(L_0^r) \neq \emptyset$. To simplify the notation, we denote the event $N_{-i}^{\kappa-1} \cap N_i(L_0^r) \neq \emptyset$ by E_0. Using this and the idea of dynamic programming, we see that

$$\begin{aligned}
a_\kappa \leq \gamma \Bigg(& P\{\left(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\right) \wedge \neg E_0\} a_\kappa \\
& + \sum_{j=0}^{\kappa-1} P\{\left(N_{-i}^{\kappa} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa} \xrightarrow{1} N_i^{j-1}\right) \wedge \neg E_0\} a_j \Bigg) + P(E_0) \\
\leq \gamma \Bigg(& P\{\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\} a_\kappa \\
& + \sum_{j=0}^{\kappa-1} P\{\left(N_{-i}^{\kappa} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa} \xrightarrow{1} N_i^{j-1}\right)\} a_j \Bigg) + P(E_0),
\end{aligned} \tag{6.11}$$

where the probability P are taken over $(L_0^s, L_0^r) \sim D$.

Since $\kappa \geq p(\kappa) \geq q \geq \frac{2n_1}{\ln(1/\lambda)} \geq \frac{2n_0}{\ln(1/\lambda)}$, by Lemma 6.A.1, we see that

$$\begin{aligned}
P(E_0) = P\{\exists j \in N_{-i}^{\kappa-1} \text{ s.t. } (j,i) \in L^r\} &\leq \sum_{q=0}^{\infty} cc_0(\kappa + q + 1)^{n_0} \lambda^{\kappa+q} \\
&\leq \frac{cc_0}{1 - \sqrt{\lambda}} (\kappa + 1)^{n_0+1} \lambda^\kappa.
\end{aligned}$$

Substituting this into (6.11) and rearranging the terms gives

$$\begin{aligned}
& \left(1 - \gamma P\{\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\}\right) a_\kappa \\
\leq \gamma & \sum_{j=\kappa-p(\kappa)+1}^{\kappa-1} P\{\left(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{j-1}\right)\} a_j
\end{aligned}$$

$$\begin{aligned}&+\gamma\sum_{j=0}^{\kappa-p(\kappa)} P\{\left(N_{-i}^{\kappa-1}\xrightarrow{1}\partial N_i^j\right)\wedge\left(\neg N_{-i}^{\kappa-1}\xrightarrow{1}N_i^{j-1}\right)\}a_j\\&+\frac{cc_0}{1-\sqrt{\lambda}}(\kappa+1)^{n_0+1}\lambda^{\kappa}.\end{aligned}\tag{6.12}$$

For simplicity, we define $\rho_\kappa := \rho^{1/(1+\ln(\kappa+1))}$. By the induction assumption, we have that

$$a_j \le C\rho^{j/(\ln(j+1)+1)} \le C\rho^{j/(\ln(\kappa+1)+1)} = C\rho_\kappa^j.$$

Substituting this into (6.12) gives that

$$\begin{aligned}&\left(1-\gamma P\{\neg N_{-i}^{\kappa-1}\xrightarrow{1}N_i^{\kappa-1}\}\right)a_\kappa\\ \le{}& C\gamma\sum_{j=\kappa-p(\kappa)+1}^{\kappa-1} P\{\left(N_{-i}^{\kappa-1}\xrightarrow{1}\partial N_i^j\right)\wedge\left(\neg N_{-i}^{\kappa-1}\xrightarrow{1}N_i^{j-1}\right)\}\rho_\kappa^j\\&+C\gamma\sum_{j=0}^{\kappa-p(\kappa)} P\{\left(N_{-i}^{\kappa-1}\xrightarrow{1}\partial N_i^j\right)\wedge\left(\neg N_{-i}^{\kappa-1}\xrightarrow{1}N_i^{j-1}\right)\}\rho_\kappa^j\\&+\frac{c_0}{1-\sqrt{\lambda}}(\kappa+1)^{n_0+1}\lambda^{\kappa}.\end{aligned}\tag{6.13}$$

By the definition of $p(\kappa)$ and q, we see that

$$\begin{aligned}\lambda^{-p(\kappa)} &\ge \lambda^{-q(1+\ln(\kappa+1))} = \lambda^{-q}\cdot(\kappa+1)^{q\ln(1/\lambda)} \ge \frac{c_2}{c_3(1-\sqrt{\gamma})^2}\cdot(\kappa+1)^{n_1}\\&\ge \frac{c_2}{c_3(1-\gamma)}\cdot(\kappa+1)^{n_1}.\end{aligned}$$

Therefore, we obtain the upper bound

$$\begin{aligned}P\{\left(N_{-i}^{\kappa-1}\xrightarrow{1}\partial N_i^j\right)\wedge\left(\neg N_{-i}^{\kappa}\xrightarrow{1}N_i^{j-1}\right)\} &\le P\{N_{-i}^{\kappa-1}\xrightarrow{1}\partial N_i^j\}\\&\le c_2(\kappa+1)^{n_1}\lambda^{(\kappa-j)}\\&\le (1-\gamma)c_3\lambda^{(\kappa-p(\kappa)-j)}.\end{aligned}$$

Using this and divide both sides of (6.13) by $\left(1-\gamma P\{\neg N_{-i}^{\kappa}\xrightarrow{1}N_i^{\kappa-1}\}\right)$, we see that

$$\begin{aligned}a_\kappa \le{}& \gamma\left(C\rho_\kappa^{\kappa-p(\kappa)+1}+Cc_3(\rho_\kappa^{\kappa-p(\kappa)}+\lambda^1\cdot\rho_\kappa^{\kappa-p(\kappa)-1}+\lambda^2\cdot\rho_\kappa^{\kappa-p(\kappa)-2}+\cdots)\right)\\&+\frac{c_0}{(1-\gamma)(1-\sqrt{\lambda})}(\kappa+1)^{n_0+1}\lambda^{\kappa},\end{aligned}\tag{6.14}$$

where we also use the fact that

$$\sum_{j=\kappa-p+1}^{\kappa-1} P\{\left(N_{-i}^{\kappa-1}\xrightarrow{1}\partial N_i^j\right)\wedge\left(\neg N_{-i}^{\kappa-1}\xrightarrow{1}N_i^{j-1}\right)\} \le 1-\gamma P\{\neg N_{-i}^{\kappa-1}\xrightarrow{1}N_i^{\kappa-1}\}.$$

By the definition of $p(\kappa), q$ and c_2, we see that the following inequalities about the exponent of λ holds:

$$\lambda^{\frac{\kappa}{4}} \leq \lambda^{\frac{p(\kappa)}{4}} \leq (\kappa+1)^{-\frac{q\ln(1/\lambda)}{4}} \leq (\kappa+1)^{-n_0-1}$$

and

$$\lambda^{\frac{\kappa}{2}} \leq \lambda^{\frac{p(\kappa)}{2}} \leq \lambda^{\frac{q}{2}} \leq \frac{(1-\sqrt{\gamma})(1-\gamma)(1-\sqrt{\lambda})}{2c_0},$$

which implies

$$\lambda^{\frac{3\kappa}{4}} \leq \frac{(1-\sqrt{\gamma})(1-\gamma)(1-\sqrt{\lambda})}{2c_0(\kappa+1)^{n_0+1}}. \tag{6.15}$$

Dividing both sides of (6.14) by $C\rho_\kappa^\kappa$ gives that

$$\begin{aligned}
\frac{a_\kappa}{C\rho_\kappa^\kappa} &\leq \gamma\left(\frac{1}{\rho_\kappa^{p(\kappa)-1}} + \frac{c_3}{\rho_\kappa^{p(\kappa)}}\cdot\frac{1}{1-(\lambda/\rho_\kappa)}\right) + \frac{c_0}{(1-\gamma)(1-\sqrt{\lambda})}(\kappa+1)^{n_0+1}\lambda^{\frac{3\kappa}{4}} && \text{(6.16a)}\\
&\leq \gamma\left(\frac{1}{\rho^q} + \frac{1}{\rho^{q+1}}\cdot\frac{c_3}{1-\sqrt{\lambda}}\right) + \frac{1}{2}(1-\sqrt{\gamma}) && \text{(6.16b)}\\
&= \frac{\gamma}{\rho^q}\left(1 + \frac{c_3}{\rho(1-\sqrt{\lambda})}\right) + \frac{1}{2}(1-\sqrt{\gamma})\\
&\leq \sqrt{\gamma}\cdot\frac{1}{2}\left(1+\frac{1}{\sqrt{\gamma}}\right) + \frac{1}{2}(1-\sqrt{\gamma}) && \text{(6.16c)}\\
&= 1,
\end{aligned}$$

where we use $\rho_\kappa = \rho^{1/(1+\ln\kappa)} \geq \rho \geq \sqrt[4]{\lambda}$ in (6.16a); we use $\rho_\kappa \geq \sqrt[4]{\lambda}$, $p = [q(1+\ln\kappa)] + 1$, and (6.15) in (6.16b); we use $c_3 = \sqrt{\lambda}(1-\sqrt{\lambda})(\sqrt{\gamma}-1) \leq \rho(1-\sqrt{\lambda})(\sqrt{\gamma}-1)$ and $\rho \geq \gamma^{1/(2q)}$ in (6.16c).

6.B Proof of Learning with Localized Observations

Proposition 6.B.1. *Suppose Assumptions 6.3.3 and 6.3.4 hold. Then for all t,*

$$\|x(t)\|_v \leq \frac{1}{1-\gamma}\left((1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right)$$

holds almost surely, where $y^* \in \mathbb{R}^{\mathcal{N}}$ is the stationary point of F.

Proof of Proposition 6.B.1. By Assumption 6.3.3, we have that for all $x \in \mathbb{R}^{\mathcal{M}}$,

$$\begin{aligned}
\|F(\Phi x)\|_v &\leq \|F(\Phi x) - F(y^*)\|_v + \|F(y^*)\|_v && \text{(6.17a)}\\
&\leq \gamma\|\Phi x - y^*\|_v + \|y^*\|_v && \text{(6.17b)}
\end{aligned}$$

$$\leq \gamma \|x\|_v + (1+\gamma)\|y^*\|_v, \tag{6.17c}$$

where we use the triangle inequality in (6.17a) and (6.17c); we use Assumption 6.3.3 in (6.17b).

Let $\bar{x} = \frac{1}{1-\gamma}\left((1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right)$. We prove $\|x(t)\|_v \leq \bar{x}$ by induction on t. Since we initialize $x(0)$ to be $\mathbf{0}$, the statement is true for $t = 0$.

Suppose the statement is true for t. By the update rule of x, we see that

$$\begin{aligned}\frac{1}{v_{h(i_t)}}\left|x_{h(i_t)}(t+1)\right| &\leq (1-\alpha_t)\frac{1}{v_{h(i_t)}}\left|x_{h(i_t)}(t)\right| + \alpha_t\left(\frac{1}{v_{h(i_t)}}\left|F_{i_t}(\Phi x(t))\right| + \frac{1}{v_{h(i_t)}}|w(t)|\right)\\ &\leq (1-\alpha_t)\|x(t)\|_v + \alpha_t\left(\|F(\Phi x(t))\|_v + \frac{\bar{w}}{\underline{v}}\right) \qquad (6.18a)\\ &\leq (1-\alpha_t)\|x(t)\|_v + \alpha_t\left(\gamma\|x(t)\|_v + (1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right) \qquad (6.18b)\\ &\leq (1-\alpha_t)\bar{x} + \alpha_t\left(\gamma\bar{x} + (1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right) \qquad (6.18c)\\ &= \bar{x},\end{aligned}$$

where we use Assumption 6.3.4 in (6.18a); (6.17) in (6.18b); the induction assumption in (6.18c).

For $j \neq h(i_t), j \in \mathcal{M}$, we have that

$$\frac{1}{v_j}\left|x_j(t+1)\right| = \frac{1}{v_j}\left|x_j(t)\right| \leq \|x(t)\|_v \leq \bar{x}. \tag{6.19}$$

Combining (6.18) and (6.19), we see that the statement also holds for $t+1$. Hence we have showed $\|x(t)\|_v \leq \bar{x}$ by induction. □

Proof of Theorem 6.3.1

In the Critic part of Algorithm 10, since the policy is fixed to be $\theta(m)$, the pair (s, a) can be viewed as the state of a Markov chain C, and $Q^{\theta(m)}(s, a)$ in the original MDP corresponds to the value function $V^*((s, a))$ on C. Define the state aggregation map h such that $h((s, a)) = (s_{N_i^\kappa}, a_{N_i^\kappa})$. By the μ-decay property, we see that if $h((s, a)) = h((s', a'))$, then

$$|V^*((s,a)) - V^*((s',a'))| = \left|Q^{\theta(m)}(s,a) - Q^{\theta(m)}(s',a')\right| \leq \mu(\kappa).$$

Note that Assumption 6.3.1 implies that Assumption 6.3.2 holds for C. Thus, we can apply Theorem 6.3.3 to finish the proof of Theorem 6.3.1.

Contraction of the Update Operator

To show that the equation $\Pi F(\Phi x) = x$ has a unique solution x^*, by the Banach–Caccioppoli fixed-point theorem, it suffices to show that operator $\Pi F(\Phi \cdot)$ is a γ-contraction in $\|\cdot\|_v$.

Proposition 6.B.2. *If Assumption 6.3.3 holds, operator* $\Pi F(\Phi \cdot)$ *is a contraction in* $\|\cdot\|_v$*, i.e., for any* $x, y \in \mathbb{R}^{\mathcal{M}}$*,* $\|\Pi F(\Phi x) - \Pi F(\Phi y)\|_v \leq \gamma \|x - y\|_v$.

To prove this proposition, we first show both operator Π and operator Φ are non-expansive in $\|\cdot\|_v$ before combining them with F.

Proof of Proposition 6.B.2. We first show that operator Π is non-expansive in $\|\cdot\|_v$, i.e., for any $x, y \in \mathbb{R}^{\mathcal{N}}$, we have

$$\|\Pi x - \Pi y\|_v \leq \|x - y\|_v. \tag{6.20}$$

Since Π is a linear operator, it suffices to show that for any $x \in \mathbb{R}^{\mathcal{N}}$, $\|\Pi x\|_v \leq \|x\|_v$.

Recall that $\forall j \in \mathcal{M}, h^{-1}(j) := \{i \in \mathcal{N} \mid h(i) = j\}$. Using this notation, the j th element of vector Πx is given by

$$(\Pi x)_j = \frac{1}{\sum_{i \in h^{-1}(j)} d_i} \left(\Phi^\top D x\right)_j = \frac{1}{\sum_{i \in h^{-1}(j)} d_i} \cdot \sum_{i \in h^{-1}(j)} d_i x_i.$$

Hence we see that

$$\frac{\left|(\Pi x)_j\right|}{v_j} \leq \frac{1}{\sum_{i \in h^{-1}(j)} d_i} \cdot \sum_{i \in h^{-1}(j)} d_i \frac{|x_i|}{v_j} \leq \sup_{i \in h^{-1}(j)} \frac{|x_i|}{v_j}. \tag{6.21}$$

By taking $\sup_j$ on both sides of (6.21), we see that

$$\|\Pi x\|_v = \sup_{j \in \mathcal{M}} \frac{\left|(\Pi x)_j\right|}{v_j} \leq \sup_{j \in \mathcal{M}} \sup_{i \in h^{-1}(j)} \frac{|x_i|}{v_j} = \sup_{i \in \mathcal{N}} \frac{|x_i|}{v_{h(i)}} = \|x\|_v, \tag{6.22}$$

where we use the definition of $\|\cdot\|_v$ on $\mathbb{R}^{\mathcal{N}}$ in the last equation. Hence we have shown that Π is non-expansive in $\|\cdot\|_v$ (inequality (6.20)).

We can also show that for any $x, y \in \mathbb{R}^{\mathcal{M}}$, we have

$$\|\Phi x - \Phi y\|_v = \|x - y\|_v. \tag{6.23}$$

Since Φ is a linear operator, we only need to show that for any $x \in \mathbb{R}^{\mathcal{M}}$, $\|\Phi x\|_v = \|x\|_v$.

Since $(\Phi x)_i = x_{h(i)}, \forall i \in \mathcal{N}$, by the definition of $\|\cdot\|_v$ on $\mathbb{R}^{\mathcal{N}}$, we see that the norm remains unchanged after applying Φ:

$$\|\Phi x\|_v = \sup_{i\in\mathcal{N}} \frac{|(\Phi x)_i|}{v_{h(i)}} = \sup_{i\in\mathcal{N}} \frac{\left|x_{h(i)}\right|}{v_{h(i)}} = \sup_{j\in\mathcal{M}} \frac{\left|x_j\right|}{v_j} = \|x\|_v.$$

Hence we have shown that Φ is non-expansive in $\|\cdot\|_v$ (equation (6.23)).

Therefore, for any $x, y \in \mathbb{R}^{\mathcal{M}}$, we have

$$\begin{aligned} \|\Pi F(\Phi x) - \Pi F(\Phi y)\|_v &\le \|F(\Phi x) - F(\Phi y)\|_v && (6.24a)\\ &\le \gamma\|\Phi x - \Phi y\|_v && (6.24b)\\ &= \gamma\|x - y\|_v, && (6.24c)\end{aligned}$$

where we use (6.20) in (6.24a); Assumption 6.3.3 in (6.24b); (6.23) in (6.24c). □

Proof of Theorem 6.3.2

The proof approach of Theorem 6.3.2 is similar to the proof of Theorem 4 in Qu and Wierman, 2020. Specifically, we show an upper bound for $\|x(t) - x^*\|_v$ by induction on time step t. To do so, we divide the whole proof into three steps: In Step 1, we manipulate the update rule (6.4) so that it can be written in a recursive form of sequence $\|x(t) - x^*\|_v$ (see Lemma 6.B.1); In Step 2, we bound the effect of noise terms in the recursive form we obtained in Step 1; In Step 3, we combine the first two steps to finish the induction.

For simplicity of notation, we use e_i to denote the indicator vector in $\mathbb{R}^n$, i.e., the i th entry is 1 and all other entries are 0. We also use ξ_i to denote the indicator vector in $\mathbb{R}^m$.

One of the main proof techniques used in Qu and Wierman, 2020 is to consider $D_t = \mathbb{E}e_{i_t}e_{i_t}^\top \mid \mathcal{F}_{t-\tau}$, which is the distribution of i_t condition on $\mathcal{F}_{t-\tau}$, in the coefficients of the recursive relationship of sequence $\|x(t) - x^*\|_v$. However, this approach does not work in the more general setting we consider because x^* may not be the stationary point of operator $(\Phi^\top D_t\Phi)^{-1}\phi^\top D_t F(\Phi\cdot)$. As a result, we cannot decompose $\|x(t) - x^*\|_v$ recursively if we use D_t in the coefficients. To overcome this difficulty, we use $D = diag(d_1, \cdots, d_n)$, which is the stationary distribution of i_t, in the coefficients of the recursive relationship (Lemma 6.B.1).

Now we begin the technical part of our proof.

Step 1: Decomposition of Error. Let $D_t = \mathbb{E}e_{i_t}e_{i_t}^\top \mid \mathcal{F}_{t-\tau}$, where τ is a parameter that we will tune later. Then D_t is a $\mathcal{F}_{t-\tau}$-measurable n-by-n diagonal random matrix, with its i'th entry being $d_{t,i} = \mathbb{P}(i_t = i \mid \mathcal{F}_{t-\tau})$. Recall that $D = diag(d_1, \cdots, d_n)$, where d is the stationary distribution of the Markov Chain $\{i_t\}$.

Notice that for all $i \in \mathcal{N}$, we have $\xi_{h(i)} = \Phi^\top e_i$. We can rewrite the update rule as

$$\begin{aligned}
x(t+1) &= x(t) + \alpha_t [e_{i_t}^\top F(\Phi x(t)) - \xi_{h(i_t)}^\top x(t) + w(t)]\xi_{h(i_t)} \\
&= x(t) + \alpha_t [\xi_{h(i_t)} e_{i_t}^\top F(\Phi x(t)) - \xi_{h(i_t)} \xi_{h(i_t)}^\top x(t) + w(t)\xi_{h(i_t)}] \\
&= x(t) + \alpha_t \Phi^\top \left[e_{i_t} e_{i_t}^\top \left(F(\Phi x(t)) - \Phi x(t) \right) + w(t) e_{i_t} \right] \qquad (6.25a) \\
&= x(t) + \alpha_t \left[\Phi^\top D F(\Phi x(t)) - \Phi^\top D \Phi x(t) \right] \\
&\quad + \alpha_t \Phi^\top \left[(e_{i_t} e_{i_t}^\top - D) \left(F(\Phi x(t)) - \Phi x(t) \right) + w(t) e_{i_t} \right] \\
&= x(t) + \alpha_t \left[\Phi^\top D F(\Phi x(t)) - \Phi^\top D \Phi x(t) \right] \\
&\quad + \alpha_t \Phi^\top \left[(e_{i_t} e_{i_t}^\top - D) \left(F(\Phi x(t-\tau)) - \Phi x(t-\tau) \right) + w(t) e_{i_t} \right] \\
&\quad + \alpha_t \Phi^\top (e_{i_t} e_{i_t}^\top - D) \left[F(\Phi x(t)) - F(\Phi x(t-\tau)) - \Phi(x(t) - x(t-\tau)) \right] \\
&= (I - \alpha_t \Phi^\top D \Phi) x(t) + \alpha_t \Phi^\top D F(\Phi x(t)) + \alpha_t (\epsilon(t) + \psi(t)), \qquad (6.25b)
\end{aligned}$$

where in (6.25a), we use $\xi_{h(i_t)} = \Phi^\top e_{i_t}$. Additionally, in (6.25b), we define

$$\epsilon(t) = \Phi^\top \left[(e_{i_t} e_{i_t}^\top - D) \left(F(\Phi x(t-\tau)) - \Phi x(t-\tau) \right) + w(t) e_{i_t} \right]$$

and

$$\psi(t) = \Phi^\top (e_{i_t} e_{i_t}^\top - D) \left[F(\Phi x(t)) - F(\Phi x(t-\tau)) - \Phi(x(t) - x(t-\tau)) \right].$$

We further decompose $\epsilon(t)$ as $\epsilon(t) = \epsilon_1(t) + \epsilon_2(t)$, where $\epsilon_1(t)$ and $\epsilon_2(t)$ are defined as

$$\epsilon_1(t) = \Phi^\top \left[(e_{i_t} e_{i_t}^\top - D_t) \left(F(\Phi x(t-\tau)) - \Phi x(t-\tau) \right) + w(t) e_{i_t} \right]$$

and

$$\epsilon_2(t) = \Phi^\top (D_t - D) \left(F(\Phi x(t-\tau)) - \Phi x(t-\tau) \right).$$

We see that condition on $\mathcal{F}_{t-\tau}$, the expected value of $\epsilon_1(t)$ is zero, i.e.,

$$\begin{aligned}
&\mathbb{E}\epsilon_1(t) \mid \mathcal{F}_{t-\tau} \\
&= \Phi^\top \mathbb{E} \left[(e_{i_t} e_{i_t}^\top - D_t) \mid \mathcal{F}_{t-\tau} \right] \left[F(\Phi x(t-\tau)) - \Phi x(t-\tau) \right] \\
&\quad + \Phi^\top \mathbb{E} \left[\mathbb{E}[w(t) \mid \mathcal{F}_t] e_{i_t} \mid \mathcal{F}_{t-\tau} \right] \\
&= 0.
\end{aligned}$$

Recall that matrix Π is defined as

$$\Pi = \left(\Phi^\top D \Phi \right)^{-1} \Phi^\top D.$$

By expanding (6.25) recursively, we obtain that

$$\begin{aligned}
x(t+1) = &\prod_{k=\tau}^{t} \left(I - \alpha_k \Phi^\top D\Phi\right) x(\tau) + \sum_{k=\tau}^{t} \alpha_k \left(\prod_{l=k+1}^{t} (I - \alpha_l \Phi^\top D\Phi)\right) \Phi^\top DF(\Phi x(k)) \\
&+ \sum_{k=\tau}^{t} \alpha_k \left(\prod_{l=k+1}^{t} (I - \alpha_l \Phi^\top D\Phi)\right) (\epsilon(k) + \psi(k)) \\
= &\tilde{B}_{\tau-1,t} x(\tau) + \sum_{k=\tau}^{t} B_{k,t} \Pi F(\Phi x(k)) + \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} (\epsilon(k) + \psi(k)), \qquad (6.26)
\end{aligned}$$

where $B_{k,t} = \alpha_k \left(\Phi^\top D\Phi\right) \prod_{l=k+1}^{t} (I - \alpha_l \Phi^\top D\Phi)$ and $\tilde{B}_{k,t} = \prod_{l=k+1}^{t} \left(I - \alpha_l \Phi^\top D\Phi\right)$.

For simplicity of notation, we define $D' = \Phi^\top D\Phi \in \mathbb{R}^{\mathcal{M}\times\mathcal{M}}$. Notice that D' is a diagonal matrix in $\mathbb{R}^{\mathcal{M}\times\mathcal{M}}$ with the j'th entry $d'_j = \sum_{j\in h^{-1}(i)} d_i$. Clearly, $B_{k,t}$ and $\tilde{B}_{k,t}$ are m-by-m diagonal matrices, with the i'th diagonal entry given by $b_{k,t,i}$ and $\tilde{b}_{k,t,i}$, where $b_{k,t,i} = \alpha_k d'_i \prod_{l=k+1}^{t} (1 - \alpha_l d'_i)$ and $\tilde{b}_{k,t,i} = \prod_{l=k+1}^{t} (1 - \alpha_l d'_i)$. Therefore, for any $i \in \mathcal{M}$, we have

$$\tilde{b}_{\tau-1,t,i} + \sum_{k=\tau}^{t} b_{k,t,i} = 1. \qquad (6.27)$$

Also, by the definition of σ', we have that for any i, almost surely

$$b_{k,t,i} \le \beta_{k,t} := \alpha_k \prod_{l=k+1}^{t} (1 - \alpha_l \sigma'), \tilde{b}_{k,t,i} \le \tilde{\beta}_{k,t} = \prod_{l=k+1}^{t} (1 - \alpha_l \sigma'),$$

where $\sigma' = \min\{d'_1, \cdots, d'_m\}$.

Recall that x^* is the unique solution of the equation $\Pi F(\Phi x^*) = x^*$. Lemma 6.B.1 shows that we can expand the error term $\|x(t) - x^*\|_v$ recursively.

Lemma 6.B.1. *Let $\Upsilon_t = \|x(t) - x^*\|_v$, we have almost surely,*

$$\Upsilon_{t+1} \le \tilde{\beta}_{\tau-1,t} \Upsilon_\tau + \gamma \sup_{i\in\mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i} \Upsilon_k + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon(k) \right\|_v + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \psi(k) \right\|_v.$$

Proof of Lemma 6.B.1. By (6.26) and the triangle inequality of $\|\cdot\|_v$, we have

$$\begin{aligned}
&\|x(t+1) - x^*\|_v \\
\le &\sup_{i\in\mathcal{M}} \frac{1}{v_i} \left| \tilde{b}_{\tau-1,t,i} x_i(\tau) + \sum_{k=\tau}^{t} b_{k,t,i} \left(\Pi F(\Phi x(k))\right)_i - x_i^* \right| \\
&+ \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon(k) \right\|_v + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \psi(k) \right\|_v. \qquad (6.28)
\end{aligned}$$

We also see that for each $i \in \mathcal{M}$,

$$\begin{aligned}
&\frac{1}{v_i}\left|\tilde{b}_{\tau-1,t,i}x_i(\tau) + \sum_{k=\tau}^{t} b_{k,t,i}\left(\Pi F(\Phi x(k))\right)_i - x_i^*\right| \\
&\le \tilde{b}_{\tau-1,t,i}\frac{1}{v_i}\left|x_i(\tau) - x_i^*\right| + \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{v_i}\left|(\Pi F(\Phi x(k)))_i - x_i^*\right| \qquad (6.29a) \\
&\le \tilde{b}_{\tau-1,t,i}\|x(\tau) - x^*\|_v + \sum_{k=\tau}^{t} b_{k,t,i}\|(\Pi F(\Phi x(k))) - x^*\|_v \\
&\le \tilde{b}_{\tau-1,t,i}\|x(\tau) - x^*\|_v + \gamma\sum_{k=\tau}^{t} b_{k,t,i}\|x(k) - x^*\|_v, \qquad (6.29b)
\end{aligned}$$

where in (6.29a), we use (6.27) which says $\tilde{b}_{\tau-1,t,i} + \sum_{k=\tau}^{t} b_{k,t,i} = 1$ holds for all $i \in \mathcal{M}$; in (6.29b), we use Proposition 6.B.2, which says $\Pi F(\Phi\cdot)$ is γ-contraction in $\|\cdot\|_v$ with fixed point x^*.

Therefore, by substituting (6.29) into (6.28), we obtain that

$$\Upsilon_{t+1} \le \tilde{\beta}_{\tau-1,t}\Upsilon_\tau + \gamma\sup_{i\in\mathcal{M}}\sum_{k=\tau}^{t} b_{k,t,i}\Upsilon_k + \left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\right\|_v + \left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\psi(k)\right\|_v .$$

□

Step 2: Bounding $\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\right\|_v$ and $\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\psi(k)\right\|_v$.

We start with a bound on each individual $\epsilon_1(k), \epsilon_2(k)$, and $\psi(k)$ in Lemma 6.B.2. For simplicity of notation, we define $\underline{v} := \inf_{j\in\mathcal{M}} v_j$.

Lemma 6.B.2. *The following bounds hold almost surely.*

1. $\|\epsilon_1(t)\|_v \le 4\bar{x} + 2C + \frac{\bar{w}}{\underline{v}} := \bar{\epsilon}$.
2. $\|\epsilon_2(t)\|_v \le (2\bar{x} + C)\cdot 2K_1\exp(-\tau/K_2)$.
3. $\|\psi(t)\|_v \le 3\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right)\sum_{k=t-\tau+1}^{t}\alpha_{k-1}$.

Proof of Lemma 6.B.2. By the definition of $\|\cdot\|_v$ in $\mathbb{R}^{\mathcal{M}}$ and its extension to $\mathbb{R}^{\mathcal{N}}$, the induced matrix norm of $\|\cdot\|$ for a matrix $A = [a_{ij}]_{i\in\mathcal{M},j\in\mathcal{N}}$ is given by $\|A\|_v = \sup_{i\in\mathcal{M}}\sum_{j\in\mathcal{N}}\frac{v_{h(j)}}{v_i}\left|a_{ij}\right|$. Recall that the i'th entry of the diagonal matrix D_t is given by $d_{t,i} = \mathbb{P}\left(i_t = i \mid \mathcal{F}_{t-\tau}\right)$. Hence we have that

$$\left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D_t)\right\|_v = \sup_{j\in\mathcal{M}}\sum_{i\in\mathcal{N}} 1(h(i) = j)\cdot\left|1(i = i_t) - d_{t,i}\right| \le 2. \qquad (6.30)$$

Therefore, we can upper bound $\|\epsilon_1(t)\|_v$ by

$$\begin{aligned}
\|\epsilon_1(t)\|_v &= \left\|\Phi^\top \left[(e_{i_t}e_{i_t}^\top - D_t)\left(F(\Phi x(t-\tau)) - \Phi x(t-\tau)\right) + w(t)e_{i_t}\right]\right\|_v \\
&\le \left\|\Phi^\top (e_{i_t}e_{i_t}^\top - D_t)\right\|_v \|F(\Phi x(t-\tau)) - \Phi x(t-\tau)\|_v + |w(t)|\left\|\Phi^\top e_{i_t}\right\|_v \\
&\le 2\|F(\Phi x(t-\tau)) - \Phi x(t-\tau)\|_v + |w(t)|\left\|\Phi^\top e_{i_t}\right\|_v \qquad (6.31a) \\
&\le 2\|F(\Phi x(t-\tau))\|_v + 2\|x(t-\tau)\|_v + \frac{\bar{w}}{\underline{v}} \qquad (6.31b) \\
&\le 4\bar{x} + 2C + \frac{\bar{w}}{\underline{v}}, \qquad (6.31c)
\end{aligned}$$

where we use (6.30) in (6.31a); the triangle inequality, the definition of $\bar{v}$, and Assumption 6.3.4 in (6.31b); Assumption 6.3.3 in (6.31c).

For $\|\epsilon_2(t)\|_v$, recall that

$$\begin{aligned}
\|\epsilon_2(t)\|_v &= \left\|\Phi^\top (D_t - D)\left(F(\Phi x(t-\tau)) - \Phi x(t-\tau)\right)\right\|_v \\
&= \sup_{j\in\mathcal{M}} \frac{1}{v_j}\left|\sum_{i\in\mathcal{N}} 1(h(i)=j)(d_{t,i}-d_i)\left(F(\Phi x(t-\tau)) - \Phi x(t-\tau)\right)_i\right| \\
&= \sup_{j\in\mathcal{M}} \frac{1}{v_j}\left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)\left(F(\Phi x(t-\tau)) - \Phi x(t-\tau)\right)_i\right|. \qquad (6.32)
\end{aligned}$$

By Assumption 6.3.2, we have that

$$\sup_{\mathcal{S}\subseteq\mathcal{N}}\left|\sum_{i\in\mathcal{S}} d_i - \sum_{i\in\mathcal{S}} d_{t,i}\right| \le K_1 \exp(-\tau/K_2). \qquad (6.33)$$

Our objective is to bound the following term in (6.32) for all $j \in \mathcal{M}$:

$$\left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)\left(F(\Phi x(t-\tau)) - \Phi x(t-\tau)\right)_i\right|.$$

Let $M_j := \sup_{i\in h^{-1}(j)} |(F(\Phi x(t-\tau)) - \Phi x(t-\tau))_i|$. Define function

$$g : [-M_j, M_j]^{\mathcal{N}} \to \mathbb{R}, \quad g(y) = \left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)y_i\right|.$$

Suppose $y_{max} \in \arg\max_y g(y)$. We know that for $i \in h^{-1}(j)$, $(y_{max})_i$ is either M_j or $-M_j$ if $d_{t,i} - d_i \neq 0$. Let $S_j := \{i \in h^{-1}(j) \mid (y_{max})_i = M_j\}$ and $S_j' := \{i \in h^{-1}(j) \mid (y_{max})_i = -M_j\}$.

Therefore, we see that

$$\begin{aligned}
&\left|\sum_{i\in h^{-1}(j)} (d_{t,i} - d_i)\left(F(\Phi x(t-\tau)) - \Phi x(t-\tau)\right)_i\right| \\
&\leq \max_{y\in[-M_j,M_j]^{\mathcal{N}}} g(y) \qquad (6.34a)\\
&= \left|\sum_{i\in S_j}(d_{t,i} - d_i)\right| M_j + \left|\sum_{i\in S_j'}(d_{t,i} - d_i)\right| M_j \\
&\leq 2K_1 \exp(-\tau/K_2) M_j, \qquad (6.34b)
\end{aligned}$$

where we use the definition of function g in (6.34a); we use (6.33) in (6.34b).

Substituting (6.34) into (6.32) gives that

$$\begin{aligned}
\|\epsilon_2(t)\|_v &\leq \|F(\Phi x(t-\tau)) - \Phi x(t-\tau)\|_v \cdot 2K_1 \exp(-\tau/K_2)\\
&\leq \left(\|F(\Phi x(t-\tau))\|_v + \|\Phi x(t-\tau)\|_v\right) \cdot 2K_1 \exp(-\tau/K_2) \qquad (6.35a)\\
&\leq (2\bar{x} + C) \cdot 2K_1 \exp(-\tau/K_2), \qquad (6.35b)
\end{aligned}$$

where we use the triangle inequality in (6.35a); we use Assumption 6.3.3 in (6.35b).

As for $\|\psi(t)\|_v$, we have the following bound

$$\begin{aligned}
&\|\psi(t)\|_v\\
&= \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\left(F(\Phi x(t)) - F(\Phi x(t-\tau))\right) - \Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi\left(x(t) - x(t-\tau)\right)\right\|_v\\
&\leq \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\left(F(\Phi x(t)) - F(\Phi x(t-\tau))\right)\right\|_v\\
&\quad + \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi\left(x(t) - x(t-\tau)\right)\right\|_v\\
&\leq \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\right\|_v \cdot \|(F(\Phi x(t)) - F(\Phi x(t-\tau)))\|_v\\
&\quad + \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi\right\|_v \cdot \|(x(t) - x(t-\tau))\|_v. \qquad (6.36)
\end{aligned}$$

Notice that

$$\left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi\right\|_v = \left\|\xi_{h(i_t)}\xi_{h(i_t)}^\top - D'\right\|_v = \sup_{j\in\mathcal{M}}\left|1(h(i_t) = j) - d_j'\right| \leq 1.$$

Substituting this into (6.36) and use (6.30), we obtain that

$$\begin{aligned}
\|\psi(t)\|_v &\leq 2\|F(\Phi x(t)) - F(\Phi x(t-\tau))\|_v + \|x(t) - x(t-\tau)\|_v\\
&\leq 3\|x(t) - x(t-\tau)\|_v\\
&\leq 3\sum_{k=t-\tau+1}^{t} \|x(k) - x(k-1)\|_v. \qquad (6.37)
\end{aligned}$$

By the update rule of x and Assumption 6.3.3, we have that

$$\begin{aligned}\|x(t)-x(t-1)\|_v &\le \alpha_{t-1}\left(\|F(\Phi x(t-1))\|_v + \|x(t-1)\|_v + \frac{\bar{w}}{\underline{v}}\right)\\ &\le \alpha_{t-1}\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right). \qquad (6.38)\end{aligned}$$

Substituting (6.38) into (6.37), we obtain that

$$\|\psi(t)\|_v \le 3\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right)\sum_{k=t-\tau+1}^{t}\alpha_{k-1}.$$

□

Lemma 6.B.3. *If $\alpha_t = \frac{H}{t+t_0}$, where $H > \frac{2}{\sigma'}$ and $t_0 \ge \max(4H, \tau)$, then $\beta_{k,t}, \tilde{\beta}_{k,t}$ satisfies the following*

1. $\beta_{k,t} \le \frac{H}{k+t_0}\left(\frac{k+1+t_0}{t+1+t_0}\right)^{\sigma' H}, \tilde{\beta}_{k,t} \le \left(\frac{k+1+t_0}{t+1+t_0}\right)^{\sigma' H}$.
2. $\sum_{k=1}^{t}\beta_{k,t}^2 \le \frac{2H}{\sigma'}\frac{1}{t+1+t_0}$.
3. $\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau+1}^{k}\alpha_{l-1} \le \frac{8H\tau}{\sigma'}\frac{1}{t+1+t_0}$.

Proof of Lemma 6.B.3. To show Lemma 6.B.3, we only need to substitute σ' for σ in the proof of Qu and Wierman, 2020[Lemma 10]. □

Lemma 6.B.4. *The following inequality holds almost surely*

$$\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\psi(k)\right\|_v \le \frac{24\left(2\bar{x}+C+\frac{\bar{w}}{\underline{v}}\right)H\tau}{\sigma'}\frac{1}{t+1+t_0} := C_\psi\frac{1}{t+1+t_0}.$$

Proof of Lemma 6.B.4. We have that

$$\begin{aligned}\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\psi(k)\right\|_v &\le \sum_{k=\tau}^{t}\alpha_k\left\|\tilde{B}_{k,t}\right\|_v\|\psi(k)\|_v\\ &\le 3\left(2\bar{x}+C+\frac{\bar{w}}{\underline{v}}\right)\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau+1}^{k}\alpha_{l-1} \qquad (6.39a)\\ &\le \frac{24\left(2\bar{x}+C+\frac{\bar{w}}{\underline{v}}\right)H\tau}{\sigma'}\frac{1}{t+1+t_0}, \qquad (6.39b)\end{aligned}$$

where we use Lemma 6.B.2 in (6.39a); Lemma 6.B.3 in (6.39b). □

Lemma 6.B.5. *For each t, with probability at least $1-\delta$, we have*

$$\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon_1(k)\right\|_v \le \frac{H\bar{\epsilon}}{t+t_0}\sqrt{2\tau t\log\left(\frac{2\tau m}{\delta}\right)}.$$

To show Lemma 6.B.5, we need to use Lemma 6.B.6, which is Lemma 13 in Qu and Wierman, 2020.

Lemma 6.B.6. *Let X_t be a $\mathcal{F}_t$-adapted stochastic process which satisfies $\mathbb{E}X_t \mid \mathcal{F}_{t-\tau} = 0$. Further, $|X_t| \le \bar{X}_t$ almost surely. Then with probability $1-\delta$, we have, $\left|\sum_{k=0}^{t}X_t\right| \le \sqrt{2\tau\sum_{k=0}^{t}\bar{X}_k^2\log\left(\frac{2\tau}{\delta}\right)}$.*

Proof of Lemma 6.B.5. Recall that $\sum_{k=\tau}\alpha_k\tilde{B}_{k,t}\epsilon_1(k)$ is a random vector in $\mathbb{R}^{\mathcal{M}}$, with its i'th entry

$$\sum_{k=\tau}^{t}\alpha_k(\epsilon_1)_i(k)\prod_{l=k+1}^{t}(1-\alpha_l d_i').$$

Since step sizes $\{\alpha_l\}$ are deterministic, we see that

$$\mathbb{E}\left[\alpha_k(\epsilon_1)_i(k)\prod_{l=k+1}^{t}(1-\alpha_l d_i') \mid \mathcal{F}_{k-\tau}\right] = \alpha_k\prod_{l=k+1}^{t}(1-\alpha_l d_i')\mathbb{E}\left[(\epsilon_1)_i(k) \mid \mathcal{F}_{k-\tau}\right] = 0.$$

Notice that

$$\alpha_k\prod_{l=k+1}^{t}(1-\alpha_l d_i') = \frac{H}{k+t_0}\prod_{l=k+1}^{t}\left(1-\frac{Hd_i'}{l+t_0}\right) \tag{6.40a}$$

$$\le \frac{H}{k+t_0}\prod_{l=k+1}^{t}\left(1-\frac{2}{l+t_0}\right) \tag{6.40b}$$

$$\le \frac{H}{k+t_0}\prod_{l=k+1}^{t}\left(1-\frac{1}{l+t_0}\right)$$

$$\le \frac{H}{t+t_0},$$

where we use $\alpha_l = \frac{H}{l+t_0}$ in (6.40a); we use $H > \frac{2}{\sigma'}$ in (6.40b).

By the definition of $\bar{\epsilon}$, we also see that $|(\epsilon_1)_i(k)| \le v_i\bar{\epsilon}$. Therefore, by Lemma 6.B.6, we obtain that

$$\left|\sum_{k=\tau}^{t}\alpha_k(\epsilon_1)_i(k)\prod_{l=k+1}^{t}(1-\alpha_l d_i')\right| \le \frac{Hv_i\bar{\epsilon}}{t+t_0}\sqrt{2\tau t\log\left(\frac{2\tau}{\delta}\right)}$$

holds with probability at least $1-\delta$. By union bound, we see that with probability at least $1-\delta$,

$$\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon_1(k)\right\|_v \le \frac{H\bar{\epsilon}}{t+t_0}\sqrt{2\tau t\log\left(\frac{2\tau m}{\delta}\right)}.$$

□

Lemma 6.B.7. *If we set τ to be an integer such that*

$$\tau \ge 2K_2\max\left(\log t, 1\right),$$

we have that

$$\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon_2(k)\right\|_v \le \frac{C_{\epsilon_2}}{t+t_0+1},$$

where $t_0 = \max(\tau, 4H)$ and $C_{\epsilon_2} = (2\bar{x}+C)\cdot 2K_1(1+2K_2+4H)$.

Proof of Lemma 6.B.7. Since $K_2 \ge 1$, the bound is trivial when $t=1$. We consider the case when $t \ge 2$ below.

Since $\alpha_k\tilde{B}_{k,t}$ is a diagonal matrix and its entries are positive and less than 1, we have that

$$\begin{aligned}\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon_2(k)\right\|_v &\le \sum_{k=\tau}^{t}\left\|\alpha_k\tilde{B}_{k,t}\right\|_v\cdot\|\epsilon_2(k)\|_v\\ &\le t\|\epsilon_2(k)\|_v &\text{(6.41a)}\\ &\le t(2\bar{x}+C)\cdot 2K_1\exp(-\tau/K_2), &\text{(6.41b)}\end{aligned}$$

where we use $\left\|\alpha_k\tilde{B}_{k,t}\right\|_v \le 1$ in (6.41a); Lemma 6.B.2 in (6.41b).

To show Lemma 6.B.7, we only need to show

$$t(2\bar{x}+C)\cdot 2K_1(t+\tau+4H)\exp(-\tau/K_2) \le C_{\epsilon_2} \tag{6.42}$$

holds for all $\tau \ge 2K_2\log t$ because $t+t_0+1 \le t+\tau+4H$.

To study how the left hand side of (6.42) changes with τ, we define function

$$g(\tau) = (\tau+t+4H)\exp(-\tau/K_2).$$

Notice that we view τ as real number in function g, so we can get the derivative of g:

$$g'(\tau) = \frac{\exp(-\tau/K_2)}{K_2}(K_2-t-4H-\tau).$$

Therefore, when $\tau \geq 2K_2 \log t$, we always have $g'(\tau) < 0$. Hence we obtain that

$$g(\tau) \leq g(2K_2 \log t) = \frac{2K_2 \log t + t + 4H}{t^2} \leq \frac{1 + 2K_2 + 4H}{t} \tag{6.43}$$

holds for all $\tau \geq 2K_2 \log t$.

Substituting (6.43) into (6.42) finishes the proof. □

Step 3: Bounding the error sequence. Based on the recursive relationship we derived in Lemma 6.B.1 and the bounds we obtained in Step 2, we want to show that, with probability $1 - \delta$,

$$\Upsilon_t \leq \frac{C_a}{\sqrt{t + t_0}} + \frac{C'_a}{t + t_0}, \tag{6.44}$$

holds for all $\tau \leq t \leq T$, where

$$C_a = \frac{2H\bar{\epsilon}}{1-\gamma} \sqrt{2\tau \log\left(\frac{2\tau m T}{\delta}\right)}, C'_a = \frac{4}{1-\gamma} \max\left(C_\psi + C_{\epsilon_2}, 2\bar{x}(\tau + t_0)\right).$$

Notice that C_a and C'_a are independent of t but may dependent on T. We set $\tau = 2K_2 \log T$.

By applying union bound to Lemma 6.B.5, we see that with probability at least $1 - \delta$, for any $t \leq T$,

$$\left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon_1(k) \right\|_v \leq \frac{C_{\epsilon_1}}{\sqrt{t + 1 + t_0}},$$

where $C_{\epsilon_1} = H\bar{\epsilon}\sqrt{2\tau \log\left(\frac{2\tau m T}{\delta}\right)}$.

Therefore, we get with probability $1 - \delta$, (6.45) holds for all $\tau \leq t \leq T$:

$$\Upsilon_{t+1} \leq \tilde{\beta}_{\tau-1,t} \Upsilon_\tau + \gamma \sup_{i \in \mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i} \Upsilon_k + \frac{C_{\epsilon_1}}{\sqrt{t + 1 + t_0}} + \frac{C_\psi + C_{\epsilon_2}}{t + 1 + t_0}. \tag{6.45}$$

We now condition on (6.45) to show (6.44) by induction. (6.44) is true for $t = \tau$, as $\frac{C'_a}{\tau + t_0} \geq \frac{8}{1-\gamma}\bar{x} \geq \Upsilon_\tau$, where we have used $\Upsilon_\tau = \|x(\tau) - x^*\|_v \leq \|x(\tau)\|_v + \|x^*\|_v \leq 2\bar{x}$. Then, assuming (6.44) is true for up to $k \leq t$. By (6.45), we have that

$$\begin{aligned}\Upsilon_{t+1} &\leq \tilde{\beta}_{\tau-1,t} \Upsilon_\tau + \gamma \sup_{i \in \mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i} \left[\frac{C_a}{\sqrt{k + t_0}} + \frac{C'_a}{k + t_0}\right] + \frac{C_{\epsilon_1}}{\sqrt{t + 1 + t_0}} + \frac{C_\psi + C_{\epsilon_2}}{t + 1 + t_0} \\ &\leq \tilde{\beta}_{\tau-1,t} \Upsilon_\tau + \gamma C_a \sup_{i \in \mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i} \frac{1}{\sqrt{k + t_0}} + \gamma C'_a \sup_{i \in \mathcal{M}} \sum_{k=\tau}^{t} \frac{1}{k + t_0} b_{k,t,i}\end{aligned}$$

$$+\frac{C_{\epsilon_1}}{\sqrt{t+1+t_0}}+\frac{C_\psi + C_{\epsilon_2}}{t+1+t_0}. \tag{6.46}$$

We use the following auxiliary lemma to handle the second and the third term in (6.46).

Lemma 6.B.8. *If* $\sigma' H(1-\sqrt{\gamma}) \geq 1, t_0 \geq 1,$ *and* $\alpha_0 \leq \frac{1}{2}$*, then, for any* $i \in \mathcal{N}$*, and any* $0 < \omega \leq 1$*, we have*

$$\sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{(k+t_0)^\omega} \leq \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}.$$

Proof of Lemma 6.B.8. Recall that $\alpha_k = \frac{H}{k+t_0}$, and $b_{k,t,i} = \alpha_k d_i' \prod_{l=k+1}^{t}(1-\alpha_l d_i')$, where $d_i' \geq \sigma'$.

Define $e_t = \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{(k+t_0)^\omega}$. We use induction on t to show that $e_t \leq \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}$.

The statement is clearly true for $t=\tau$. Assume it is true for $t-1$. Notice that

$$\begin{aligned}
e_t &= \sum_{k=\tau}^{t-1} b_{k,t,i}\frac{1}{(k+t_0)^\omega} + b_{t,t,i}\frac{1}{(t+t_0)^\omega} \\
&= (1-\alpha_t d_i')\sum_{k=\tau}^{t-1} b_{k,t-1,i}\frac{1}{(k+t_0)^\omega} + \alpha_t d_i' \frac{1}{(t+t_0)^\omega} \qquad (6.47a)\\
&= (1-\alpha_t d_i') e_{t-1} + \alpha_t d_i' \frac{1}{(t+t_0)^\omega} \\
&\leq (1-\alpha_t d_i')\frac{1}{\sqrt{\gamma}(t+t_0)^\omega} + \alpha_t d_i' \frac{1}{(t+t_0)^\omega} \qquad (6.47b)\\
&= \left[1-\alpha_t d_i'(1-\sqrt{\gamma})\right]\frac{1}{\sqrt{\gamma}(t+t_0)^\omega},
\end{aligned}$$

where we use $b_{t,t,i} = \alpha_t d_i'$ in (6.47a); we use the induction assumption in (6.47b).

Plugging in $\alpha_t = \frac{H}{t+t_0}$, we see that

$$\begin{aligned}
e_t &\leq \left[1-\frac{\sigma' H}{t+t_0}(1-\sqrt{\gamma})\right]\frac{1}{\sqrt{\gamma}(t+t_0)^\omega} \qquad (6.48a)\\
&= \left[1-\frac{\sigma' H}{t+t_0}(1-\sqrt{\gamma})\right]\left(1+\frac{1}{t+t_0}\right)^\omega \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} \\
&\leq \left(1-\frac{1}{t+t_0}\right)\left(1+\frac{1}{t+t_0}\right)^\omega \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} \qquad (6.48b)\\
&\leq \left(1-\frac{1}{t+t_0}\right)\left(1+\frac{1}{t+t_0}\right) \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} \qquad (6.48c)
\end{aligned}$$

$$\leq \frac{1}{\sqrt{\gamma}(t+1+t_0)^{\omega}},$$

where we use $d_i' \geq \sigma'$ in (6.48a); we use the assumption that $\sigma' H(1-\sqrt{\gamma}) \geq 1$ in (6.48b); we use $0 < \omega \leq 1$ in (6.48c). □

Applying Lemma 6.B.8 to (6.46), we see that

$$\begin{aligned}\Upsilon_{t+1} &\leq \tilde{\beta}_{\tau-1,t}\Upsilon_\tau + \sqrt{\gamma}C_a\frac{1}{\sqrt{t+1+t_0}} + \sqrt{\gamma}C_a'\frac{1}{t+1+t_0}\\ &\quad + C_{\epsilon_1}\frac{1}{\sqrt{t+1+t_0}} + (C_\psi + C_{\epsilon_2})\frac{1}{t+1+t_0} \qquad (6.49a)\\ &\leq \left(\sqrt{\gamma}C_a\frac{1}{\sqrt{t+1+t_0}} + C_{\epsilon_1}\frac{1}{\sqrt{t+1+t_0}}\right)\\ &\quad + \left(\sqrt{\gamma}C_a'\frac{1}{t+1+t_0} + (C_\psi + C_{\epsilon_2})\frac{1}{t+1+t_0} + \left(\frac{\tau+t_0}{t+1+t_0}\right)^{\sigma' H}\Upsilon_\tau\right), \qquad (6.49b)\end{aligned}$$

where we use Lemma 6.B.8 in (6.49a); we use the bound on $\tilde{\beta}_{\tau-1,t}$ in Lemma 6.B.3 in (6.49b).

To bound the two terms in (6.49b), we define

$$\chi_t := \sqrt{\gamma}C_a\frac{1}{\sqrt{t+1+t_0}} + C_{\epsilon_1}\frac{1}{\sqrt{t+1+t_0}}$$

and

$$\chi_t' = \sqrt{\gamma}C_a'\frac{1}{t+1+t_0} + (C_\psi + C_{\epsilon_2})\frac{1}{t+1+t_0} + \left(\frac{\tau+t_0}{t+1+t_0}\right)^{\sigma' H} a_\tau.$$

To finish the induction, it suffices to show that $\chi_t \leq \frac{C_a}{\sqrt{t+1+t_0}}$ and $\chi_t' \leq \frac{C_a'}{t+1+t_0}$. To see this

$$\begin{aligned}\chi_t\frac{\sqrt{t+1+t_0}}{C_a} &= \sqrt{\gamma} + \frac{C_{\epsilon_1}}{C_a},\\ \chi_t'\frac{t+1+t_0}{C_a'} &= \sqrt{\gamma} + \frac{C_\psi + C_{\epsilon_2}}{C_a'} + \frac{\Upsilon_\tau(\tau+t_0)}{C_a'}\left(\frac{\tau+t_0}{t+1+t_0}\right)^{\sigma' H-1}.\end{aligned}$$

It suffices to show that $\frac{C_{\epsilon_1}}{C_a} \leq 1-\sqrt{\gamma}$, $\frac{C_\psi + C_{\epsilon_2}}{C_a'} \leq \frac{1-\sqrt{\gamma}}{2}$, and $\frac{\Upsilon_\tau(\tau+t_0)}{C_a'} \leq \frac{1-\sqrt{\gamma}}{2}$. Recall that

$$C_a = \frac{2H\bar{\epsilon}}{1-\gamma}\sqrt{2\tau\log\left(\frac{2\tau m T}{\delta}\right)}, C_a' = \frac{4}{1-\gamma}\max\left(C_\psi + C_{\epsilon_2}, 2\bar{x}(\tau+t_0)\right),$$

and

$$C_{\epsilon_1} = H\bar{\epsilon}\sqrt{2\tau\log\left(\frac{2\tau m T}{\delta}\right)}.$$

Using that $\Upsilon_\tau \leq 2\bar{x}$, one can check that C_a and C_a' satisfy the above three inequalities.

Asymptotic Convergence of TD Learning with State Aggregation

Our asymptotic convergence result for TD learning with state aggregation builds upon the asymptotic convergence result for TD learning with linear function approximation shown in Tsitsiklis and Van Roy, 1997. For completeness, we first present the main result of Tsitsiklis and Van Roy, 1997 in Theorem 6.B.9. In order to do this, we must first state a few definitions and assumptions made in Tsitsiklis and Van Roy, 1997.

We use $\phi(i) \in \mathbb{R}^m$ to denote the feature vector associated with state $i \in \mathcal{N}$. Feature matrix Φ is a n-by-m matrix whose i'th row is $\phi(i)^\top$. Starting from $\theta(0) = \mathbf{0}$, the $TD(\lambda)$ algorithm keeps updating θ, ψ by the following update rule,

$$\theta(t+1) = \theta(t) + \alpha_t d_t \psi_t,$$
$$\psi_{t+1} = \gamma\lambda\psi_t + \phi(i_{t+1}),$$

where ψ_t is named *eligible vector* in Tsitsiklis and Van Roy, 1997 and satisfies $\psi_0 = \phi(i_0)$.

Recall that $D = diag(d_1, d_2, \cdots, d_n)$ denotes the stationary distribution of Markov chain $\{i_t\}$. For vectors $x, y \in \mathbb{R}^n$, we define inner product $\langle x, y\rangle = x^\top D y$. The induced norm of this inner product is $\|\cdot\|_D = \sqrt{\langle\cdot,\cdot\rangle_D}$. Let $L_2(\mathcal{N}, D)$ denote the set of vectors $V \in \mathbb{R}^n$ such that $\|V\|_D$ is finite.

Recall that we define $\Pi = (\Phi^\top D\Phi)^{-1}\Phi^\top D$. As shown in Tsitsiklis and Van Roy, 1997, the projection matrix that projects an arbitrary vector in $\mathbb{R}^n$ to the set $\{\Phi\theta \mid \theta \in \mathbb{R}^m\}$ is given by $\Phi\Pi$, i.e. for any $V \in L_2(\mathcal{N}, D)$, we have

$$\Phi\Pi V = \operatorname*{arg\,min}_{\bar{V}\in\{\Phi\theta|\theta\in\mathbb{R}^m\}} \left\|V - \bar{V}\right\|_D.$$

Notice that our definition of matrix Π is slightly different with Tsitsiklis and Van Roy, 1997 because we want to be consistent with Section 6.3.

To characterize the TD(λ) algorithm's dynamics, Tsitsiklis and Van Roy, 1997 defines $T^{(\lambda)} : L_2(\mathcal{N}, D) \to L_2(\mathcal{N}, D)$ operator as following: for all $V \in \mathbb{R}^n$, let the i'th dimension of $\left(T^{(\lambda)}V\right)$ be defined as

$$\left(T^{(\lambda)}V\right)_i = \begin{cases} (1-\lambda)\sum_{m=0}^{\infty}\lambda^m \mathbb{E}\left[\sum_{t=0}^{m}\gamma^t r(i_t, i_{t+1}) + \gamma^{m+1}V_{i_{m+1}} \mid i_0 = i\right] & \text{if } \lambda < 1 \\ \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(i_t, i_{t+1}) \mid i_0 = i\right] & \text{if } \lambda = 1. \end{cases}$$

If V is an approximation of the value function V^*, $T^{(\lambda)}$ can be viewed as an improved approximation to V^*. Notice that when $\lambda = 0$, $T^{(\lambda)}$ is identical with the Bellman operator.

Formally, Tsitsiklis and Van Roy, 1997 made four necessary assumptions for their main result (Theorem 6.B.9). We omit the third assumption (Tsitsiklis and Van Roy, 1997[Assumption 3]) in our summary because it must hold when the state space $\mathcal{N}$ is finite.

The first assumption (Tsitsiklis and Van Roy, 1997[Assumption 1]) concerns the stationary distribution and the reward function of the Markov chain $\{i_t\}$. It must hold when Assumption 6.3.2 holds and every stage reward r_t is upper bounded by $\bar{r}$, as assumed by Theorem 6.3.3.

Assumption 6.B.1. *The transition probability and cost function satisfies the following two conditions:*

1. *The Markov chain $\{i_t\}$ is irreducible and aperiodic. Furthermore, there is a unique distribution d that satisfies $d^\top P = d^\top$ with $d_i > 0$ for all $i \in \mathcal{N}$. Let $\mathbb{E}_0$ stand for expectation with respect to this distribution.*
2. *The reward function $r(i_t, i_{t+1})$ satisfies $\mathbb{E}_0\left[r^2(i_t, i_{t+1})\right] < \infty$.*

The second assumption (Tsitsiklis and Van Roy, 1997[Assumption 2]) concerns the feature vectors and the feature matrix. It must hold when Φ is defined as (6.5).

Assumption 6.B.2. *The following two conditions hold for Φ:*

1. *The matrix Φ has full column rank; that is, the m columns (named basis functions in Tsitsiklis and Van Roy, 1997) $\{\phi_k \mid k = 1, \cdots, m\}$ are linearly independent.*
2. *For every k, the basis function ϕ_k satisfies $\mathbb{E}_0\left[\phi_k^2(i_t)\right] < \infty$.*

The third assumption (Tsitsiklis and Van Roy, 1997[Assumption 4]) concerns the learning step size. It must hold if the learning step sizes are as defined in Theorem 6.3.3.

Assumption 6.B.3. *The step sizes α_t are positive, nonincreasing, and chosen prior to execution of the algorithm. Furthermore, they satisfy $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.*

Now we are ready to present the main asymptotic convergence result given in Tsitsiklis and Van Roy, 1997.

Theorem 6.B.9. *Under Assumptions 6.B.1, 6.B.2, 6.B.3, the following hold.*

1. *The value function V is in $L_2(\mathcal{N}, D)$.*
2. *For any $\lambda \in [0,1]$, the TD(λ) algorithm with linear function approximation converges with probability one.*
3. *The limit of convergence θ^* is the unique solution of the equation*

$$\Pi T^{(\lambda)}\left(\Phi\theta^*\right) = \theta^*.$$

4. *Furthermore, θ^* satisfies*

$$\|\Phi\theta^* - V^*\|_D \leq \frac{1-\lambda\gamma}{1-\gamma}\|\Phi\Pi V^* - V^*\|_D. \tag{6.50}$$

Notice that (6.50) is not exactly the result we want to obtain. Specifically, we want the both sides of (6.50) to be in $\|\cdot\|_\infty$ instead of $\|\cdot\|_D$. Although this kind of result is not obtainable for general TD learning with linear function approximation, we can leverage the special assumptions for state aggregation, which are summarized below:

Assumption 6.B.4. *$h : \mathcal{N} \to \mathcal{M}$ is a surjective function from set $\mathcal{N}$ to $\mathcal{M}$. The feature matrix Φ is as defined in* (6.5)*, i.e., the feature vector associated with state $i \in \mathcal{N}$ is given by*

$$\phi_k(i) = \begin{cases} 1 & \text{if } k = h(i) \\ 0 & \text{otherwise} \end{cases}, \forall k \in \mathcal{M}.$$

Further, if $h(i) = h(i')$ for $i, i' \in \mathcal{N}$, we have $|V^*(i) - V^*(i')| \leq \zeta$ for a fixed positive constant ζ.

Under Assumption 6.B.4, we can show the asymptotic error bound in the infinity norm as we desired:

Theorem 6.B.10. *Under Assumptions 6.B.1, 6.B.2, 6.B.3, if Assumption 6.B.4 also holds, the limit of convergence θ^* of the $TD(\lambda)$ algorithm satisfies*

$$\|\Phi\theta^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\|\Phi\Pi V^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\zeta.$$

To show Theorem 6.B.10, we need to prove several auxiliary lemmas first.

Lemma 6.B.11. *Under Assumption 6.B.1, for any $V \in L_2(\mathcal{N}, D)$, we have $\|PV\|_\infty \leq \|V\|_\infty$.*

Proof of Lemma 6.B.11. This lemma holds because the transition matrix P is non-expansive in infinity norm. □

Lemma 6.B.12. *Under Assumption 6.B.1, for any $V, \bar{V} \in L_2(\mathcal{N}, D)$, we have*

$$\left\|T^{(\lambda)}V - T^{(\lambda)}\bar{V}\right\|_\infty \leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|V - \bar{V}\|_\infty.$$

Proof of Lemma 6.B.12. By the definition of $T^{(\lambda)}$, we have that

$$\begin{aligned}\left\|T^{(\lambda)}V - T^{(\lambda)}\bar{V}\right\|_\infty &= \left\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma P)^{m+1}\left(V - \bar{V}\right)\right\|_\infty \\ &\leq (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\|V - \bar{V}\|_\infty \qquad (6.51a) \\ &\frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|V - \bar{V}\|_\infty,\end{aligned}$$

where inequality (6.51a) holds because $\left\|V - \bar{V}\right\|_\infty < \infty$ so we use Lemma 6.B.11. □

Lemma 6.B.13. *Under Assumption 6.B.1 and 6.B.4, we have*

$$\|\Phi\Pi V^* - V^*\|_\infty \leq \zeta \qquad (6.52)$$

and for any $V \in L_2(\mathcal{N}, D)$

$$\|\Phi\Pi V\|_\infty \leq \|V\|_\infty. \qquad (6.53)$$

Proof of Lemma 6.B.13. For $j \in \mathcal{M}$, we use $h^{-1}(j) \subseteq \mathcal{N}$ to denote all the elements in $\mathcal{N}$ whose feature is e_j, i.e., $h^{-1}(j) = \{i \mid i \in \mathcal{N}, h(i) = j\}$. Since h is surjection, $h^{-1}(j) \neq \emptyset, \forall j \in \mathcal{M}$. Since $\Phi\Pi$ is the projection matrix that projects a vector in $\mathbb{R}^n$ to the set $\{\Phi\theta \mid \theta \in \mathbb{R}^m\}$, we have

$$\Pi V = \arg\min_{\theta \in \mathbb{R}^m} \sum_{j \in \mathcal{M}} \sum_{i \in h^{-1}(j)} d_i\left(V_i - \theta_j\right).$$

Hence the optimal θ_j must be in the range $\left[\min_{i \in h^{-1}(j)} V_i, \max_{i \in h^{-1}(j)} V_i\right]$. Therefore, we see that

$$|(\Phi\Pi V)_i| = \left|(\Pi V)_{h(i)}\right| \leq \max_{i' \in h^{-1}(h(i))} |V_{i'}|,$$

which shows (6.53). Besides, we also have

$$|(\Phi\Pi V)_i - V_i| \leq \max\left(\left|\min_{i'\in h^{-1}(h(i))} V_{i'} - V_i\right|, \left|\max_{i'\in h^{-1}(h(i))} V_{i'} - V_i\right|\right). \tag{6.54}$$

holds for all $z \in \mathcal{Z}$. Let $V = V^*$ and use Assumption 6.B.4 in (6.54) gives (6.52). □

Now we come back to the proof of Theorem 6.B.10.

Notice that

$$\begin{aligned} \|\Phi\theta^* - V^*\|_\infty &\leq \|\Phi\theta^* - \Phi\Pi V^*\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty && (6.55a)\\ &= \left\|\Phi\Pi T^{(\lambda)}\left(\Phi\theta^*\right) - \Phi\Pi V^*\right\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty && (6.55b)\\ &\leq \left\|T^{(\lambda)}\left(\Phi\theta^*\right) - V^*\right\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty && (6.55c)\\ &\leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|\Phi\theta^* - V^*\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty, && (6.55d) \end{aligned}$$

where we use the triangle inequality in (6.55a); Theorem 6.B.9 in (6.55b); Lemma 6.B.13 in (6.55c); Lemma 6.B.12 in (6.55d).

Therefore, we obtain that

$$\|\Phi\theta^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\|\Pi V^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\zeta,$$

where we use Lemma 6.B.13 in the second inequality.

Proof of Theorem 6.3.3

Before presenting the proof of Theorem 6.3.3, we first show two upper bounds that are needed in the assumptions of Theorem 6.3.2.

Proposition 6.B.3. *Under the same assumptions as Theorem 6.3.3, we have* $\|\theta(t)\|_\infty \leq \bar{\theta} := \frac{\bar{r}}{1-\gamma}$ *holds for all* t *almost surely and* $\|\theta^*\|_\infty \leq \bar{\theta}$. $|w(t)| \leq \bar{w} := \frac{2\bar{r}}{1-\gamma}$ *also holds for all* t *almost surely.*

Proof of Proposition 6.B.3. We show $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ by induction on t. The statement holds for $t = 0$ because we initialize $\theta(0) = \mathbf{0}$. Suppose the statement holds for t. By the induction assumption, we see that

$$\begin{aligned} \theta_{h(i_t)}(t+1) &= (1-\alpha_t)\theta_{h(i_t)}(t) + \alpha_t\left[r_t + \gamma\theta_{h(i_{t+1})}(t)\right]\\ &\leq (1-\alpha_t)\|\theta(t)\|_\infty + \alpha_t\left[r_t + \gamma\|\theta(t)\|_\infty\right]\\ &\leq (1-\alpha_t)\frac{\bar{r}}{1-\gamma} + \alpha_t\left[r_t + \gamma\cdot\frac{\bar{r}}{1-\gamma}\right]\\ &\leq \frac{\bar{r}}{1-\gamma}. \end{aligned}$$

For $j \neq h(i_t), j \in \mathcal{M}$, we have that

$$\theta_j(t+1) = \theta_j(t) \leq \|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}.$$

Hence the statement also holds for $t+1$. Therefore, we have showed $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ by induction.

By Theorem 6.B.9, we know $\theta^* = \lim_{t\to\infty} \theta(t)$. Since we have already shown that $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ holds for all t, we must have $\|\theta^*\|_\infty \leq \frac{\bar{r}}{1-\gamma}$.

Using $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$, we see that

$$\begin{aligned}|w(t)| &\leq |r_t| + \gamma\left|\theta_{h(i_{t+1})}(t)\right| - \left|\mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\left[r(i_t, i') + \gamma\theta_{h(i')}(t)\right]\right| \\ &\leq 2\bar{r} + 2\gamma\bar{\theta} \\ &= \frac{2\bar{r}}{1-\gamma}.\end{aligned}$$

□

Now we come back to the proof of Theorem 6.3.3. Recall that we define F as the Bellman Policy Operator and the noise sequence $w(t)$ as

$$w(t) = r_t + \gamma\theta_{h(i_{t+1})}(t) - \mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\left[r(i_t, i') + \gamma\theta_{h(i')}(t)\right].$$

Let θ^* be the unique solution of the equation

$$\Pi F(\Phi\theta^*) = \theta^*.$$

By the triangle inequality, we have that

$$\begin{aligned}\|\Phi\cdot\theta(T) - V^*\|_\infty &\leq \|\Phi\cdot\theta(T) - \Phi\cdot\theta^*\|_\infty + \|\Phi\cdot\theta^* - V^*\|_\infty \\ &\leq \|\theta(T) - \theta^*\|_\infty + \|\Phi\cdot\theta^* - V^*\|_\infty. \end{aligned} \tag{6.56}$$

We first bound the first term of (6.56) by Theorem 6.3.2. To do this, we first rewrite the update rule of TD learning with state aggregation (6.7) in the form of the SA update rule (6.4):

$$\begin{aligned}\theta_{h(i_t)}(t+1) &= \theta_{h(i_t)}(t) + \alpha_t\left(F_{i_t}\left(\Phi\theta(t)\right) - \theta_{h(i_t)}(t) + w(t)\right), \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(i_t), j \in \mathcal{M}.\end{aligned}$$

Now we verify all the assumptions of Theorem 6.3.2. Assumption 6.3.2 is assumed to be satisfied in the body of Theorem 6.3.3. As for Assumption 6.3.3, F is γ-contraction in the infinity norm because it is the Bellman operator, and we can set

$C = \frac{2\bar{r}}{1-\gamma}$ so that $C \geq (1+\gamma)\|y^*\|_\infty$ (see the discussion below Assumption 6.3.3). As for Assumption 6.3.4, by the definition of noise sequence $w(t)$, we see that

$$\begin{aligned}\mathbb{E}\left[w(t) \mid \mathcal{F}_t\right] &= \mathbb{E}\left[r_t + \gamma\theta_{h(i_{t+1})}(t) - \mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\left[r(i_t, i') + \gamma\theta_{h(i')}(t)\right] \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[r_t + \gamma\theta_{h(i_{t+1})}(t) \mid \mathcal{F}_t\right] - \mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\left[r(i_t, i') + \gamma\theta_{h(i')}(t)\right] \\ &= 0.\end{aligned}$$

In addition, we can set $\bar{w} = \frac{2\bar{r}}{1-\gamma}$ according to Proposition 6.B.3. Finally, we can set $\bar{\theta} = \frac{\bar{r}}{1-\gamma}$ according to Proposition 6.B.3.

Therefore, by Theorem 6.3.2, we see that

$$\|\theta(T) - \theta^*\|_\infty \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0}, \text{ where} \tag{6.57}$$

$$\begin{aligned}C_a &= \frac{40H\bar{r}}{(1-\gamma)^2}\sqrt{K_2\log T} \cdot \sqrt{\log T + \log\log T + \log\left(\frac{4mK_2}{\delta}\right)}, \\ C_a' &= \frac{8\bar{r}}{(1-\gamma)^2}\max\left(\frac{144K_2H\log T}{\sigma'} + 4K_1(1+2K_2+4H), 2K_2\log T + t_0\right).\end{aligned}$$

As for the second term of (6.56), by Theorem 6.B.10, we have that

$$\|\Phi \cdot \theta^* - V^*\|_\infty \leq \frac{\zeta}{1-\gamma}. \tag{6.58}$$

Substituting (6.57) and (6.58) into (6.56) finishes the proof.

Application of the SA Scheme to Q-learning with State and Action Aggregation

We study Q-learning with state and action aggregation in a setting that is a generalization of the tabular setting studied in Qu and Wierman, 2020. Specifically, we consider an MDP M with a finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. Suppose the transition probability is given by $\mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a) = \mathbb{P}(s' \mid s, a)$, and the stage reward at time step t is a random variable r_t with its expectation given by R_{s_t,a_t}. Under a stochastic policy π, the Q function (vector) $Q^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is defined as

$$Q^\pi_{s,a} = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t \middle| (s_0, a_0) = (s, a)\right],$$

where $0 \leq \gamma < 1$ is the discounting factor. We use Q^* to denote the Q function corresponding to the optimal policy π^*.

Similar to Qu and Wierman, 2020, we assume the trajectory $\{(s_t, a_t, r_t)\}_{t=0}^{\infty}$ is sampled by implementing a fixed behavioral stochastic policy π. In Q-learning with state and action aggregation, the state abstraction ψ_1 operates on the state space $\mathcal{S}$ and the action abstraction ψ_2 operates on action space $\mathcal{A}$. For simplicity of notation, we define the abstraction space as $\mathcal{M} = \psi_1(\mathcal{S}) \times \psi_2(\mathcal{A})$ and the abstraction operator $h : \mathcal{S} \times \mathcal{A} \to \mathcal{M}$ as $h(s, a) = (\psi_1(s), \psi_2(a))$. The update rule for Q-learning with state and action aggregation is then given by

$$\begin{aligned} \theta_{h(s_t,a_t)}(t+1) &= (1-\alpha_t)\theta_{h(s_t,a_t)}(t) + \alpha_t \left[r_t + \gamma \max_{a\in\mathcal{A}} \theta_{h(s_{t+1},a)}(t) \right], \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(s_t, a_t). \end{aligned} \tag{6.59}$$

As a remark, some previous work considers abstraction on the state space $\mathcal{S}$ but does not compress the action space (see Jiang, 2018). In contrast, our setting also compresses the action space, and when ψ_2 is the identity map, our setting reduces to the case with only state aggregation.

We define function F as the *Bellman Optimality Operator*, i.e.,

$$F_{s,a}(Q) = R_{s,a} + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \max_{a'\in\mathcal{A}} Q_{s',a'}.$$

It is shown in Bertsekas and Tsitsiklis, 1996 that Q^* is the unique fixed point of function F. By viewing $\mathcal{S} \times \mathcal{A}$ as $\mathcal{N}$, we can define matrix $\Phi \in \mathcal{N} \times \mathcal{M}$ as in (6.5). We can rewrite the update rule (6.59) as

$$\begin{aligned} \theta_{h(s_t,a_t)}(t+1) &= \theta_{h(s_t,a_t)}(t) + \alpha_t \left[F_{s_t,a_t}(\Phi\theta(t)) - \theta_{h(s_t,a_t)}(t) + w(t) \right], \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(s_t, a_t), \end{aligned}$$

where

$$\begin{aligned} w(t) &= r_t + \gamma \max_{a\in\mathcal{A}} \theta_{h(s_{t+1},a)}(t) - F_{s_t,a_t}(\Phi\theta(t)) \\ &= (r_t - R_{s_t,a_t}) + \gamma \left[\max_{a\in\mathcal{A}} \theta_{h(s_{t+1},a)}(t) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)} \max_{a'\in\mathcal{A}} \theta_{h(s',a')}(t) \right]. \end{aligned}$$

Hence we have $\mathbb{E}[w(t) \mid \mathcal{F}_t] = 0$. In order to apply Theorem 6.3.2, we need the following assumption on the induced Markov chain of stochastic policy π which is standard, cf. Qu and Wierman, 2020.

Assumption 6.B.5. *The following conditions hold:*

1. *For each time step t, the stage reward r_t satisfies $|r_t| \leq \bar{r}$ almost surely.*

2. *Under the behavioral policy π, the induced Markov chain (s_t, a_t) with state space $\mathcal{S} \times \mathcal{A}$ satisfies Assumption 6.3.2 with stationary distribution d and parameters σ', K_1, K_2.*

The next assumption is approximate Q^*-irrelevant abstraction, which measures the quality of the abstraction map and is standard in the literature (see Jiang, 2018).

Assumption 6.B.6. *There exists an abstract Q function $q : \mathcal{M} \to \mathbb{R}$ such that $\|\Phi q - Q^*\|_\infty \le \epsilon_{Q^*}$.*

We can now state our theorem for Q-learning with state aggregation.

Theorem 6.B.14. *Under Assumption 6.B.5 and 6.B.6, suppose the step size of Q-learning with state aggregation is given by $\alpha_t = \frac{H}{t+t_0}$, where $t_0 = \max(4H, 2K_2 \log T)$ and $H \ge \frac{2}{\sigma'(1-\gamma)}$. Then, with probability at least $1 - \delta$,*

$$\|\Phi \cdot \theta(T) - Q^*\|_\infty \le \frac{C_a}{\sqrt{T+t_0}} + \frac{C'_a}{T+t_0} + \frac{2\epsilon_{Q^*}}{1-\gamma}, \textit{ where}$$

$$C_a = \frac{40H\bar{r}}{(1-\gamma)^2}\sqrt{K_2 \log T} \cdot \sqrt{\log T + \log\log T + \log\left(\frac{4mK_2}{\delta}\right)},$$

$$C'_a = \frac{8\bar{r}}{(1-\gamma)^2}\max\left(\frac{144K_2 H \log T}{\sigma'} + 4K_1(1+2K_2+4H), 2K_2\log T + t_0\right).$$

Proof of Theorem 6.B.14. Define θ^* as the unique solution of equation $\theta = \Pi F(\Phi\theta)$, where the definition of Π is given in (6.6). Under Assumption 6.B.5, we see that $\|\theta^*\|_\infty \le \frac{\bar{r}}{1-\gamma}$: otherwise, by assuming that $|\theta_i^*| = \|\theta^*\|_\infty > \frac{\bar{r}}{1-\gamma}$, we can derive a contradiction that $\|\Pi F(\Phi\theta^*)\|_\infty < |\theta_i^*|$. To see this, recall that linear operators Π and Φ are non-expansions in the infinity norm (see Proposition 6.B.2), and $\|F(v)\|_\infty < \|v\|_\infty$ for a vector $v \in \mathbb{R}^{\mathcal{N}}$ if $\|v\|_\infty > \frac{\bar{r}}{1-\gamma}$.

Further, using a similar approach with the proof of Proposition 6.B.3, we also see that

$$\|\theta(t)\|_\infty \le \bar{\theta} := \frac{\bar{r}}{1-\gamma}, |w(t)| \le \bar{w} := \frac{2\bar{r}}{1-\gamma}$$

hold for all t almost surely.

Therefore, by Theorem 6.3.2, we obtain that

$$\|\theta(T) - \theta^*\|_\infty \le \frac{C_a}{\sqrt{T+t_0}} + \frac{C'_a}{T+t_0}. \tag{6.60}$$

To finish the proof of Theorem 6.B.14, we only need to show that

$$\|\Phi\theta^* - Q^*\| \leq \frac{2\epsilon_{Q^*}}{1-\gamma}. \tag{6.61}$$

Given the behavioral policy π, we use $\{d_{s,a} \mid (s,a) \in \mathcal{S}\times\mathcal{A}\}$ to denote the stationary distribution under policy π. Recall that we define $\mathcal{M} = \psi_1(\mathcal{S}) \times \psi_2(\mathcal{A})$. For each abstract state-action pair $(x,y) \in \mathcal{M}$, we define a distribution $p_{(x,y)}$ over $h^{-1}(x,y)$ such that

$$p_{(x,y)}(s,a) = \frac{d_{s,a}}{\sum_{(\tilde{s},\tilde{a})\in h^{-1}(x,y)} d_{\tilde{s},\tilde{a}}}, \forall (s,a) \in h^{-1}(x,y).$$

Using the set of distributions $\{p_{(x,y)} \mid (x,y) \in \mathcal{M}\}$, we define two new MDPs:

$$M_\psi = \left(\psi_1(\mathcal{S}), \psi_2(\mathcal{A}), P_\psi, R_\psi, \gamma\right), \tag{6.62}$$

where $(R_\psi)_{x,y} = \mathbb{E}_{(s,a)\sim p_{(x,y)}}[R_{s,a}]$, and $P_\psi(x' \mid x,y) = \mathbb{E}_{(s,a)\sim p_{(x,y)}}[P(x' \mid s,a)]$; and

$$M'_\psi = (\mathcal{S}, \mathcal{A}, P'_\psi, R'_\psi, \gamma), \tag{6.63}$$

where $(R'_\psi)_{s,a} = \mathbb{E}_{(\tilde{s},\tilde{a})\sim p_{h(s,a)}}[R_{\tilde{s},\tilde{a}}], P'_\psi(s' \mid s,a) = \mathbb{E}_{(\tilde{s},\tilde{a})\sim p_{h(s,a)}}[P(s' \mid \tilde{s},\tilde{a})]$.

We use Γ to denote the Bellman Optimality Operator. For simplicity, we use the subscript to distinguish the value functions (V^*), the state-action value functions (Q^*), and the Bellman Optimality Operators (Γ) of the three MDPs M, M_ψ and M'_ψ. Notice that Γ_M is identical with F.

We can show that θ^* is identical with the state-action value function of M_ψ, i.e.,

$$\theta^* = Q^*_{M_\psi}. \tag{6.64}$$

To see this, we notice that $(\Phi\theta^*)_{s,a} = \theta^*_{h(s,a)}$. Hence we get that

$$\begin{aligned} F(\Phi\theta^*)_{s,a} &= [\Gamma_M \Phi\theta^*]_{s,a} \\ &= R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\left[\max_a (\Phi\theta^*)_{s',a}\right] \\ &= R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\left[\max_a \theta^*_{h(s',a)}\right]. \end{aligned}$$

Using this, we further obtain that

$$\begin{aligned} (\Pi F(\Phi\theta^*))_{x,y} &= \sum_{(s,a)\in h^{-1}(x,y)} \frac{d_{s,a}}{\sum_{(\tilde{s},\tilde{a})\in h^{-1}(x,y)} d_{\tilde{s},\tilde{a}}} \left(R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\left[\max_a \theta^*_{h(s',a)}\right]\right) \\ &= \sum_{(s,a)\in h^{-1}(x,y)} p_{(x,y)}(s,a) \left(R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\left[\max_a \theta^*_{h(s',a)}\right]\right) \end{aligned}$$

$$= (R_\psi)_{x,y} + \sum_{(s,a)\in h^{-1}(x,y)} p_{(x,y)}(s,a) \sum_{x'\in\psi_1(\mathcal{S})} P(x' \mid s,a) \max_a \theta^*_{x',\psi_2(a)}$$
$$= (R_\psi)_{x,y} + \sum_{x'\in\psi_1(\mathcal{S})} P_\psi(x' \mid x,y) \max_{y'} \theta^*_{x',y'}$$
$$= [\Gamma_{M_\psi}\theta^*]_{x,y}.$$

Since we have $\Pi F(\Phi\theta^*) = \theta^*$ by definition, we see that

$$[\Gamma_{M_\psi}\theta^*]_{x,y} = \theta^*_{x,y}, \forall (x,y) \in \mathcal{M}.$$

Thus we have shown that $\theta^* = Q^*_{M_\psi}$.

Next, we observe that the state-value function of MDP M'_ψ is given by

$$Q^*_{M'_\psi} = \Phi Q^*_{M_\psi}. \tag{6.65}$$

This is because

$$\left(\Gamma_{M'_\psi}(\Phi Q^*_{M_\psi})\right)_{s,a} = (R'_\psi)_{s,a} + \gamma \sum_{s'\in\mathcal{S}} P'_\psi(s' \mid s,a) \max_{a'} (\Phi Q^*_{M_\psi})_{s',a'}$$
$$= (R'_\psi)_{s,a} + \gamma\langle P'_\psi(s,a), \Phi V^*_{M_\psi}\rangle$$
$$= \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a}) \left(R_{\tilde{s},\tilde{a}} + \gamma\langle P(\tilde{s},\tilde{a}), \Phi V^*_{M_\psi}\rangle\right) \tag{6.66a}$$
$$= \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a}) R_{\tilde{s},\tilde{a}}$$
$$+ \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})\gamma\langle P(\tilde{s},\tilde{a}), \Phi V^*_{M_\psi}\rangle$$
$$= (R_\psi)_{h(s,a)} + \gamma\langle P_\psi(h(s,a)), V^*_{M_\psi}\rangle \tag{6.66b}$$
$$= (Q^*_{M_\psi})_{h(s,a)}$$
$$= (\Phi Q^*_{M_\psi})_{s,a},$$

where we use the definition of M'_ψ (see (6.63)) in (6.66a); we use the definition of M_ψ (see (6.62)) in (6.66b).

By (6.65), we see that

$$\left\|\Phi Q^*_{M_\psi} - Q^*_M\right\|_\infty = \left\|Q^*_{M'_\psi} - Q^*_M\right\|_\infty \le \frac{1}{1-\gamma}\left\|\Gamma_{M'_\psi} Q^*_M - Q^*_M\right\|_\infty. \tag{6.67}$$

We further notice that

$$\left|(\Gamma^*_{M_\psi} Q^*_M)_{s,a} - (Q^*_M)_{s,a}\right|$$

$$
\begin{aligned}
&= \left|(R'_\psi)_{s,a} + \gamma\langle P_\psi(s,a), V_M^*\rangle - (Q_M^*)_{s,a}\right| \\
&= \left|\left(\sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})(R_{\tilde{s},\tilde{a}} + \gamma\langle P(\tilde{s},\tilde{a}), V_M^*\rangle)\right) - (Q_M^*)_{s,a}\right| \qquad (6.68a) \\
&= \left|\sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})\left((Q_M^*)_{\tilde{s},\tilde{a}} - (Q_M^*)_{s,a}\right)\right| \\
&\le \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})\left|(Q_M^*)_{\tilde{s},\tilde{a}} - (Q_M^*)_{s,a}\right| \\
&\le \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})(2\epsilon_{Q^*}) \qquad (6.68b) \\
&= 2\epsilon_{Q^*},
\end{aligned}
$$

where we use the definition of M_ψ in (6.68a); we use Assumption 6.B.6 in (6.68b).

Substituting (6.68) into (6.67) gives that

$$\left\|\Phi Q_{M_\psi}^* - Q_M^*\right\|_\infty \le \frac{2\epsilon_{Q^*}}{1-\gamma}. \qquad (6.69)$$

Combining (6.64) and (6.69) finishes the proof. □

Part IV

Conclusions

Chapter 7

CONCLUSIONS AND FUTURE DIRECTIONS

To leverage the groundbreaking advances on ML-based predictors in real-world online decision making tasks, it is critical to build a strong theoretical foundation for understanding why the predictions help and how to use them optimally. The goal of this thesis is to provide analytical frameworks for characterizing the benefit of using predictions in control under general prediction/dynamical models and propose efficient/scalable policy optimization algorithms with provable guarantees.

The first part of this thesis investigates the benefits of leveraging predictions in online decision making under two predictive modeling paradigms: the adversarial model and the stochastic model. Under the adversarial model, we focus on MPC-style approaches and quantify the improvements in worst-case performance metrics—such as dynamic regret and competitive ratio—relative to settings without predictions. In contrast, under the stochastic model, we analyze the structure of the optimal predictive control policy and establish sufficient conditions under which the prediction power admits meaningful lower bounds. These results collectively provide a theoretical foundation for understanding when and how predictions can improve decision-making performance across a range of problem settings.

The second part of this thesis focuses on policy optimization for general policy classes in both online settings and multi-agent networked systems—encompassing the problem of finding or tracking the optimal predictive policy as a special case. In the online setting, we consider a single-trajectory framework, where key challenges stem from time-varying environments and limited feedback. In contrast, our study of policy optimization in networked systems adopts an episodic setting, where scalability arises as the primary concern in large-scale cooperation. Together, these contributions address core practical challenges in learning effective decision policies, thereby enabling users to better realize the potential benefits of prediction in online decision making.

7.1 Summary of Chapters

In Chapter 3, we present a proof framework for MPC-style predictive control based on the perturbation analysis under an adversarial prediction model. If one can establish a exponentially decaying perturbation bound for the underlying optimal control problem, the framework helps derive finite-time performance guarantees for the MPC policy. Our results highlight the insight that, while it becomes harder to predict system parameters (e.g., disturbances) further into the future, the parameters also become less important for approximating the current optimal decision in a wide class of settings. We further extend the perturbation-based proof framework to online convex optimization in networked systems, and we demonstrate how to use the theoretical insight for algorithm design in the application of adaptive bitrate streaming.

Chapter 4 studies the prediction power under a stochastic prediction model. The goal is to characterize the benefit of leveraging "weak" predictions, whose potential be overlooked by the adversarial prediction model in Chapter 3. In a general setting, we provide sufficient conditions under which we can derive a meaningful lower bound of the prediction power. We demonstrate the effectiveness of this general lower bound by instantiating it in more specific online control problem such as LQR. For prediction power evaluation, we provide examples to show that using standard accuracy metrics like the mean-squared error is not enough, because one should also consider the specific online control problem.

Chapter 5 considers online policy optimization on a single trajectory. The theoretical foundation of our results is the contractive perturbation property, which enables us to evaluate the current policy without the need to re-simulate what would happen if we keep using the same policy from time 0. Under contractive perturbation, we design an efficient policy optimization algorithm, M-GAPS, by differentiating through the actual trajectory experienced by the controller. We first show M-GAPS can adapt quickly in changing environments with provable guarantees when the Jacobians of the dynamics are known. When the Jacobians are unknown, we propose a meta-framework that can combine an online policy optimization algorithm like M-GAPS with an online estimator of the unknown component in the dynamical model. Lastly, we demonstrate the effectiveness of M-GAPS in the application of quadcopter control, which involves multiple challenges such as nonlinear dynamics, periodic disturbances, and limited computing hardware.

In Chapter 6, we study policy optimization in the setting of MARL in networked systems. We exploit the localized interaction structure among agents to show a decay property of each agent's local Q function. As a result, each agent only needs to gather information within its κ-hop neighborhood to evaluate the current joint policy. We leverage the decay property to design a scalable actor-critic algorithm with finite-time sample complexity bounds. Our results also reveal a trade-off between the observation radius κ and the sub-optimality gap of the learned policy. We test the proposed algorithm in the applications of wireless networks and spreading networks.

7.2 Future Directions

The results presented in this thesis on prediction power characterization and online policy optimization focus on settings with continuous state and action spaces under controllable dynamics, where stability is not the primary concern. The analysis further relies on structural properties such as exponentially decaying or contractive perturbations, which facilitate both theoretical analysis and algorithm design. However, these properties may not hold in many real-world systems, which often involve discontinuities, partial controllability, or complex dynamics. Extending the theoretical framework to address such complexities represents an important and challenging future direction in general.

In the following, we outline concrete challenges that aim to extend the scope and applicability of our results on prediction power and policy optimization, respectively.

Prediction power. There are multiple interesting future directions based on our results for prediction power in Part II. First, we have studied how to evaluate the prediction power of a certain prediction sequence in Chapter 4, and a natural next step is to develop an end-to-end approach to select the predictive model that works best for a specific online control task. Second, the sufficient conditions for characterizing the prediction power in Theorem 4.2.3 relies on the properties of the optimal predictive policy, which could be challenging to verify beyond linear dynamics and quadratic costs. It is desirable to further relax these conditions to be based on an arbitrary predictive policy that may be sub-optimal (e.g., MPC). Third, our networked online convex optimization setting in Chapter 3 does not consider the underlying states and dynamics. Given the recent advances on decentralized control (Shin, Lin, et al., 2023; Zhang, Li, and Li, 2023), it is promising to generalize our setting and perturbation analysis to allow such dynamical constraints, where a node's next local state can be affected by its direct neighbors' current state.

Policy optimization. There are many future directions for improving our results on policy optimization in Part III. First, our theoretical guarantees for M-GAPS rely on the contractive perturbation as well as other properties such as the stability. As we have seen in the application of quadcopter control, the theoretical verification of such properties becomes challenging as the dynamics/policy class get more complicated. Inspired by the approach of Li, Preiss, et al., 2023, a potential solution could be monitoring whether the desired properties holds online and eliminate a policy parameter if a property is violated. This is an interesting direction for the future research. Second, M-GAPS is model-based because it requires the (exact or approximate) Jacobians of the dynamics to compute the gradients of the surrogate costs. Zeroth-order gradient estimation provides a model-free way for evaluating the gradients. Potential combination of zeroth-order optimization and M-GAPS is a direction that worth exploring. Finally, a limitation of our results for networked MARL in Chapter 6 is that the underlying network $\mathcal{G}$ which defines the distance metric between any two nodes must be fixed, so the κ-hop neighborhood of each node cannot change throughout the learning process. However, in real-world applications, the neighborhood relationships may change, for example, as each user moves around different locations in the wireless network example in Section 6.4. Therefore, studying more general time-varying networks is an interesting future direction.

BIBLIOGRAPHY

Agarwal, Naman et al. (2019). "Online control with adversarial disturbances." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 111–119. URL: `https://proceedings.mlr.press/v97/agarwal19c.html`.

Amos, Brandon et al. (2018). "Differentiable MPC for end-to-end planning and control." In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper_files/paper/2018/file/ba6d843eb4251a4526ce65d1807a9309-Paper.pdf`.

Anava, Oren, Elad Hazan, and Shie Mannor (2015). "Online learning for adversaries with memory: Price of past mistakes." In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper_files/paper/2015/file/38913e1d6a7b94cb0f55994f679f5956-Paper.pdf`.

Argue, C.J., Anupam Gupta, and Guru Guruganesh (2020). "Dimension-free bounds for chasing convex functions." In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 219–241. URL: `https://proceedings.mlr.press/v125/argue20a.html`.

Astrom, Karl Johan and Richard M. Murray (2008). *Feedback systems: An introduction for scientists and engineers*. Princeton University Press. ISBN: 0691135762.

Bansal, Nikhil and Anupam Gupta (2019). "Potential-function proofs for gradient methods." In: *Theory of Computing* 15.4, pp. 1–32. DOI: `10.4086/toc.2019.v015a004`. URL: `https://theoryofcomputing.org/articles/v015a004`.

Bayer, Florian, Mathias Bürger, and Frank Allgöwer (2013). "Discrete-time incremental ISS: A framework for robust NMPC." In: *2013 European Control Conference (ECC)*. IEEE, pp. 2068–2073. DOI: `10.23919/ECC.2013.6669322`.

Beck, Amir (2017). *First-order methods in optimization*. SIAM-Society for Industrial and Applied Mathematics. ISBN: 1611974984.

Bertsekas, Dimitri P. (2007). *Dynamic programming and optimal control, Vol. II*. 3rd. Athena Scientific. ISBN: 1886529302.

Bertsekas, Dimitri P. and John N. Tsitsiklis (1996). *Neuro-dynamic programming*. 1st. Athena Scientific. ISBN: 1886529108.

Bhandari, Jalaj, Daniel Russo, and Raghav Singal (2018). “A finite time analysis of temporal difference learning with linear function approximation.” In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 1691–1692. URL: https://proceedings.mlr.press/v75/bhandari18a.html.

Bothra, Chandan et al. (2023). “Veritas: Answering causal queries from video streaming traces.” In: *Proceedings of the ACM SIGCOMM 2023 Conference*. ACM SIGCOMM ’23. New York, NY, USA, pp. 738–753. DOI: 10.1145/3603269.3604828. URL: https://doi.org/10.1145/3603269.3604828.

Brémaud, Pierre (2001). *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Vol. 31. Springer Science & Business Media.

Candogan, Ozan, Kostas Bimpikis, and Asuman Ozdaglar (2012). “Optimal pricing in networks with externalities.” In: *Operations Research* 60.4, pp. 883–905.

Chen, Niangjun, Anish Agarwal, et al. (2015). “Online convex optimization using predictions.” In: *ACM SIGMETRICS Performance Evaluation Review* 43.1, pp. 191–204. DOI: 10.1145/2796314.2745854.

Chen, Niangjun, Joshua Comden, et al. (2016). “Using predictions in online optimization: Looking forward with an eye on the past.” In: *ACM SIGMETRICS Performance Evaluation Review* 44.1, pp. 193–206. DOI: 10.1145/2896377.2901464.

Chen, Niangjun, Gautam Goel, and Adam Wierman (2018). “Smoothed online convex optimization in high dimensions via online balanced descent.” In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 1574–1594. URL: https://proceedings.mlr.press/v75/chen18b.html.

Chen, Tianyu, Yiheng Lin, et al. (2024). “SODA: An adaptive bitrate controller for consistent high-quality video streaming.” In: *Proceedings of the ACM SIGCOMM 2024 Conference*, pp. 613–644. DOI: 10.1145/3651890.3672260.

Chen, Xinyi and Elad Hazan (2021). “Black-box control for linear dynamical systems.” In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 1114–1143. URL: https://proceedings.mlr.press/v134/chen21c.html.

Cheng, Sheng et al. (2024). “DiffTune: Autotuning through autodifferentiation.” In: *IEEE Transactions on Robotics* 40, pp. 4085–4101. DOI: 10.1109/TRO.2024.3429191.

Dann, Christoph et al. (2018). “On oracle-efficient PAC RL with rich observations.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/5f0f5e5f33945135b874349cfbed4fb9-Paper.pdf.

Dean, Sarah et al. (2020). “On the sample complexity of the linear quadratic regulator.” In: *Foundations of Computational Mathematics* 20.4, pp. 633–679. DOI: 10.1007/s10208-019-09426-y.

Demko, Stephen, William F. Moss, and Philip W. Smith (1984). “Decay rates for inverses of band matrices.” In: *Mathematics of Computation* 43.168, pp. 491–499. ISSN: 00255718, 10886842. URL: http://www.jstor.org/stable/2008290.

Donti, Priya L., Brandon Amos, and J. Zico Kolter (2017). “Task-based end-to-end model learning in stochastic optimization.” In: *Advances in Neural Information Processing Systems*. Vol. 30. Long Beach, CA, USA: Curran Associates, Inc. URL: http://papers.nips.cc/paper/7132-task-based-end-to-end-model-learning-in-stochastic-optimization.

Easley, David, Jon Kleinberg, et al. (2012). “Networks, crowds, and markets: Reasoning about a highly connected world.” In: *Significance* 9, pp. 43–44.

Elmachtoub, Adam N. and Paul Grigas (2022). “Smart “Predict, then optimize”.” In: *Management Science* 68.1, pp. 9–26. ISSN: 0025-1909. DOI: 10.1287/mnsc.2020.3922. URL: https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3922.

Ferguson, Joel et al. (2020). “Matched disturbance rejection for a class of nonlinear systems.” In: *IEEE Transactions on Automatic Control* 65.4, pp. 1710–1715. DOI: 10.1109/TAC.2019.2933398. URL: https://ieeexplore.ieee.org/document/8788577.

Forsgren, Anders, Philip E. Gill, and Margaret H. Wright (2002). “Interior methods for nonlinear optimization.” In: *SIAM Review* 44.4, pp. 525–597. DOI: 10.1137/S0036144502414942. eprint: https://doi.org/10.1137/S0036144502414942. URL: https://doi.org/10.1137/S0036144502414942.

García, Carlos E., David M. Prett, and Manfred Morari (1989). “Model predictive control: Theory and practice—A survey.” In: *Automatica* 25.3, pp. 335–348. ISSN: 0005-1098. DOI: https://doi.org/10.1016/0005-1098(89)90002-2. URL: https://www.sciencedirect.com/science/article/pii/0005109889900022.

Garofalo, Gianluca, Christian Ott, and Alin Albu-Schäffer (2012). “Walking control of fully actuated robots based on the Bipedal SLIP model.” In: *2012 IEEE International Conference on Robotics and Automation*, pp. 1456–1463. DOI: 10.1109/ICRA.2012.6225272. URL: https://ieeexplore.ieee.org/document/6225272.

Ghadimi, Saeed and Guanghui Lan (2013). "Stochastic first-and zeroth-order methods for nonconvex stochastic programming." In: *SIAM Journal on Optimization* 23.4, pp. 2341–2368.

Goel, Gautam, Yiheng Lin, et al. (2019). "Beyond online balanced descent: An optimal algorithm for smoothed online optimization." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper_files/paper/2019/file/9f36407ead0629fc166f14dde7970f68-Paper.pdf`.

Goel, Gautam and Adam Wierman (2019). "An online algorithm for smoothed regression and LQR control." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 2504–2513. URL: `https://proceedings.mlr.press/v89/goel19a.html`.

Gradu, Paula, Elad Hazan, and Edgar Minasyan (2023). "Adaptive regret for control of time-varying dynamics." In: *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*. Ed. by Nikolai Matni, Manfred Morari, and George J. Pappas. Vol. 211. Proceedings of Machine Learning Research. PMLR, pp. 560–572. URL: `https://proceedings.mlr.press/v211/gradu23a.html`.

Hazan, Elad (2016). "Introduction to online convex optimization." In: *Foundations and Trends® in Optimization* 2.3-4, pp. 157–325. ISSN: 2167-3888. DOI: `10.1561/2400000013`. URL: `http://dx.doi.org/10.1561/2400000013`.

Hazan, Elad, Sham Kakade, and Karan Singh (2020). "The nonstochastic control problem." In: *Proceedings of the 31st International Conference on Algorithmic Learning Theory*. Ed. by Aryeh Kontorovich and Gergely Neu. Vol. 117. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 408–421. URL: `https://proceedings.mlr.press/v117/hazan20a.html`.

Hazan, Elad and Comandur Seshadhri (2007). "Adaptive algorithms for online decision problems." In: *Electronic colloquium on computational complexity (ECCC)*. Vol. 14. 088.

Hazan, Elad and Karan Singh (2022). "Introduction to online nonstochastic control." In: *arXiv preprint arXiv:2211.09619*.

Hazan, Elad, Karan Singh, and Cyril Zhang (2017). "Efficient regret minimization in non-convex games." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 1433–1441. URL: `https://proceedings.mlr.press/v70/hazan17a.html`.

Jiang, Nan (2018). *Notes on State Abstractions*. http://nanjiang.web.engr.illinois.edu/files/cs598/note4.pdf.

Jiang, Nan, Alex Kulesza, and Satinder Singh (2015). "Abstraction selection in model-based reinforcement learning." In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: Proceedings of Machine Learning Research, pp. 179–188. URL: https://proceedings.mlr.press/v37/jiang15.html.

Jong, Nicholas K. and Peter Stone (2005). "State abstraction discovery from irrelevant state variables." In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*. IJCAI'05. Edinburgh, Scotland: Morgan Kaufmann Publishers Inc., pp. 752–757.

Kakade, Sham and John Langford (2002). "Approximately optimal approximate reinforcement learning." In: *Proceedings of the Nineteenth International Conference on Machine Learning*. ICML '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 267–274. ISBN: 1558608737.

Kempe, David, Jon Kleinberg, and Éva Tardos (2015). "Maximizing the spread of influence through a social network." In: *Theory of Computing* 11. ISSN: 15572862. DOI: 10.4086/toc.2015.v011a004.

Kumar, Raunak, Sarah Dean, and Robert Kleinberg (2023). "Online convex optimization with unbounded memory." In: *Advances in Neural Information Processing Systems* 36, pp. 26229–26270.

Li, Lihong, Thomas J. Walsh, and Michael L. Littman (2006). "Towards a unified theory of state abstraction for MDPs." In: *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*.

Li, Tongxin, Ruixiao Yang, Guannan Qu, Yiheng Lin, et al. (2023). "Certifying Black-Box Policies With Stability for Nonlinear Control." In: *IEEE Open Journal of Control Systems* 2, pp. 49–62. DOI: 10.1109/OJCSYS.2023.3241486.

Li, Tongxin, Ruixiao Yang, Guannan Qu, Guanya Shi, et al. (2022). "Robustness and consistency in linear quadratic control with untrusted predictions." In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 6.1, pp. 1–35. DOI: 10.1145/3508038. URL: https://doi.org/10.1145/3508038.

Li, Yingying, Xin Chen, and Na Li (2019). "Online optimal control with linear dynamics and predictions: Algorithms and regret analysis." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/6d1e481bdcf159961818823e652a7725-Paper.pdf.

Li, Yingying, James A. Preiss, et al. (2023). "Online switching control with stability and regret guarantees." In: *Learning for Dynamics and Control Conference*. Proceedings of Machine Learning Research, pp. 1138–1151. URL: https://proceedings.mlr.press/v211/li23a/li23a.pdf.

Li, Yingying, Guannan Qu, and Na Li (2018). “Using predictions in online optimization with switching costs: A fast algorithm and a fundamental limit.” In: *2018 Annual American Control Conference (ACC)*. IEEE, pp. 3008–3013.

– (2021). “Online optimization with predictions and switching costs: Fast algorithms and the fundamental limit.” In: *IEEE Transactions on Automatic Control* 66.10, pp. 4761–4768. DOI: 10.1109/TAC.2020.3040249.

Lin, Yiheng, Zaiwei Chen, et al. (2025). “Maximizing the value of stochastic predictions in control: Accuracy is not enough.” In: *Under submission*.

Lin, Yiheng, Judy Gan, et al. (2022). “Decentralized online convex optimization in networked systems.” In: *International Conference on Machine Learning*. Proceedings of Machine Learning Research, pp. 13356–13393. URL: https://proceedings.mlr.press/v162/lin22c/lin22c.pdf.

Lin, Yiheng, Gautam Goel, and Adam Wierman (2020). “Online optimization with predictions and non-convex losses.” In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 4.1, pp. 1–32. DOI: 10.1145/3393691.3394208.

Lin, Yiheng, Yang Hu, Guannan Qu, et al. (2022). “Bounded-regret MPC via perturbation analysis: Prediction error, constraints, and nonlinearity.” In: *Advances in Neural Information Processing Systems* 35, pp. 36174–36187. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/eadeef7c51ad86989cc3b311cb49ec89-Paper-Conference.pdf.

Lin, Yiheng, Yang Hu, Guanya Shi, et al. (2021). “Perturbation-based regret analysis of predictive control in linear time varying systems.” In: *Advances in Neural Information Processing Systems* 34, pp. 5174–5185. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/298f587406c914fad5373bb689300433-Paper.pdf.

Lin, Yiheng, James A. Preiss, Emile Anand, et al. (2023). “Online adaptive policy selection in time-varying systems: No-regret via contractive perturbations.” In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., pp. 53508–53521. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/a7a7180fe7f82ff98eee0827c5e9c141-Paper-Conference.pdf.

Lin, Yiheng, James A. Preiss, Fengze Xie, et al. (2024). “Online policy optimization in unknown nonlinear systems.” In: *The Thirty Seventh Annual Conference on Learning Theory*. Proceedings of Machine Learning Research, pp. 3475–3522. URL: https://proceedings.mlr.press/v247/lin24a.html.

Lin, Yiheng, Guannan Qu, et al. (2021). “Multi-agent reinforcement learning in stochastic networked systems.” In: *Advances in Neural Information Processing Systems* 34, pp. 7825–7837. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/412604be30f701b1b1e3124c252065e6-Paper.pdf.

Mandi, Jayanta et al. (2024). "Decision-focused learning: Foundations, state of the art, benchmark and future opportunities." In: *Journal of Artificial Intelligence Research* 80, pp. 1623–1701. ISSN: 1076-9757. DOI: 10.1613/jair.1.15320. URL: https://www.jair.org/index.php/jair/article/view/15320.

Mania, Horia, Stephen Tu, and Benjamin Recht (2019). "Certainty equivalence is efficient for linear quadratic control." In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/5dbc8390f17e019d300d5a162c3ce3bc-Paper.pdf.

Marsden, Jerrold E. and Anthony Tromba (2003). *Vector calculus*. 5th. W. H. Freeman and Company. ISBN: 9780716749929.

Mokhtari, Aryan et al. (2016). "Online optimization in dynamic environments: Improved regret rates for strongly convex problems." In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, pp. 7195–7201.

Na, Sen and Mihai Anitescu (2022). "Superconvergence of online optimization for model predictive control." In: *IEEE Transactions on Automatic Control* 68.3, pp. 1383–1398.

O'Connell, Michael et al. (2022). "Neural-fly enables rapid learning for agile flight in strong winds." In: *Science Robotics* 7.66, eabm6597.

Preiss, James A. et al. (2025). "Fast non-episodic adaptive tuning of robot controllers with model-based online policy optimization." In: *Under submission*.

Qu, Guannan, Yiheng Lin, et al. (2020). "Scalable multi-agent reinforcement learning for networked systems with average reward." In: *Advances in Neural Information Processing Systems* 33, pp. 2074–2086. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/168efc366c449fab9c2843e9b54e2a18-Paper.pdf.

Qu, Guannan and Adam Wierman (2020). "Finite-time analysis of asynchronous stochastic approximation and Q-learning." In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 3185–3205. URL: https://proceedings.mlr.press/v125/qu20a.html.

Qu, Guannan, Adam Wierman, and Na Li (2020). "Scalable reinforcement learning of localized policies for multi-agent networked systems." In: *Proceedings of the 2nd Conference on Learning for Dynamics and Control*. Ed. by Alexandre M. Bayen et al. Vol. 120. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 256–266. URL: https://proceedings.mlr.press/v120/qu20a.html.

Qu, Guannan, Chenkai Yu, et al. (2021). "Exploiting linear models for model-free nonlinear control: A provably convergent policy gradient approach." In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, pp. 6539–6546.

Rajappa, Sujit et al. (2015). “Modeling, control and design optimization for a fully-actuated hexarotor aerial vehicle with tilted propellers.” In: *IEEE International Conference on Robotics and Automation, ICRA 2015, Seattle, WA, USA, 26-30 May, 2015*. IEEE, pp. 4006–4013. DOI: 10.1109/ICRA.2015.7139759. URL: https://doi.org/10.1109/ICRA.2015.7139759.

Roberts, Lawrence G (1975). “ALOHA packet system with and without slots and capture.” In: *ACM SIGCOMM Computer Communication Review* 5.2, pp. 28–42.

Roughgarden, Tim (2020). “Resource Augmentation.” In: *arXiv preprint*. URL: https://arxiv.org/abs/2007.13234.

Ruhi, Navid Azizan, Christos Thrampoulidis, and Babak Hassibi (2016). “Improved bounds on the epidemic threshold of exact SIS models on complex networks.” In: *2016 IEEE 55th Conference on Decision and Control (CDC)*. IEEE, pp. 3560–3565.

Rutten, Daan et al. (2023). “Smoothed online optimization with unreliable predictions.” In: *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7.1, pp. 1–36.

Schneider, Rolf (2014). *Convex bodies: the Brunn–Minkowski theory*. Cambridge University Press.

Shah, Devavrat and Qiaomin Xie (2018). “Q-learning with nearest neighbors.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/309fee4e541e51de2e41f21bebb342aa-Paper.pdf.

Shi, Guanya et al. (2020). “Online optimization with memory and competitive control.” In: *Advances in Neural Information Processing Systems*, pp. 20636–20647. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/ed46558a56a4a26b96a68738a0d28273-Paper.pdf.

Shin, Sungho, Mihai Anitescu, and Victor M. Zavala (2022). “Exponential decay of sensitivity in graph-structured nonlinear programs.” In: *SIAM Journal on Optimization* 32.2, pp. 1156–1183.

Shin, Sungho, Yiheng Lin, et al. (2023). “Near-optimal distributed linear-quadratic regulator for networked systems.” In: *SIAM Journal on Control and Optimization* 61.3, pp. 1113–1135. DOI: 10.1137/22M1489836.

Shin, Sungho and Victor M. Zavala (2021). “Controllability and observability imply exponential decay of sensitivity in dynamic optimization.” In: *IFAC-PapersOnLine* 54.6. 7th IFAC Conference on Nonlinear Model Predictive Control NMPC 2021, pp. 179–184. ISSN: 2405-8963. DOI: https://doi.org/10.1016/j.ifacol.2021.08.542. URL: https://www.sciencedirect.com/science/article/pii/S2405896321013173.

Shin, Sungho, Victor M. Zavala, and Mihai Anitescu (2020). “Decentralized schemes with overlap for solving graph-structured optimization problems.” In: *IEEE Transactions on Control of Network Systems* 7.3, pp. 1225–1236. ISSN: 2372-2533. DOI: 10.1109/tcns.2020.2967805. URL: http://dx.doi.org/10.1109/TCNS.2020.2967805.

Siciliano, Bruno et al. (2008). *Robotics: Modelling, planning and control*. Springer. ISBN: 1846286417. URL: https://link.springer.com/book/10.1007/978-1-84628-642-1.

Simchowitz, Max, Karan Singh, and Elad Hazan (2020). “Improper learning for non-stochastic control.” In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 3320–3436. URL: https://proceedings.mlr.press/v125/simchowitz20a.html.

Singh, Satinder P., Tommi Jaakkola, and Michael I. Jordan (1995). “Reinforcement learning with soft state aggregation.” In: *Advances in neural information processing systems*, pp. 361–368.

Sinha, Rohan et al. (2022). “Adaptive robust model predictive control with matched and unmatched uncertainty.” In: *2022 American Control Conference (ACC)*, pp. 906–913. DOI: 10.23919/ACC53348.2022.9867457. URL: https://ieeexplore.ieee.org/document/9867457.

Slotine, Jean-Jacques E., Weiping Li, et al. (1991). *Applied nonlinear control*. Vol. 199. 1. Prentice hall Englewood Cliffs, NJ.

Srikant, R. and Lei Ying (2019). “Finite-time error bounds for linear stochastic approximation and TD learning.” In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by Alina Beygelzimer and Daniel Hsu. Vol. 99. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 2803–2830. URL: https://proceedings.mlr.press/v99/srikant19a.html.

Stanica, Pantelimon (2001). “Good lower and upper bounds on binomial coefficients.” In: *Journal of Inequalities in Pure and Applied Mathematics* 2.3, p. 30.

Suggala, Arun Sai and Praneeth Netrapalli (2020). “Online non-convex learning: Following the perturbed leader is optimal.” In: *Proceedings of the 31st International Conference on Algorithmic Learning Theory*. Ed. by Aryeh Kontorovich and Gergely Neu. Vol. 117. Proceedings of Machine Learning Research. Proceedings of Machine Learning Research, pp. 845–861. URL: https://proceedings.mlr.press/v117/suggala20a.html.

Sunehag, Peter et al. (2018). “Value-decomposition networks for cooperative multiagent learning based on team reward.” In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. AAMAS ’18. Stockholm, Sweden: International Foundation for Autonomous Agents and Multiagent Systems, pp. 2085–2087.

Sutton, Richard S. et al. (1999). “Policy gradient methods for reinforcement learning with function approximation.” In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*. NIPS’99. Denver, CO: MIT Press, pp. 1057–1063.

Tretter, Christiane (2008). “Spectral theory of block operator matrices and applications.” In: *Spectral Theory of Block Operator Matrices and Applications*. DOI: 10.1142/P493.

Tsitsiklis, John and Benjamin Van Roy (1996). “Analysis of temporal-diffference learning with function approximation.” In: *Advances in Neural Information Processing Systems*. Ed. by M.C. Mozer, M. Jordan, and T. Petsche. Vol. 9. MIT Press, pp. 1075–1081. URL: https://proceedings.neurips.cc/paper_files/paper/1996/file/e00406144c1e7e35240afed70f34166a-Paper.pdf.

Tsitsiklis, John N. (1994). “Asynchronous stochastic approximation and Q-learning.” In: *Machine learning* 16.3, pp. 185–202.

Tsitsiklis, John N. and Benjamin Van Roy (1997). “An analysis of temporal-difference learning with function approximation.” In: *IEEE Transactions on Automatic Control* 42.5, pp. 674–690.

Tsukamoto, Hiroyasu, Soon-Jo Chung, and Jean-Jaques E. Slotine (2021). “Contraction theory for nonlinear stability analysis and learning-based control: A tutorial overview.” In: *Annual Reviews in Control* 52, pp. 135–169.

Wainwright, Martin J (2019). “Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning.” In: *arXiv preprint arXiv:1905.06265*.

Wang, Irina et al. (2023). “Learning decision-focused uncertainty sets in robust optimization.” In: *arXiv preprint arXiv:2305.19225*.

Xu, Wanting and Mihai Anitescu (2019). “Exponentially convergent receding horizon strategy for constrained optimal control.” In: *Vietnam Journal of Mathematics* 47.4, pp. 897–929.

Yan, Francis Y. et al. (2020). “Learning in Situ: A randomized experiment in video streaming.” In: *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*, pp. 495–511. ISBN: 9781939133137. URL: https://www.usenix.org/conference/nsdi20/presentation/yan.

Yeh, Christopher et al. (2024). “End-to-end conformal calibration for optimization under uncertainty.” In: *arXiv preprint arXiv:2409.20534*.

Yin, Xiaoqi et al. (2015). “A control-theoretic approach for dynamic adaptive video streaming over HTTP.” In: *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*. New York, NY, USA: ACM, pp. 325–338. ISBN: 9781450335423. DOI: 10.1145/2785956.2787486. URL: https://dl.acm.org/doi/10.1145/2785956.2787486.

Yu, Chenkai et al. (2020). “The power of predictions in online control.” In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 1994–2004. URL: `https://proceedings.neurips.cc/paper_files/paper/2020/file/155fa09596c7e18e50b58eb7e0c6ccb4-Paper.pdf`.

– (2022). “Competitive control with delayed imperfect information.” In: *2022 American Control Conference (ACC)*. IEEE, pp. 2604–2610.

Zhang, Runyu, Yingying Li, and Na Li (2021). “On the regret analysis of online LQR control with predictions.” In: *2021 American Control Conference (ACC)*. IEEE, pp. 697–703.

Zhang, Runyu Cathy, Weiyu Li, and Na Li (2023). “On the optimal control of network LQR with spatially-exponential decaying structure.” In: *2023 American Control Conference (ACC)*, pp. 1775–1780. DOI: `10.23919/ACC55779.2023.10156208`.

Zhang, Yan, Yi Zhou, et al. (2024). “Boosting one-point derivative-free online optimization via residual feedback.” In: *IEEE Transactions on Automatic Control* 69.9, pp. 6309–6316. DOI: `10.1109/TAC.2024.3382358`.

Zheng, Peter et al. (2020). “TiltDrone: A fully-actuated tilting quadrotor platform.” In: *IEEE Robotics and Automation Letters* 5.4, pp. 6845–6852. DOI: `10.1109/LRA.2020.3010460`. URL: `https://ieeexplore.ieee.org/document/9144371`.

Zocca, Alessandro (2019). “Temporal starvation in multi-channel CSMA networks: An analytical framework.” In: *Queueing Systems* 91.3-4, pp. 241–263.