Deep Stochastic Control for Regime-dependent Utility Maximization Problems

Thesis by
Harshkooshal Kamlesh Gandhi

In Partial Fulfillment of the Requirements for the
Degree of
PhD

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended 20th May 2026

Harshkooshal Kamlesh Gandhi
ORCID: 0000-0002-4624-2406

ACKNOWLEDGEMENTS

Some debts can never be repaid, and some forms of gratitude can never be fully expressed, but I will try to do some form of justice to the people I owe in this section. Firstly, my time at Caltech and academia wouldn't be possible had it not been for my mother, Sri Laxmi Durbha, whose push for me to challenge boundaries of my capabilities. I can't recollect a single moment in my 26 years where she had a doubt in my ability. Over the years, my experience at Caltech had its ups and downs, and without her words of affirmation, her support, her patience, her tolerance and her blessings, I don't think I would have been able to finish my thesis or my PhD nor even come to Caltech without her sacrifices. Her presence in my life is a debt I can never repay. I also want to dedicate this work to my late grandmother Vijay Laxmi Durbha, whose blessing and strength has guided me over the years and I hope I have made her proud.

My advisor Jakša Cvitanić saw potential in me in 2022 when no one else did and I will always remember the first meeting I had with him in his office in Baxter. His mentorship, insight, guidance and his patience shaped my research and I will always cherish the meetings I had in his office. I would also like to express my utmost gratitude my committee chair Prof. Yaser Abu-Mostafa, whose advice and conversations were vital in my research and my PhD. Furthermore, it was my privilege to have Prof. Steven Low and Prof. John Doyle in my candidacy committee and their feedback was instrumental in refining my research. I am also very grateful to Prof. Pietro Perona and Prof. Steven Low for being part of my defense.

I spent a large chunk of my time since fall of 2022 TAing classes for funding my PhD. Over the years, Prof. Yaser Abu-Mostafa, Prof. Konstantin Zuev, Prof. Houman Owhadi, Prof. Jakša Cvitanić and Prof. Glen George were very kind to give me this responsibility and I will always be grateful to them. Every term had a moment where my funding for my PhD was uncertain and had it not been for them, I wouldn't have been able to finish my PhD. I also want to thank David Chan and Darrell Peterson at the Dean's office who in 2023 allowed me to fund my PhD via TAing. I found TAing not an obligation but one of the most rewarding experiences. I was fortunate to be given the opportunity head TA almost all of the classes that I was involved with over the years, and I had the opportunity to interact and get to know a lot of students and my Co-TAs in the process.

Other than my research and TAing, my most memorable time at Caltech was every Thursday night at 9pm Student Investment Fund (SIF) meetings. SIF was my excuse to take a break from everything. At SIF, I met a lot of people who went onto become very good friends of mine and relationships I intend to keep even after Caltech. I thank Matthew Hajjar (my first vote to becoming president in 2022), Sarah Hashash, Aaron Zhao, Dhruv Verma, Varun Gabbita, Aman Burman, Gautam Kappaganthula, Joseph Pieper, Arjun Pradhan, and Ali Niazi for making my time so memorable. Being the co-president of such an extraordinary club was a privilege and I will have a soft corner for any SIF member.

Over the years, I also had friends who have now become family to me, who have supported me throughout the entire journey. Without them I would have never been able to complete my thesis, explored LA, found my inner circle I enjoy being with and made memories which I hold on dearly throughout my life. I thank the Matthew Hajjar, Grace Hajjar, the entire Hajjar family, Perry (the Platypus) Samimy, Alex (T-rex) Bozorgi, Tyler (T) Fox, Helen Siavalis, Gautham Chawla, Kristina Stoyanova, Ian Fowler, Arjun Pradhan, Ali Niazi and Anirudh Gajula. Everyone had a very dear place in my heart and played their role and even missing one of them in this process would have changed my time at Caltech.

Finally, I will be forever indebted to my partner Marilyn Recarte. She was my support system, who never lost faith in me in my most darkest time and who helped me through every obstacle, every uncertainty and every setback over the years. Had it not been for her emotional support and cushion, I would have never been able to weather through the marathon that I call my PhD.

This PhD is dedicated to everyone who has touched my life over the years, for teaching me valuable lessons which I carry forward.

ABSTRACT

We develop a unified framework for solving utility maximization problems (UMPs) in continuous-time stochastic control under high-dimensional, partially observed, and path-dependent uncertainty. Such settings arise in quantitative and engineering control problems where classical dynamic programming breaks down due to curse of dimensionality and non-Markovian structure. Our approach combines forward–backward stochastic differential equations (FBSDEs) with deep learning to construct scalable, simulation-based algorithms operating directly on trajectories. A central principle is the interplay between primal and dual formulations of stochastic control. We show that the duality gap, traditionally a post-hoc diagnostic, can be elevated to a principled driver of learning, yielding stable training and computable near-optimality. In regime-switching models, dynamic programming, convex duality and FBSDE representations yield a complete analytical and computational characterization. The framework extends to non-Markovian settings via filtering under latent regime uncertainty and functional Itô calculus for long-memory dynamics, with transformer-based architectures enabling efficient computation directly from trajectories. Structural correctness and adaptability is confirmed on classical engineering problems including the linear-quadratic regulator. Finally, we introduce a robust Minmax formulation under model ambiguity delivers algorithms that jointly learn optimal strategies and worst-case models. Overall, this work provides a unified, scalable approach to stochastic control in complex environments, bridging together modern machine learning and classical control theory for high-dimensional decision-making.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Background: Classical Utility Maximization Problems

Utility maximization problems (UMPs) play a central role in continuous-time stochastic control and mathematical finance. Ever since the work of Merton (Merton, 1971), UMP has provided for a principled framework for modeling optimal decision-making under uncertainty, with applications ranging from portfolio selection, consumption planning to insurance, risk management and macroeconomic policy. More recently, growing awareness of model uncertainty in financial markets has motivated a further extension of the classical UMP framework to *robust* utility maximization, in which the investor optimizes against a worst-case probability measure drawn from an ambiguity set of plausible models, providing resilience to misspecification of the underlying market dynamics (Hansen and Sargent, 2001; Krach, Teichmann, and Wutte, 2025). In its more classical formulation, an investor controls a portfolio process so as to maximize the expected utility of terminal wealth or more generally, the expected utility derived from both consumption and terminal payoff.

From a mathematical standpoint, classical UMPs are typically solved using dynamic programming principle (DPP). Under standard assumptions (including Markovian state dynamics, sufficiently smooth coefficients, and concave utility functions), the value function associated with the control problem satisfied a Hamilton-Jacobi-Bellman (HJB) equation. Verification theorem then provide sufficiency under which a smooth solution of the HJB equation coincides with the true optimal value function and yields an optimal control policy (Pham, 2009; Karatzas and Shreve, 1998; Fleming and Soner, 2006). In highly stylized setting, this approach leads to closed-form solutions with clear economic interpretations, such as constant-proportion investment strategies in the classical Merton problem. These classical techniques rely on the collective of restrictive structural assumptions. The DPP approach fundamentally depends on the Markovian nature of the underlying state process and, smoothness of the value function along with low-dimensional state spaces. In many real world economic and other relevant settings, these assumptions are violated. State variable coefficients may be stochastic, path dependent, utility

function may induce non-smooth value functions, and the effective state dimension may be large or even infinite-dimensional. In such cases, the associated HJB may become path-dependent, and not admit a classical solution, often resulting in analytically intractable PDEs rendering verification-based approaches infeasible.

A less widely studied extension of the classical UMP framework arises when the underlying economical or market environment is allowed to evolve through distinct regimes. Empirical evidence in literature suggests that systems exhibit structural changes over time, such as shifts between expansionary and recessionary phases, volatility clustering, or change in monetary policy (Sotomayor and Cadenillas, 2009). These are commonly modeled using regime-switching dynamics, where the coefficients depend on an exogenous finite-state Markov chain. In the context of UMP, regime switching allows asset returns, volatilities, interest rates, and even preferences (utility functions) to vary across regimes, leading to optimal strategies that adapt endogenously to change in market conditions. Again, from a mathematical perspective, incorporating regime-switching dynamics fundamentally alters the structure of the UMP. The presence of an exogenous Markov chain renders the market incomplete, even when the regime process is fully observable. As a result, the HJB becomes a system of coupled partial differential equation, one for each regime linked through the generator of Markov chain. This coupling destroys the separability that underlies many classical UMP solutions and introduces additional jump-related terms into the stochastic calculus. Closed form solutions are therefore available only in highly specialized settings involving a specific utility functions (such as power or logarithmic utility), and restrictive assumptions on market dynamics (Sotomayor and Cadenillas, 2009). In general, solving the resulting coupled HJB system analytically is not feasible.

Beyond analytical intractability, regime-switching UMP also pose severe difficulties if we were to use classical numerical methods. The issue is that the coupled system of HJB equations must be solved simultaneously over the full state space, leading to a rapid growth in computational complexity as the dimension of the state variables or the number of regimes increases. Traditional numerical approaches, including finite difference methods, finite element schemes, and semi-Lagrangian methods, suffer from the curse of dimensionality, with computational cost growing exponentially in the number of state variables (Fleming and Soner, 2006; Z. Chen and Forsyth, 2008). In practice, these methods rarely scale beyond a small number of dimensions and become prohibitively expensive or unstable when applied to

realistic regime-dependent models. As a result, classical grid-based numerical solvers remain ill-suited for solving general regime-switching UMPs.

An alternative and fundamentally different approach to such UMPs is provided by probabilistic formulations based on backward stochastic differential equations (BSDEs). Since the work of Pardoux and Peng (Pardoux and S.G. Peng, 1990), BSDEs have been recognized as a powerful tool for representing solutions of nonlinear partial differential equations through stochastic processes. In the context of stochastic control, this connection leads to forward-BSDEs (FBSDE) formulations, where the forward component describes the controlled state dynamics and the backward component characterizes the value function and associated optimality conditions (El Karoui, S. Peng, and Quenez, 1997). Crucially, BSDE-based formulations avoid explicit discretizations of the state space and instead operate directly on simulated state trajectories. This removed the need for tensor product grids whose size grows exponentially with state dimension, thereby mitigating the curse of dimensionality that restricts the use of classical grid-based PDE solvers infeasible in higher-dimensional settings.

1.2 Probabilistic Reformulations via BSDEs

In the context of UMP, BSDE formulations provide a natural representation of the first-order optimality conditions associated with the control problem. Specifically, the adjoint process appearing in the backward equation captures the sensitivity of the value function with respect to the controlled state, while the optimal control is characterized through a pointwise maximization of the associated Hamiltonian. This leads to a fully coupled FBSDE, in which the forward dynamics depend on the control and the backward dynamics depend implicitly on the forward state. Unlike valuation problems, where the backward component is driven by an exogenous payoff, UMPs involve endogenous feedback between the forward and backward equations. This forward–backward coupling significantly increases both analytical and numerical complexity, but also provides a principled probabilistic alternative to HJB-based approaches.

Despite their theoretical base, FBSDE formulations of UMP are difficult to solve numerically. The backward equation is typically nonlinear, high dimensional, and coupled to the forward dynamics through the control, making classical time-discretizations or regression based schemes difficult to implement reliably and with stability. Moreover, the unknown components of the solution, such as the adjoint

process or optimal control as a function of the state space must be approximated in the entire domain. Traditional numerical methods for BSDEs, therefore, suffer from instability and do not scale well with dimension, particularly in presence of regime switching or nonlinear utilities. This leaves us with the motivation to develop a flexible functional approximation technique that can represent high-dimensional solutions that can be trained from a data set of simulated trajectories. Recent advances in machine learning (ML) have allowed us to look at a new class of numerical methods for solving high-dimensional BSDEs leveraging deep neural networks (NN) as flexible functional approximates. In particular, deep BSDE methods introduced by E. Han and Jentzen (E, Han, and Jentzen, 2017; Han, Jentzen, and E, 2018) parametrize the unknown components in BSDE solution using neural networks trained via stochastic optimization on simulated trajectories. By avoiding explicit discretizations of state space and learning functional relationship directly from data, we can scale to problems with large dimension of state variables. From an engineering perspective, this re-frames the solution of the UMPs as a learning problem over trajectory data where computation, scalability, and stability become primary design considerations.

While these learning-based approaches significantly improve scalability, their direct application to stochastic control problems introduces additional algorithmic challenges related to endogenous feedback between forward and backward components. In particular, independently trained primal and dual representations may converge differently, motivating the need for adaptive coupling mechanisms that explicitly enforce contraction of the duality gap during learning which further allows for policy convergence under structural assumptions.

1.3 Learning-Based BSDE Methods: Pros and Cons

Learning-based BSDE methods have achieved notable success in high-dimensional problems arising in quantitative finance and applied mathematics, most prominently in option pricing and nonlinear PDEs. In these settings, the backward equation is driven by an exogenous terminal payoff, and the forward dynamics are independent of the backward components. This structural decoupling allows the backward process to be learned efficiently as a function of the state variables, leading to stable and scalable algorithms even in moderately high dimensions (E, Han, and Jentzen, 2017; Han, Jentzen, and E, 2018; Becker, Cheridito, and Jentzen, 2019). As a result, deep BSDE solvers have become a practical alternative to grid-based PDE methods for a wide range of valuation problems. The structural properties that underpin the ease

of use of learning-based BSDE methods in valuation problems are fundamentally absent in UMP. In option pricing and related applications, the backward equation is driven by an exogenous terminal payoff and evolves independently of the control, while the forward dynamics are fixed and unaffected by the backward components. In contrast, UMPs give rise to fully coupled forward–backward systems in which the control enters the forward dynamics and must be determined endogenously from the backward equation. Errors in the learned control therefore propagate forward in time, altering the state distribution and directly affecting the training targets for the backward process. This feedback introduces instability during training and inhibits the straightforward extensions of deep BSDE methods, particularly in the presence of regime switching or non-conventional utility functions.

Furthermore, a challenge arises in learning based implementations of stochastic control through FBSDE formulations. In practice, the approaches (primal and dual) are typically trained independently, producing a lower and upper value estimates, whose difference forms an empirical *duality gap*. While the duality gap provides a useful certificate of sub-optimality, it is merely a diagnostic quantity which does not participate in the learning dynamics. As a result, both the primal and dual estimators converge at different rates and stabilize at different values particular in problems involving regime switching and structural nonlinearities. This observation motivates the development of *adaptive primal–dual learning mechanisms*, where the duality gap is incorporated directly as a input directly into the optimization object in order to stabilize training and enforce progressive tightening of the value bounds.

The presence of regime-switching dynamics further amplifies the difficulty of learning-based approaches to UMP. Due to the presence of stochastic coefficients with jump discontinuities, the associated FBSDE system is coupled not only through the control but across regimes. From a ML perspective, this induces a multi-modal structure in the optimal policy and value function as optimal response may change abruptly across regimes. Standard neural network approximations, which typically assume smooth dependence on state variables, struggle to capture such regime-dependent behavior without careful architectural design or explicit conditioning. Beyond regime switching, many physically relevant systems exhibit intrinsic path dependence which further complicates the application of ML based methods. Examples, includes stochastic volatility driven by long-memory processes, habit formations, learning effects, and delayed responses to exogenous shocks. In such

UMPS, the optimal control and value functions depend not only on the current state, but on the entire trajectory of the system. This destroys the Markovian nature of the problems and effectively renders finite-dimensional state representatives insufficient. As a result, the associated control problem is naturally formulated on an infinite-dimensional path space, governed by functional HJB equations or path-dependent FBSDEs.

Thinking from an algorithmic standpoint, solving such path-dependent regime-switching control problems would need functional appropriators that can represent complex temporal dependencies without relying on finite-dimensional state embeddings. A further and distinct challenge arises when the investor is also uncertain about the probability measure governing the market dynamics. In this setting, the learning problem acquires an adversarial dimension: rather than approximating a single value function under a fixed model, the algorithm must simultaneously learn an optimal investment policy and the worst-case model from an ambiguity set. This min–max structure introduces additional instability into the training dynamics and requires a dual representation of the robust problem that provides theoretical guarantees on the quality of the learned saddle point. The development of such a dual min–max learning framework, grounded in the Legendre–Fenchel duality machinery and the adaptive gap framework, constitutes the final and most general contribution of this thesis. Here standard feedforward NN are inherently memory-less, while recurrent architectures such as long-short-term memory (LSTM) or Gated Recurrent Unit (GRU) often struggle to retain information over long horizons due to vanishing gradient effects. In contrast, sequence-based models that operate directly on trajectories provide a solution to addressing these problems. By treating the control problem as a sequence-to-function learning task, such model can approximate value function and policies that depend on the entire observed history rather than a compressed state summary. Among sequence-based architectures, transformer models provide a particularly well-suited framework for addressing the challenges posed by path-dependent and regime-switching control problems. Unlike recurrent networks, transformers rely on attention mechanism to model dependencies across an entire sequence without explicit recursion (Vaswani et al., 2017). This ability is especially important in control problems driven by long-memory dynamics, where relevant information may be distributed across distant time steps. When combined with the BSDE-based formulation, transformer based architectures allow us to solve for path-dependent value functionals and optimal control policies directly from simulated trajectories.

1.4 Contributions of This Thesis

Motivated by the limitations of classical analytical, numerical, and learning-based approaches, this thesis develops a unified framework for solving regime-switching and path-dependent UMPs using aforementioned formulations and modern machine learning techniques. The primary contributions of this thesis are summarized as follows:

- We develop a learning-based framework for solving regime-switching utility maximization problems using FBSDEs, providing a probabilistic alternative to HJB methods in settings where classical verification arguments fail.
- We identify and analyze the fundamental algorithmic challenges that arise when applying learning-based BSDE methods to utility maximization, including endogenous control feedback, regime-dependent dynamics, and non-Markovian path dependence.
- We propose NN parameterizations for both primal and dual formulations of regime-switching UMPs, extending learning-based BSDE solvers beyond valuation problems to fully coupled stochastic control settings.
- We introduce an adaptive primal–dual learning framework in which the empirical duality gap is incorporated directly into the training objective. This coupling upgrades the duality gap from a post-hoc diagnostic quantity into a dynamic contraction mechanism that stabilizes learning and improves convergence of primal and dual value estimates. We further establish theoretical connections between duality gap reduction and policy convergence under structural concavity assumptions.
- We introduce sequence-based architectures, including transformer models, for approximating path-dependent value functionals and optimal control policies in non-Markovian environments without explicit state augmentation.
- We demonstrate through numerical experiments that the proposed methods scale to high-dimensional state spaces and complex regime-dependent dynamics where classical PDE-based, regression-based, and grid-based numerical methods are computationally infeasible or unstable.
- We develop a dual min–max framework for robust UMP under regime switching, in which model ambiguity is addressed by dualizing min–max primal

problem. This yields a novel dual robust functional whose saddle-point structure admits a full FBSDE characterization, and whose numerical solution is achieved through a GAN-style dual min–max deep learning algorithm for both the investor and the adversary simultaneously.

1.5 Related Literature and Connects

The work in this thesis culminates several different fields of study, including classical utility maximization, regime-switching stochastic control, BSDEs, path-dependent analysis, and ML-based numerical methods for high-dimensional problems.

Utility maximization and regime-switching control: Classical continuous-time utility maximization originates with the seminal work of Merton (Merton, 1971) and has been extensively developed using dynamic programming and martingale/duality methods (Karatzas and Shreve, 1998; Pham, 2009). Extensions to regime-switching models, where market coefficients depend on an exogenous finite-state Markov chain, have been studied to capture structural economic shifts and market heterogeneity; see, e.g., explicit-solution regimes in consumption–investment–insurance settings (Sotomayor and Cadenillas, 2009) and further developments in regime-switching duality-based control (e.g. margin/constraint variants) (Heunis, 2015). While these models provide richer economic insight, they typically lead to systems of coupled HJB equations or coupled FBSDE systems that admit closed-form solutions only under restrictive utility function and/or market specifications.

BSDEs and stochastic control: BSDE were introduced by Pardoux and Peng (Pardoux and S.G. Peng, 1990) and later connected to stochastic control and nonlinear PDEs through stochastic maximum principles and FBSDE representations (El Karoui, S. Peng, and Quenez, 1997; Shige Peng, 1990). These formulations yield powerful characterizations of optimal controls, but do not by themselves resolve the computational challenges posed by high-dimensional, regime coupling, and path dependence.

Path dependence - functional Itô calculus and PPDE theory (Dupire): Path-dependent control problems are naturally formulated on spaces of trajectories. A foundational calculus for such problems is *functional Itô calculus*, initially developed by Dupire and developed further through rigorous formulations and applications to (infinite-dimensional) Kolmogorov equations and finance (Cont and Fournié, 2013).

From the PDE perspective, this leads to *path-dependent PDEs* (PPDEs) and their viscosity-solution theory, providing a principled analytical counterpart to classical finite-dimensional HJB methods (see, e.g., viscosity solutions for fully nonlinear PPDEs) (Ekren, Touzi, and Zhang, 2014; Ekren, Touzi, and Zhang, 2016). These papers clarifies why state augmentation is often insufficient or impractical and motivates solvers that operate *directly on sampled paths* rather than on a finite-dimensional Markov state.

Deep learning in mathematical finance: Deep learning has been used as an extensive tool in mathematical finance, ranging from early neural approximations for pricing/hedging (Buehler et al., 2019; Horvath, Teichmann, and Žurič, 2021) to modern high-dimensional solvers and data-driven calibration. A widely cited early work on the role of deep learning in finance is (Heaton, Polson, and Witte, 2017). Recently however, learning-based BSDE methods have enabled scalable approximation of high-dimensional nonlinear PDEs and BSDEs in valuation problems (E, Han, and Jentzen, 2017; Han, Jentzen, and E, 2018; Becker, Cheridito, and Jentzen, 2019). The thesis directly reuses the core engineering ingredients of this line of work: (i) time discretizations of (B)SDEs on simulated trajectories, (ii) neural parameterizations of unknown conditional objects (e.g. gradients), and (iii) stochastic-gradient training of terminal-matching losses. However, we move away from valuation-style setups by addressing *endogenous control feedback* and *regime coupling*, which require policy parameterizations and better learning strategies beyond the decoupled-payoff setting. Recent work has also explored primal–dual and adversarial formulations for solving stochastic control problems using deep learning, often inspired by variational representations or saddle-point optimization techniques (Krach, Teichmann, and Wutte, 2025). These approaches highlight the importance of coupling complementary estimators in order to improve stability and accuracy of learning-based solvers.

Transformers and sequence modeling - insights relevant to path-dependent control: Sequence models provide a natural representation of path-dependent value functions. Transformers (Vaswani et al., 2017) replace recurrence with attention, enabling direct modeling of long-range temporal dependencies and parallelizable training. This thesis draws on several transformer insights established in adjacent areas: (i) *long-horizon time-series forecasting* transformers introduce architectural choices to handle long sequences efficiently (e.g. sparse/probabilistic attention and

decomposition-based inductive biases) (H. Zhou et al., 2021; T. Zhou et al., 2022; Wu et al., 2021), (ii) *sequence modeling for control* shows that policies can be learned as sequence-to-action mappings, where attention provides a flexible mechanism for using history (e.g. returns-to-go style conditioning) (L. Chen et al., 2021); and (iii) interpretable multi-horizon forecasting variants highlight conditioning and temporal fusion mechanisms for heterogeneous covariates (Lim et al., 2021). In our setting, these ideas motivate *causal attention* over simulated trajectories, explicit conditioning on regime information, and architectures that can represent long-memory features without forcing a finite-dimensional Markov state embedding.

Robust utility maximization and model ambiguity: Robust UMP, in which the investor maximizes expected utility against a worst-case probability measure drawn from an ambiguity set, has been studied extensively in the mathematical finance literature since the foundational work of (Gilboa and Schmeidler, 1989) on maxmin expected utility and (Hansen and Sargent, 2001) on robust control in macroeconomics. In continuous-time settings, robust portfolio optimization has been analyzed using stochastic control methods and min–max duality techniques, with explicit solutions available under logarithmic utility and specific ambiguity set structures (Guo et al., 2022). More recently, deep learning approaches have been proposed to address the computational intractability of robust UMPs in realistic market settings. In particular, (Krach, Teichmann, and Wutte, 2025) introduced a GAN-based algorithm that trains investor and adversary networks in a primal min–max game, demonstrating empirical success in multi-asset markets with transaction costs and path-wise constraints. However, this approach operates entirely within the primal formulation and does not exploit any dual representation of the robust problem, leaving the theoretical structure of the dual problem, including weak and strong duality, the FBSDE characterization of the saddle point, and the associated duality gap certificate, completely unexplored. In our work, we address this gap by developing the first dual min–max FBSDE framework for robust UMP under regime switching, grounding the GAN-style training in a rigorous duality theory and providing computable near-optimality certificates for both the investor and the adversary.

Positioning: Overall, existing literature provides either theoretical backbone without scalable numerical methods for regime-switching and path-dependent UMP, or ML-based solvers tailored primarily to decoupled valuation problems. This thesis

bridges these gaps by developing learning-based BSDE methods for *fully coupled* utility maximization under *regime switching* and *path dependence*, including both primal and dual parameterizations and sequence-based (transformer) approximators for history-dependent policies.

1.6 Organization of the Thesis

The remainder of this thesis is organized as follows.

Chapter 2 introduces the mathematical formulation of regime-switching utility maximization problems, including the probabilistic framework, admissible control sets, and the associated forward–backward stochastic differential equation (FBSDE) representation. Both primal and dual formulations are developed, together with neural parametrizations of the associated FBSDE systems, a comparison theorem establishing monotonicity of value estimates with respect to the admissible control class, and numerical experiments illustrating stability and scalability across five benchmark problems of increasing structural complexity.

Chapter 3 introduces an adaptive primal–dual learning framework in which the empirical duality gap is incorporated directly into the optimization objective. The resulting coupled learning dynamics are analyzed theoretically and shown to induce Lyapunov-type contraction of the squared duality gap and convergence of the learned control in $L^2([0,T])$ under structural concavity assumptions on the Hamiltonian. Numerical experiments on the regime-switching Merton benchmark validate the predicted geometric gap contraction and linear scaling between control error and gap magnitude.

Chapter 4 extends the framework to non-Markovian environments, where the value function and optimal policies depend on the full trajectory of the system. Two canonical sources of non-Markovianity are addressed: latent regime uncertainty, handled via the Wonham filter and a filtered FBSDE representation, and path-dependent long-memory dynamics, handled via Dupire's functional Itô calculus and a functional HJB system on path space. Sequence-based architectures, including transformer models with causal self-attention, are introduced to approximate path-dependent solutions without explicit state augmentation, and numerical experiments across rough volatility dynamics, path-dependent drift, and linear-quadratic examples validate the framework.

Chapter 5 develops a dual min–max framework for robust UMP under regime switching. The robust UMP is dualized via the Legendre–Fenchel transform, yield-

ing a dual robust functional whose min–max structure is characterized theoretically through weak duality, a formal min–max equality theorem, and an FBSDE representation of the saddle-point solution. A GAN-style dual min–max deep learning algorithm provides a joint solutions of near-optimality for both the investor and the adversary. Numerical experiments validate the framework against the non-robust benchmark.

Chapter 6 summarizes the main findings of the thesis, discusses the connections across chapters, acknowledges the limitations of the proposed framework, and outlines several directions for future research.

References

Becker, Sebastian, Patrick Cheridito, and Arnulf Jentzen (2019). "Deep optimal stopping". In: *Journal of Machine Learning Research* 20.74, pp. 1–25.

Buehler, H. et al. (2019). "Deep hedging". In: *Quantitative Finance* 19.8, pp. 1271–1291. DOI: 10.1080/14697688.2019.1571683. eprint: https://doi.org/10.1080/14697688.2019.1571683. URL: https://doi.org/10.1080/14697688.2019.1571683.

Chen, Lili et al. (2021). "Decision Transformer: Reinforcement Learning via Sequence Modeling". In: *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2021)*, pp. 15084–15097.

Chen, Zhuliang and Peter A. Forsyth (2008). "A Semi-Lagrangian Approach for Natural Gas Storage Valuation and Optimal Operation". In: *SIAM Journal on Scientific Computing* 30.1, pp. 339–368.

Cont, Rama and David-Antoine Fournié (2013). "Functional Itô calculus and functional Kolmogorov equations". In: *The Annals of Probability* 41.1, pp. 109–133.

E, Weinan, Jiequn Han, and Arnulf Jentzen (Dec. 2017). "Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations". In: *Communications in Mathematics and Statistics* 5.4, pp. 349–380.

Ekren, Ibrahim, Nizar Touzi, and Jianfeng Zhang (2014). "Viscosity Solutions of Fully Nonlinear Parabolic Path Dependent PDEs: Part I". In: *Annals of Probability* 42.1, pp. 204–236.

– (2016). "Viscosity solutions of fully nonlinear parabolic path dependent PDEs: Part II". In: *The Annals of Probability* 44.4, pp. 2507–2553. DOI: 10.1214/15-AOP1027. URL: https://doi.org/10.1214/15-AOP1027.

El Karoui, N., S. Peng, and M. C. Quenez (1997). "Backward Stochastic Differential Equations in Finance". In: *Mathematical Finance* 7.1, pp. 1–71. DOI: https://doi.org/10.1111/1467-9965.00022. eprint: https://

onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9965.00022. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9965.00022.

Fleming, Wendell H. and Halil Mete Soner (2006). "Deterministic Optimal Control". In: *Controlled Markov Processes and Viscosity Solutions*. New York, NY: Springer New York, pp. 1–55. ISBN: 978-0-387-31071-8. DOI: 10.1007/0-387-31071-1_1. URL: https://doi.org/10.1007/0-387-31071-1_1.

Gilboa, Itzhak and David Schmeidler (1989). "Maxmin expected utility with non-unique prior". In: *Journal of Mathematical Economics* 18.2, pp. 141–153. ISSN: 0304-4068. DOI: https://doi.org/10.1016/0304-4068(89)90018-9. URL: https://www.sciencedirect.com/science/article/pii/0304406889900189.

Guo, Ivan et al. (Jan. 2022). "Robust utility maximization under model uncertainty via a penalization approach". In: *Mathematics and Financial Economics* 16.1, pp. 51–88. ISSN: 1862-9660. DOI: 10.1007/s11579-021-00301-5. URL: https://doi.org/10.1007/s11579-021-00301-5.

Han, Jiequn, Arnulf Jentzen, and Weinan E (2018). "Solving high-dimensional partial differential equations using deep learning". In: *Proceedings of the National Academy of Sciences* 115.34, pp. 8505–8510. DOI: 10.1073/pnas.1718942115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1718942115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1718942115.

Hansen, Lars Peter and Thomas J. Sargent (2001). "Robust Control and Model Uncertainty". In: *The American Economic Review* 91.2, pp. 60–66. ISSN: 00028282. URL: http://www.jstor.org/stable/2677734 (visited on 03/29/2026).

Heaton, J. B., N. G. Polson, and J. H. Witte (2017). "Deep learning for finance: deep portfolios". In: *Applied Stochastic Models in Business and Industry* 33.1, pp. 3–12. DOI: https://doi.org/10.1002/asmb.2209. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asmb.2209. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2209.

Heunis, Andrew J. (2015). "Utility Maximization in a Regime Switching Model with Convex Portfolio Constraints and Margin Requirements: Optimality Relations and Explicit Solutions". In: *SIAM Journal on Control and Optimization* 53.4, pp. 2608–2656.

Horvath, Blanka, Josef Teichmann, and Žan Žurič (2021). "Deep Hedging under Rough Volatility". In: *Risks* 9.7. ISSN: 2227-9091. DOI: 10.3390/risks9070138. URL: https://www.mdpi.com/2227-9091/9/7/138.

Karatzas, Ioannis and Steven E. Shreve (1998). "Martingales, Stopping Times, and Filtrations". In: *Brownian Motion and Stochastic Calculus*. New York, NY: Springer New York, pp. 1–46.

Krach, Florian, Josef Teichmann, and Hanna Wutte (2025). *Robust Utility Optimization via a GAN Approach*. arXiv: 2403.15243 [q-fin.CP]. URL: https://arxiv.org/abs/2403.15243.

Lim, Bryan et al. (2021). "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting". In: *International Journal of Forecasting* 37.4, pp. 1748–1764. DOI: 10.1016/j.ijforecast.2021.03.012. URL: https://arxiv.org/abs/1912.09363.

Merton, Robert C (1971). "Optimum consumption and portfolio rules in a continuous-time model". In: *Journal of Economic Theory* 3.4, pp. 373–413. ISSN: 0022-0531. DOI: https://doi.org/10.1016/0022-0531(71)90038-X. URL: https://www.sciencedirect.com/science/article/pii/002205317190038X.

Pardoux, E. and S.G. Peng (1990). "Adapted solution of a backward stochastic differential equation". In: *Systems & Control Letters* 14.1, pp. 55–61. ISSN: 0167-6911. DOI: https://doi.org/10.1016/0167-6911(90)90082-6. URL: https://www.sciencedirect.com/science/article/pii/0167691190900826.

Peng, Shige (1990). "A General Stochastic Maximum Principle for Optimal Control Problems". In: *SIAM Journal on Control and Optimization* 28.4, pp. 966–979. DOI: 10.1137/0328054. eprint: https://doi.org/10.1137/0328054. URL: https://doi.org/10.1137/0328054.

Pham, Huyên (2009). "The classical PDE approach to dynamic programming". In: *Continuous-time Stochastic Control and Optimization with Financial Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–60. ISBN: 978-3-540-89500-8. DOI: 10.1007/978-3-540-89500-8_3. URL: https://doi.org/10.1007/978-3-540-89500-8_3.

Sotomayor, Luz Rocío and Abel Cadenillas (2009). "Explicit Solutions of Consumption-Investment Problems in Financial Marekts with Regime Switching". In: *Mathematical Finance* 19.2, pp. 251–279.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wu, Haixu et al. (2021). "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting". In: *Advances in Neural Information Processing Systems* 34, pp. 22419–22430.

Zhou, Haoyi et al. (2021). "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12, pp. 11106–11115.

Zhou, Tian et al. (2022). "FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research, pp. 27268–27286.

Chapter 2

MARKOVIAN STOCHASTIC CONTROL PROBLEM

The thesis focuses on studying different classes of stochastic maximization problems and testing various mathematical and numerical methods to obtain the optimal policy. We begin with a bottom-up approach, starting with the simplest form of optimization problems, where the state dynamics is Markovian in nature. What we mean by Markovian is that the drift and diffusion terms in the processes are deterministic in that they do not depend on $\omega \in \Omega$. This will form the foundational stepping stone on top of which we will be building to add more complication to the SMP by adding non-Markovian dynamics, path dependence, etc.

2.1 Setting up the Utility Maximization Problem (UMP)

UMP at its core is a stochastic control problem where the system is linear in control process. The setup for a simple regime-dependent UMP problem conceptualizes that the evolution of a controlled system is influenced not only by its current state but also by a prevailing regime, which can be a latent (hidden) or observable environment that governs the dynamics of the system. The most simple example of this would be a system or a wealth function which evolves with system parameters (drift, diffusion) based on an exogenous S finite-states continuous time Markov chain $\epsilon := \{\epsilon_t, t \in [0,T]\}$ with strongly irreducible generator $Q = [q_{i,j}]_{S\times S}$ where $q_{i,i} := -\lambda_i < 0$ and $\sum_{j\in S} q_{i,j} = 0$. This is a step beyond the standard literature, where stochasticity in state dynamics originates solely from a d-dimensional continuous-time Brownian motion $W = \{W_t, t \in [0,T]\}$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In fact, the regime-switching models allow for a richer description of both short-term and long-term uncertainties persistent in a given state. In financial markets, one can see parallels in which the bull-bear regimes influence stock prices to evolve differently over time. Although regime-switching models using Markov chains is a more accurate description of the markets, the added stochastic nature of the Markov chain violates the completeness of the market (Sotomayor and Cadenillas, 2009).

Remark 2.1.1. The condition to have a strongly irreducible generator for the Markov chain is largely technical in essence to ensure that the Markov chain visits all the regimes infinitely often across time, and that there is no state which is absorbing. Moreover, with this we also ensure that the transitions are rich to allow regime

interaction in the value function.

Let's now start defining the market model, which begins with understanding $\mathbf{F} = \{\mathcal{F}_t, t \in [0,T]\}$ to be the $\mathbb{P}$-augmentation of the filtration $\{\mathcal{F}^{W,\epsilon} := \sigma\{W_s, \epsilon_s : 0 \leq s \leq t\} \vee \mathcal{N}, t \in [0,T]\}$ generated by the stochastic process W, ϵ, where the $\mathcal{N}$ refers to the $\mathbb{P}$-null sets in $\mathcal{F}$. This filtration $\mathcal{F}^{W,\epsilon}$ at any time t represents the information available to the investor, containing all the continuous uncertainty from the W and ϵ up to time t. Moreover, this filtration satisfies the usual conditions (right-continuous and completeness), and all the stochastic processes considered further-on from here are assumed to be adapted to $\mathbf{F}$ (Heunis, 2015).

A regime-independent UMP is obtained as a special case of the regime-dependent UMP by assuming that all coefficients are independent of the regime variable ϵ_t and are identical across regimes. In this case, the regime process does not influence the state dynamics, although it may still be included in the underlying probability space. Throughout this study, we assume that the Brownian motion W and the regime process ϵ are independent.

For our study, we will consider the market to comprise of a risk-free asset S_0 and d-risky assets S_i for $i = 1, 2, 3, ...d$ modeled by the relations

$$\begin{aligned} dS_0(t) &= S_0(t) r_{\epsilon(t)}\, dt, && t \in [0,T], \quad S_0(0) = 1, \\ dS_i(t) &= S_i(t)\left(\mu_{i,\epsilon(t)}(t)\, dt + \sigma_{i,\epsilon(t)}(t)\, dW^\top(t)\right), && t \in [0,T], \quad S_i(0) > 0, \end{aligned} \tag{2.1}$$

We assume the initial condition $\epsilon_{t=0} = \epsilon_0$. Note that the diffusion term for the risky assets involves a *matrix–vector multiplication*, and the superscript $^\top$ denotes the *transpose operator*, not to be confused with T, which refers to the *terminal time* in this finite-horizon model. Without matrix notation, the term $\sigma_{i,\epsilon(t)}(t)\, dW^\top(t)$ can be written as

$$\sum_{n=1}^{d} \sigma_{i,\epsilon(t),n}(t)\, dW_n(t), \tag{2.2}$$

where $\sigma_{i,\epsilon(t),n}(t)$ refers to the entry in the i-th row and n-th column of the $d \times d$ matrix $\sigma_{\epsilon(t)}(t)$, representing the volatility of the risky assets in regime $\epsilon(t)$. In the (2.1), one can see that the coefficients of the SDEs all depend on the regime of the economy. Under this pretense, the investor has to choose a *portfolio policy* $\pi = \{\pi_t = (\pi_0(t), \pi_1(t)...\pi_d(t))^\top, t \in [0,T]\}$, where each element $\pi_i(t)$ of the policy vector refers to the fraction of total wealth that is invested in the i-th asset. Now, this policy vector is modeled as a progressively measured stochastic process,

belonging to a constrained set $K \subseteq \mathbb{R}^d$ which in our case will be convex. In case $K = R^d$, the control is unconstrained and the problem is reduced to an unconstrained UMP.

Remark 2.1.2. We will assume that our π will be square integrable, as is the case in the previous literature (Wiedermann, 2022). Although this looks like a technical condition, the assumption is critical so as to protect our wealth dynamics from diverging in finite time. Economic interpretation would posit that square integrable prevents the investor from taking arbitrarily large-sized position leading to unbounded risk.

Consolidating all the conditions for π we can then define a set of *admissible strategies* as:

$$\mathcal{A} := \left\{ \pi \in \mathcal{A}_{\text{prog}} \,\middle|\, \pi(t) \in K \text{ a.s. for a.e. } t \in [0,T], \quad \mathbb{E}\left[\int_0^T |\pi(t)|^2 \, dt \right] < \infty \right\}. \tag{2.3}$$

where $\mathcal{A}_{\text{prog}}$ refers to a set of all progressively measurable processes mapping $\Omega \times [0,T] \to \mathbb{R}^d$. Furthermore, for any strategy $\pi \in \mathcal{A}$, we can finally define the associated *wealth process* $X^\pi(t)$ following the *self-financing condition* by:

$$dX^\pi(t) = \sum_{i=1}^{d} \frac{X^\pi(t)\pi_i(t)}{S_i(t)} \, dS_i(t) + \frac{X^\pi(t)}{S_0(t)} \left(1 - \sum_{j=1}^{d} \pi_j(t) \right) dS_0(t), \quad t \in [0,T]. \tag{2.4}$$

By substituting the asset dynamics from equation (2.1) into equation (2.4), which yields:

$$dX^\pi(t) = \alpha(t, \epsilon_t, X_t, \pi_t) \, dt + \beta(t, \epsilon_t, X_t, \pi_t)^\top dW(t), \quad X^\pi(0) = x_0. \tag{2.5}$$

Here, $\alpha(\cdot)$ and $\beta(\cdot)$ represent the drift and diffusion terms of the wealth process, respectively. The current formulation of the wealth process does not restrict the nature of the coefficients of the SDE. However, if the coefficients were to be deterministic in nature, the the UMP retains its Markovian structure. This would include the case where the coefficients are time-homogeneous in nature where the coefficients are measurable function of t (Øksendal, 2003). If the coefficients are random, i.e., they depend on sources of randomness not fully captured by the current state or regime, then the problem becomes a Non-Markovian UMP as the state dynamics now exhibit dependence on the full path or filtration. That being said, the only restriction that we

ascertain as of now is that these $\alpha(t, \epsilon_t, X_t, \pi_t)$, and $\beta(t, \epsilon_t, X_t, \pi_t)$ remain uniformly bounded, and progressively measurable processes with respect to filtration $\mathcal{F}^{W,\epsilon}$ thereby ensuring that the SDE admits strong unique solutions.

We are now at a juncture where we can formally introduce the central problem studied in this thesis.

Definition 2.1.3. Let g be a real-valued utility function that is strictly increasing and strictly concave, and let X^π be the solution to the SDE in (2.5) for every $\pi \in \mathcal{A}$. We then define the associated *gain function* $J(t, \epsilon, x, \pi)$ as:

$$J(t, \epsilon, x, \pi) := \mathbb{E}\left[\int_t^T f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + g\left(X^{t,x,\pi}(T)\right)\right] \tag{2.6}$$

where the running utility function f and the terminal utility function g are assumed to be bounded. One objective of the utility maximization problem is to determine the *value function*:

$$v(t, \epsilon, x) := \sup_{\pi \in \mathcal{A}} J(t, \epsilon, x, \pi), \tag{2.7}$$

which represents the maximal expected utility attainable at time t, regime ϵ, and wealth x, by following an admissible-optimal control π^*. Definition 2.1.3, serves as the *primal version* of our optimization problem for the rest of the thesis.

Remark 2.1.4. Do note that the value function $v(t, \epsilon, x)$ is defined for every $t \in [0, T]$, which is standard in continuous-time stochastic control and is essential for the application of dynamic programming principles. A commonly studied special case in the literature corresponds to UMP from terminal wealth only. This setting is recovered from (2.6) by choosing the running utility function $f \equiv 0$ (and $g \equiv U$), in which case the objective reduces to

$$\mathbb{E}\big[U(X^\pi(T))\big].$$

There is also scope for more complex formulations to be studied. One such example can be that of (Sotomayor and Cadenillas, 2009; Heunis, 2015) which introduces a setting that incorporates a *consumption process* with its own separate consumption utility function U_c. For these problems we are then needed to maximize the expected utility derived from both the terminal gain and running consumption process, leading to gain function of the form:

$$\mathbb{E}\left[\int_t^T U_c(c(s))\, ds + U(X^\pi(T))\right].$$

All of these formulations fall within the broader class of *expected utility maximization problems*, and notably, they do not explicitly incorporate risk metrics such as variance or standard deviation in the objective. Thus, these variations are, in essence, structured modifications of the general problem presented here.

While the theoretical results and proofs in this thesis are developed within the context of the generalized setting introduced above, we will draw connections to these alternative formulations and, where relevant, present numerical examples that demonstrate how they can be approached using the deep learning methodologies developed herein.

Remark 2.1.5. It is assumed that the utility function is strictly increasing to reflect the preference for higher wealth, and strictly concave to capture risk aversion. Moreover, a strictly concave utility function U ensures the existence and uniqueness of the optimal solution in utility maximization problems (Cvitanić and Karatzas, 1992).

2.2 Structural Characterization of the Primal Problem

Focusing on the Markovian structure of the problem, we observe that solving for the value function at each point (t, ϵ, x) amounts to analyzing a family of control problems, each initialized at a different time, regime, and wealth level. This naturally leads to the application of the *dynamic programming principle* (DPP), which provides a recursive framework for characterizing the value function and deriving the associated Hamilton–Jacobi–Bellman (HJB) equation (Pham, 2009).

Dynamic Programming Principle

Theorem 2.2.1. Let (t, ϵ, x) be any initial state, where $t \in [0, T]$, $\epsilon \in S$, and $x \in \mathbb{R}$. Suppose that the coefficients in equation (2.4) are deterministic, so as to preserve the Markovian structure of the UMP defined in Definition 2.1.3. Then, for any stopping time $\tau : \Omega \to [t, T]$, the value function satisfies the DPP:

$$v(t, \epsilon, x) = \sup_{\pi \in \mathcal{A}} \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + v\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau)\right)\right] \tag{2.8}$$

Proof. The proof for the DPP has been well explained and studied in the classical case with out regime-switching in (Pham, 2009; Fleming and Soner, 2006). The same has been used in (Wiedermann, 2022). The key distinction however, is that in our proof, we would extend it to the regime-switching with the idea that the

augmented state processes (X_t^π, ϵ_t) is Markovian in nature, and thereby the value function inherits a recursive optimality property. The need for the Markovian structure here is not just a mere technicality but rather critical as it allows us to solve the for the value function via DPP.

Let us say that $\pi \in \mathcal{A}$ is an admissible control strategy, and for any stopping time $\tau \in [t, T]$, owing to the Markovian nature using the Law of Iterated Conditional Expectations write

$$J(t, \epsilon, x, \pi) = \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + J\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau), \pi\right)\right] \quad (2.9)$$

Since the remaining expected value from time τ onward is bounded above by the value function

$$J(t, \epsilon, x, \pi) \leq \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + v\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau)\right)\right] \quad (2.10)$$

Taking the supremum over all $\pi \in \mathcal{A}$ and using the Equation (2.7) for the left hand side, we obtain:

$$v(t, \epsilon, x) \leq \sup_{\pi \in \mathcal{A}} \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + v\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau)\right)\right]. \quad (2.11)$$

Equation (2.11) establishes one direction of the dynamic programming principle and is relatively straightforward to prove. However, the converse inequality is more subtle. The challenge arises from the fact that, at any given time t, we do not know which control $\pi \in \mathcal{A}$ yields the optimal value — we only know that the value function represents the supremum over all such controls. To overcome this, we follow a standard technique in the literature: constructing an ε-optimal control π^ε, which achieves performance within ε of the value function.

Lets set the value of $\varepsilon > 0$. For each possible set $(\epsilon_\tau, X^{t,x,\pi}(\tau))$, there exists a control π^ε on $[\tau, T]$ such that:

$$J\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau), \pi^\varepsilon\right) \geq v\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau)\right) - \varepsilon. \quad (2.12)$$

Construct a new admissible control $\tilde{\pi}$ by combining a near-optimal control π^* on $[t, \tau]$ with π^ε on $[\tau, T]$. This concatenation is valid using the measurable selection theorem under the admissibility and measurability assumptions for $\mathcal{A}$, the $\tilde{\pi} \in \mathcal{A}$ (Bertsekas and Shreve, 1978).

Remark 2.2.2. The rationale behind introducing an ε-optimal control lies in our approach to approximate the optimal policy over the entire interval $[t, T]$. While the exact π may be unknown, $\tilde{\pi}$ strategy approximates the expected utility within ε of the true value function.

Then by using the law of conditional expectations we can write:

$$J(t, \epsilon, x, \tilde{\pi}) \geq \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\tilde{\pi}}(s), \tilde{\pi}(s)\right)\, ds + v\left(\tau, \epsilon_\tau, X^{t,x,\tilde{\pi}}(\tau)\right)\right] - \varepsilon \quad (2.13)$$

Furthermore by taking the supremum over all such concatenated controls, and limiting $\varepsilon \to 0$, we obtain:

$$v(t, \epsilon, x) \geq \sup_{\pi \in \mathcal{A}} \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + v\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau)\right)\right] \quad (2.14)$$

Finally, with the combination of Equations (2.11), (2.14) we can finally conclude with the equality which is the regime-switching form of DPP:

$$v(t, \epsilon, x) = \sup_{\pi \in \mathcal{A}} \mathbb{E}\left[\int_t^\tau f\left(s, \epsilon_s, X^{t,x,\pi}(s), \pi(s)\right)\, ds + v\left(\tau, \epsilon_\tau, X^{t,x,\pi}(\tau)\right)\right] \quad (2.15)$$

□

Theorem 2.2.1 hints at a very intuitive idea: To maximize expected gains over the full horizon $[t, T]$, policy needs to be optimal across all the sub-intervals belonging to the entire time-horizon. The more direct interpretation is that the value function at time t is produced via the locally optimal policy from $[t, \tau]$ and then over $[\tau, T]$ recursively, leading to a globally optimal policy.

Remark 2.2.3. Although Theorem 2.2.1 provides a straightforward conceptual framework for solving the value function, the literature shows that it is very difficult to apply in practice. (Pham, 2009) notes that solving the integral formulation, especially in continuous time, is extremely challenging and often intractable. References such as (Pham, 2009; Fleming and Soner, 2006) often motivate the derivation of the *Hamilton–Jacobi–Bellman* (HJB) equation as a practical alternative to overcome the computational intractability of working directly with the DPP. The HJB equation is an infinitesimal, local version of the DPP—rather than reasoning over the entire interval $[t, T]$, it describes the evolution of the value function over an infinitesimally small time window.

Despite the HJB equation being a local formulation of the DPP, certain technical conditions must be satisfied in order to rigorously express the system dynamics in terms of partial derivatives. Specifically, we require that $v \in C^{1,2}([0,T] \times \mathbb{R})$, meaning it is once differentiable in time and twice differentiable in the state variable. This smoothness assumption allows us to apply Itô's lemma to the value function as part of the HJB derivation.

Derivation of the Regime-Switching HJB Equation

Theorem 2.2.4. Under suitable technical conditions to be specified below (in particular equation (2.25)), consider the stochastic control problem defined in Definition 2.1.3, where the state evolves under regime switching. Assume that the value function $v \in C^{1,2}([0,T] \times \mathbb{R})$ for each observable regime $\epsilon_i \in S$, is evaluated along the process (s, ϵ_s, X_s^π), and that the control $\pi \in K$ is Markovian and admissible.

Then the value function satisfies the following coupled system of Hamilton–Jacobi–Bellman (HJB) equations: for each $\epsilon_i \in S$,

$$\begin{aligned} -\frac{\partial v}{\partial t}(t, \epsilon_i, x) = &\sup_{\pi \in K} \{ f(t, \epsilon_i, x, \pi) + \mathcal{L}^\pi v(t, \epsilon_i, x) \} \\ &+ \sum_{j \in S} q_{ij} \left(v(t, \epsilon_j, x) - v(t, \epsilon_i, x) \right), \quad (t, x) \in [0, T) \times \mathbb{R}. \end{aligned} \tag{2.16}$$

for which the terminal condition is set as:

$$v(T, \epsilon_i, x) = g(\epsilon_i, x), \qquad x \in \mathbb{R}, \quad \epsilon_i \in S. \tag{2.17}$$

Here, $\mathcal{L}^\pi$ denotes the differential generator of the diffusion process under control π, defined as:

$$\mathcal{L}^\pi v(t, \epsilon_i, x) := \alpha(t, \epsilon_i, x, \pi) \frac{\partial v}{\partial x}(t, \epsilon_i, x) + \frac{1}{2} |\beta(t, \epsilon_i, x, \pi)|^2 \frac{\partial^2 v}{\partial x^2}(t, \epsilon_i, x). \tag{2.18}$$

Proof. Let us fix a point $(t, \epsilon_i, x) \in [0, T) \times S \times \mathbb{R}$, and define $h \in (0, T - t]$ arbitrarily. Moreover, let us assume that $\pi \in K$ is a fixed, arbitrary control. While $\pi \in \mathcal{A}$ typically includes all progressively measurable or Markovian controls—thus representing a broader class in the context of deriving the HJB equation, it is necessary for the controlled state process X^π to possess the Markov property and admit a direct application of Itô's formula. To that end, restricting attention to constant controls over a small time interval $[t, t + h]$ enables a clean and rigorous application of dynamic programming arguments. Furthermore, we assume that

$\pi \in K$ is admissible, which is trivially satisfied under the conditions provided in Definition 2.1.3.

Now, let us take the following stopping time:

$$\tau_{t,\epsilon_i,x,1} := \inf\left\{s \in [t,T] : \left|X_s^\pi - x\right| \geq 1\right\} \wedge T. \tag{2.19}$$

Since the process X_s^π is continuous and adapted, the stopping time from the equation above is well defined. The reason we define a stopping time is to ensure that the state process remains within a bounded neighborhood of its initial value x over the time interval. In order to work with a short time interval, we define $\tau_h := \tau_{t,\epsilon_i,x,1} \wedge (t+h)$, such that the process is now locally controlled in space and confined to the interval $[t, t+h]$, where π is assumed to be constant.

Using the Theorem 2.2.1 applied to the stopping time τ_h, we can write

$$v(t,\epsilon_i,x) \geq \mathbb{E}\left[\int_t^{\tau_h} f(s,\epsilon_i,X_s^\pi,\pi)\,ds + v(\tau_h,\epsilon_{\tau_h},X_{\tau_h}^\pi)\right] \tag{2.20}$$

The inequality appears in the equations as we are calculating the expectation under a particular π which might not always be optimal. Moreover, as $v \in C^{1,2}([0,T]\times\mathbb{R})$ by assumption, we can apply Itô-Dynkin's formula

$$\begin{aligned} dv(s,\epsilon_s,X_s^\pi) = \Bigg[& \frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \alpha(s,\epsilon_s,X_s^\pi,\pi)\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi) \\ & + \frac{1}{2}|\beta(s,\epsilon_s,X_s^\pi,\pi)|^2\frac{\partial^2 v}{\partial x^2}(s,\epsilon_s,X_s^\pi) \\ & + \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)\Bigg] ds \\ & + \beta(s,\epsilon_s,X_s^\pi,\pi)\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi)\,dW(s) + dM^\epsilon(s) \end{aligned} \tag{2.21}$$

In this equation, in addition to the drift and diffusion terms, there appears an additional martingale term $dM^\epsilon(s)$, which arises from the jumps in the regime-switching process. This term captures the stochastic fluctuations associated with the transitions. Although it contributes to the stochastic dynamics of the value function, it vanishes in expectation and does not affect the drift term. Furthermore, using the Equation (2.18) we can write:

$$dv(s,\epsilon_s,X_s^\pi) = \left[\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \mathcal{L}^\pi v(s,\epsilon_s,X_s^\pi) + \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)\right]ds + \beta(s,\epsilon_s,X_s^\pi,\pi)\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi)\,dW(s) + dM^\epsilon(s) \tag{2.22}$$

$$\begin{aligned} v(\tau_h,\epsilon_{\tau_h},X_{\tau_h}^\pi) - v(t,\epsilon_i,x) &= \int_t^{\tau_h}\left(\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \mathcal{L}^\pi v(s,\epsilon_s,X_s^\pi)\right)ds \\ &+ \int_t^{\tau_h}\sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)ds \\ &+ \int_t^{\tau_h}\beta(s,\epsilon_s,X_s^\pi,\pi)\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi)dW(s) + \int_t^{\tau_h} dM^\epsilon(s) \end{aligned} \tag{2.23}$$

Note that at time t, we assume that we are in regime ϵ_i. Furthermore, if we take conditional expectation given information up to time t, i.e., with respect to the filtration $\mathcal{F}_t$ on either sides of the expression above, we can get

$$\begin{aligned} \mathbb{E}\left[v(\tau_h,\epsilon_{\tau_h},X_{\tau_h}^\pi)\right] - v(t,\epsilon_i,x) &= \mathbb{E}\left[\int_t^{\tau_h}\left(\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \mathcal{L}^\pi v(s,\epsilon_s,X_s^\pi)\right)ds\right] \\ &+ \mathbb{E}\left[\int_t^{\tau_h}\sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)ds\right] \\ &+ \mathbb{E}\left[\int_t^{\tau_h}\beta(s,\epsilon_s,X_s^\pi,\pi)\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi)\,dW(s)\right] \\ &+ \mathbb{E}\left[\int_t^{\tau_h} dM^\epsilon(s)\right]. \end{aligned} \tag{2.24}$$

As the Markov Chain is observable, the regime at time t is also known. As a result the HJB equation is derived separately for each regime $\epsilon_i \in S$, yielding a coupled system of PDEs indexed by the regime. Let us now examine certain terms on the right-hand side of equation (2.24).

Remark 2.2.5. In particular, consider the third expectation, whose argument is the Itô integral

$$\int_t^{\tau_h}\beta(s,\epsilon_s,X_s^\pi,\pi)\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi)\,dW(s),$$

which can easily be seen to be a martingale with zero expectation. The only condition required for the martingale property to hold is that the integrand in the Itô integral is square-integrable. This is ensured by the assumption that

$$\mathbb{E}\left[\int_t^{\tau_h} |\beta(s,\epsilon_s,X_s^\pi,\pi)|^2 \left(\frac{\partial v}{\partial x}(s,\epsilon_s,X_s^\pi)\right)^2 ds\right] < \infty. \tag{2.25}$$

Remark 2.2.6. We can also show that the process

$$M^\epsilon(s) := \int_t^s dM^\epsilon(r)$$

is a progressively measurable martingale with initial condition $M^\epsilon(t) = 0$. This term is a purely discontinuous local martingale that arises from the jump structure of the regime-switching process. By the martingale property, it satisfies

$$\mathbb{E}[M^\epsilon(\tau) \mid \mathcal{F}_s] = M^\epsilon(s), \quad \text{for all } s \le \tau.$$

Since the initial condition is zero, we have $\mathbb{E}[M^\epsilon(\tau)] = 0$ as long as M^ϵ is integrable. This condition is ensured by localization, since the stopped process $v(s \wedge \tau_h, \epsilon_{s\wedge\tau_h}, X^\pi_{s\wedge\tau_h})$ is bounded by construction.

Combining the remarks 2.2.5 and 2.2.6 we can obtain:

$$\begin{aligned}\mathbb{E}\left[v(\tau_h,\epsilon_{\tau_h},X^\pi_{\tau_h})\right] - v(t,\epsilon_i,x) = \mathbb{E}\left[\int_t^{\tau_h}\left(\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \mathcal{L}^\pi v(s,\epsilon_s,X_s^\pi)\right)ds\right] \\ + \mathbb{E}\left[\int_t^{\tau_h}\sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)ds\right]\end{aligned} \tag{2.26}$$

Finally, we will put in the equation (2.20) to obtain:

$$\begin{aligned}v(t,\epsilon_i,x) \ge \mathbb{E}\Bigg[\int_t^{\tau_h}\Bigg(f(s,\epsilon_s,X_s^\pi,\pi) + \frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \mathcal{L}^\pi v(s,\epsilon_s,X_s^\pi) \\ + \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)\Bigg)ds\Bigg] + v(t,\epsilon_i,x)\end{aligned} \tag{2.27}$$

Subtracting $v(t,\epsilon_i,x)$ from both sides and dividing by h for some $h > 0$, we get:

$$\begin{aligned}0 \ge \frac{1}{h}\mathbb{E}\Bigg[\int_t^{\tau_h}\Bigg(f(s,\epsilon_s,X_s^\pi,\pi) + \frac{\partial v}{\partial s}(s,\epsilon_s,X_s^\pi) + \mathcal{L}^\pi v(s,\epsilon_s,X_s^\pi) \\ + \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^\pi) - v(s,\epsilon_s,X_s^\pi)\right)\Bigg)ds\Bigg].\end{aligned} \tag{2.28}$$

Since the stopping time $\tau_h := \tau_{t,\epsilon_i,x,1} \wedge (t+h)$ confines the state process to a bounded neighborhood of x over the interval $[t, t+h]$, all terms inside the expectation are uniformly bounded on $[t, \tau_h]$. Letting $h \to 0$, and using the continuity of all relevant terms together with the facts that $\epsilon_s \to \epsilon_i$, $X_s^\pi \to x$, and $s \to t$, we may apply the dominated convergence theorem (Karatzas and Shreve, 1998) to obtain:

$$0 \geq f(t, \epsilon_i, x, \pi) + \frac{\partial v}{\partial t}(t, \epsilon_i, x) + \mathcal{L}^\pi v(t, \epsilon_i, x) + \sum_{j \in S} q_{ij} \left(v(t, \epsilon_j, x) - v(t, \epsilon_i, x) \right) \tag{2.29}$$

Since this inequality holds for any constant admissible control $\pi \in K$, taking the supremum over $\pi \in K$ gives:

$$-\frac{\partial v}{\partial t}(t, \epsilon_i, x) \geq \sup_{\pi \in K} \{ f(t, \epsilon_i, x, \pi) + \mathcal{L}^\pi v(t, \epsilon_i, x) \} + \sum_{j \in S} q_{ij} \left(v(t, \epsilon_j, x) - v(t, \epsilon_i, x) \right). \tag{2.30}$$

Under optimality conditions, the inequality becomes an equality, yielding the final HJB equation.

$$-\frac{\partial v}{\partial t}(t, \epsilon_i, x) = \sup_{\pi \in K} \{ f(t, \epsilon_i, x, \pi) + \mathcal{L}^\pi v(t, \epsilon_i, x) \} + \sum_{j \in S} q_{ij} \left(v(t, \epsilon_j, x) - v(t, \epsilon_i, x) \right). \tag{2.31}$$

Given the HJB equation, the terminal condition for the PDE can be derived from the Theorem 2.2.1. At the end of the period T, one can see that there is no further opportunity to control the state process and the investor simply receives the terminal payoff $g(\epsilon_T, X_T)$. Moreover, at time T, the state X_t is known and so is the regime ϵ_T, and therefore, one can conclude that $v(T, \epsilon_i, x) = g(\epsilon_i, x), \forall \epsilon_i \in S,\ x \in \mathbb{R}$. This serves as the terminal boundary condition required to fully characterize the solutions of Equation (2.31). □

The proof of Theorem 2.2.4 is relatively straightforward, and the intuition behind it shows that, for a given value function $v(t, \epsilon_i, x)$, it can be shown to satisfy the coupled HJB system given by Equation (2.31). While Theorem 2.2.4 provides sufficient conditions to solve the stochastic control problem, it fails to establish the necessary conditions. In simpler terms, it does not prove that the candidate value function $v(t, \epsilon_i, x)$ is indeed the true optimal value function of the stochastic control problem. Furthermore, it does not guarantee the existence of an optimal policy π^*.

To complete the theoretical framework of the coupled HJB system, it is necessary to establish a *verification theorem*, as in (Sotomayor and Cadenillas, 2009). This theorem must address two key components:

- First, that any sufficiently smooth candidate function w which satisfies the coupled HJB equations in an inequality form—along with appropriate boundary, terminal, and growth conditions—coincides with the true value function v, provided that a control $\pi^* \in \mathcal{A}$ exists that attains the supremum in the HJB system almost surely.
- Second, that if such a control π^* exists and attains the supremum, then it is an optimal control, and the candidate function represents the cost of following that control policy. This establishes a *sufficient condition* for optimality.

Verification Theorem for the Coupled HJB System

Theorem 2.2.7. Let $w(t, \epsilon_i, x)$ be a candidate value function $\in C^{1,2}([0,T) \times \mathbb{R}) \cap C([0,T] \times \mathbb{R})$ such that for every $\epsilon_i \in S$,

1. It satisfies the **HJB inequality equation** (derived from Equation (2.31)) for all $(t,x) \in [0,T) \times \mathbb{R}$,

$$-\frac{\partial w}{\partial t}(t, \epsilon_i, x) \geq \sup_{\pi \in K} \left\{ f(t, \epsilon, x, \pi) + \mathcal{L}^{\pi} w(t, \epsilon_i, x) \right\} + \sum_{j \in S} q_{ij} \left(w(t, \epsilon_j, x) - w(t, \epsilon_i, x) \right).$$

2. It satisfies the **terminal condition** for the coupled HJB equations for all $x \in \mathbb{R}$:

$$w(T, \epsilon_i, x) \geq g(\epsilon_i, x),$$

3. It also satisfies a **polynomial growth condition**, where there exist constants $C > 0$, $p \geq 1$ such that

$$|w(t, \epsilon, x)| \leq C(1 + |x|^p), \quad \text{for all } (t, x, \epsilon) \in [0,T] \times \mathbb{R} \times S.$$

Furthermore, assume that there exists $\pi^* \in \mathcal{A}$, where $\mathcal{A}$ follows the definition given previously in Equation (2.3), such that, almost surely for almost every $s \in [t,T]$,

$$\pi^*(s) \in \arg\max_{\pi \in K} \left\{ f(s, \epsilon_s, X_s^{\pi^*}, \pi) + \mathcal{L}^{\pi} w(s, \epsilon_s, X_s^{\pi^*}) \right\}.$$

Additionally, we assume that the controlled state process $(X_s^{\pi^*}, \epsilon_s)$ exists uniquely, and that for each admissible control $\pi \in \mathcal{A}$, there exists p such that

$$\mathbb{E}\left[\sup_{s \in [t,T]} |X_s^{\pi}|^p \right] < \infty,$$

where p is the same exponent appearing in the polynomial growth condition of the candidate value function w. Then:

1. For any admissible control $\pi \in \mathcal{A}$, the candidate function w dominates the performance functional:
$$w(t, \epsilon_i, x) \geq J(t, \epsilon_i, x; \pi) := \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X_s^\pi, \pi(s))\, ds + g(\epsilon_T, X_T^\pi)\right].$$
2. Moreover, if the control π^* is used, then it is a *Markovian optimal control*, and w coincides with the true value function:
$$v(t, \epsilon_i, x) := \sup_{\pi \in \mathcal{A}} J(t, \epsilon_i, x; \pi) = J(t, \epsilon_i, x; \pi^*) = w(t, \epsilon_i, x).$$

Proof. Let $(t, \epsilon_i, x) \in [0, T] \times S \times \mathbb{R}$ be fixed, and let $\pi \in \mathcal{A}$ be any admissible control. Let (X_s^π, ϵ_s) denote the corresponding controlled state and regime process starting from $X_t^\pi = x$, $\epsilon_t = \epsilon_i$.

Since the candidate value function $w \in C^{1,2}([0, T) \times \mathbb{R})$ for each regime $\epsilon_i \in S$, and satisfies the polynomial growth condition, we may apply the Itô-Dynkin formula to the process $s \mapsto w(s, \epsilon_s, X_s^\pi)$. To rigorously handle the stochastic integrals, we define a sequence of stopping times $(\tau_n)_{n \geq 1}$ following Pham's approach (Pham, 2009):

$$\tau_n := \inf\left\{s \geq t : \int_t^s \left|\beta(u, \epsilon_u, X_u^\pi, \pi(u)) \frac{\partial w}{\partial x}(u, \epsilon_u, X_u^\pi)\right|^2 du \geq n\right\} \wedge T. \quad (2.32)$$

This ensures that the stochastic integral in the Itô formula is square-integrable on $[t, \tau_n]$, and that $\tau_n \uparrow T$ almost surely as $n \to \infty$.

Remark 2.2.8. While we were working with the proof for theorem 2.2.4, we directly assumed that the Itô integral was square-integrable by imposing the condition of integrability on the integrand (Remark 2.2.5). One may still use that approach and apply Itô's formula directly on the full interval $[t, T]$ without the use of the stopping times.

The approach of using the stopping times as expressed in equation (2.32) avoids assumption of square-integrability. Instead, with the help of the stopping times, we are localizing the state-processes and guarantee that the stochastic integrals are well-defined. Both of these approaches are equivalent in the current setting due to the presence of the assumptions such as the polynomial growth condition and integrability of the state process.

Applying the regime-switching Itô-Dynkin formula on $[t, \tau_n]$, we obtain:

$$\begin{aligned} w(\tau_n, \epsilon_{\tau_n}, X^\pi_{\tau_n}) = w(t, \epsilon_i, x) &+ \int_t^{\tau_n} \left(\frac{\partial w}{\partial s} + \mathcal{L}^{\pi(s)} w \right)(s, \epsilon_s, X^\pi_s)\, ds \\ &+ \int_t^{\tau_n} \sum_{j \in S} q_{sj} \left(w(s, \epsilon_j, X^\pi_s) - w(s, \epsilon_s, X^\pi_s) \right) ds \\ &+ \int_t^{\tau_n} \beta(s, \epsilon_s, X^\pi_s, \pi(s)) \frac{\partial w}{\partial x}(s, \epsilon_s, X^\pi_s)\, dW(s) \\ &+ \int_t^{\tau_n} dM^\epsilon(s), \end{aligned} \tag{2.33}$$

where $M^\epsilon(s)$ is a local martingale associated with the compensated jumps of the finite-state Markov chain ϵ_s. Taking the conditional expectation given the information available up to time t, and invoking Remarks 2.2.8 and 2.2.6, we observe that both the stochastic Itô integral and the regime-switching martingale term are true martingales under the imposed integrability and growth conditions.

As already mentioned, the structure of the stopping-times (τ_n) is vital to ensure that the integrand of the Itô integral is square-integrable, which forms the basis of the expectation to be 0 (Remark 2.2.8). Similarly, the jump process $M^\epsilon(s)$, associated with discontinuous regime-switching, is also a martingale with zero expectation provided the value function is bounded in each regime (Remark 2.2.6). Finally, we can write:

$$\begin{aligned} \mathbb{E}\left[w(\tau_n, \epsilon_{\tau_n}, X^\pi_{\tau_n}) \right] = w(t, \epsilon_i, x) &+ \mathbb{E}\left[\int_t^{\tau_n} \left(\frac{\partial w}{\partial s} + \mathcal{L}^{\pi(s)} w \right)(s, \epsilon_s, X^\pi_s)\, ds \right] \\ &+ \mathbb{E}\left[\int_t^{\tau_n} \sum_{j \in S} q_{sj} \left(w(s, \epsilon_j, X^\pi_s) - w(s, \epsilon_s, X^\pi_s) \right) ds \right]. \end{aligned} \tag{2.34}$$

Now, by the HJB inequality satisfied by the candidate value function w, we have for all $(s, x) \in [t, T) \times \mathbb{R}$ and almost every path:

$$\frac{\partial w}{\partial s}(s, \epsilon_s, X^\pi_s) + \mathcal{L}^{\pi(s)} w(s, \epsilon_s, X^\pi_s) + \sum_{j \in S} q_{sj} \left(w(s, \epsilon_j, X^\pi_s) - w(s, \epsilon_s, X^\pi_s) \right) \leq -f(s, \epsilon_s, X^\pi_s, \pi(s)) \tag{2.35}$$

Substituting this into Equation (2.34), we obtain:

$$\mathbb{E}\left[w(\tau_n, \epsilon_{\tau_n}, X^\pi_{\tau_n}) \right] \leq w(t, \epsilon_i, x) - \mathbb{E}\left[\int_t^{\tau_n} f(s, \epsilon_s, X^\pi_s, \pi(s))\, ds \right]. \tag{2.36}$$

Now, since w satisfies a polynomial growth condition, there exist constants $C > 0$ and $p \geq 1$ such that

$$|w(s, \epsilon, x)| \leq C(1 + |x|^p), \quad \forall (s, \epsilon, x) \in [t, T] \times S \times \mathbb{R}.$$

In particular, for each n,

$$|w(\tau_n, \epsilon_{\tau_n}, X^\pi_{\tau_n})| \leq C\left(1 + |X^\pi_{\tau_n}|^p\right) \leq C\left(1 + \sup_{s\in[t,T]} |X^\pi_s|^p\right).$$

By assumption, we have

$$\mathbb{E}\left[\sup_{s\in[t,T]} |X^\pi_s|^p\right] < \infty,$$

so the upper bound is integrable and independent of n. Therefore, by the dominated convergence theorem, we can pass to the limit in the left-hand side of Equation (2.36):

$$\lim_{n\to\infty} \mathbb{E}\left[w(\tau_n, \epsilon_{\tau_n}, X^\pi_{\tau_n})\right] = \mathbb{E}\left[w(T, \epsilon_T, X^\pi_T)\right].$$

Moreover, since $f(s, \epsilon_s, X^\pi_s, \pi(s))$ is integrable on $[t, T]$, and $\tau_n \uparrow T$, we can again apply dominated convergence to obtain:

$$\lim_{n\to\infty} \mathbb{E}\left[\int_t^{\tau_n} f(s, \epsilon_s, X^\pi_s, \pi(s))\, ds\right] = \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X^\pi_s, \pi(s))\, ds\right].$$

Taking the limit as $n \to \infty$ in Equation (2.36), we conclude:

$$\mathbb{E}\left[w(T, \epsilon_T, X^\pi_T)\right] \leq w(t, \epsilon_i, x) - \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X^\pi_s, \pi(s))\, ds\right]. \tag{2.37}$$

Finally, using the terminal condition $w(T, \epsilon, x) \geq g(\epsilon, x)$, we obtain:

$$\mathbb{E}\left[g(\epsilon_T, X^\pi_T)\right] \leq w(t, \epsilon_i, x) - \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X^\pi_s, \pi(s))\, ds\right]. \tag{2.38}$$

which can be rewritten as

$$w(t, \epsilon_i, x) \geq \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X^\pi_s, \pi(s))\, ds + g(\epsilon_T, X^\pi_T)\right] = J(t, \epsilon_i, x; \pi).$$

This steps marks the end of the proof of the first part of the theorem where we show that the candidate value function is a *classical supersolution* for any admissible control $\pi \in \mathcal{A}$

Now for the second part of the theorem we are going to consider control $\pi^* \in \mathcal{A}$ such that for almost every $s \in [t, T]$, we have

$$\pi^*(s) \in \arg\max_{\pi\in K} \left\{f(s, \epsilon_s, X^{\pi^*}_s, \pi) + \mathcal{L}^\pi w(s, \epsilon_s, X^{\pi^*}_s)\right\}$$

Then, along the path of the process $(X_s^{\pi^*}, \epsilon_s)$, the HJB inequality now attains an equality:

$$\frac{\partial w}{\partial s}(s, \epsilon_s, X_s^{\pi^*}) + \mathcal{L}^{\pi^*(s)} w(s, \epsilon_s, X_s^{\pi^*}) + \sum_{j \in S} q_{sj} \left(w(s, \epsilon_j, X_s^{\pi^*}) - w(s, \epsilon_s, X_s^{\pi^*}) \right) = -f(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s))$$

From here, we can repeat the same steps as we did in the earlier part of the proof from Hence, repeating the same steps as in Part 1, but instead of an inequality in equation (2.35) we take the equality and obtain:

$$w(t, \epsilon_i, x) = \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s))\, ds + g(\epsilon_T, X_T^{\pi^*}) \right] = J(t, \epsilon_i, x; \pi^*). \quad (2.39)$$

Thus, π^* is a *Markovian optimal control policy*, and w coincides with the true value function v. □

The verification theorem above provides a critical sufficiency condition to confirm whether a given candidate value function w is the true value function and that the associated control π^* is indeed optimal policy. However, it fails to offer a method for obtaining such a candidate function or the corresponding optimal control.

Limitations of Classical Verification and Motivation for Viscosity Solutions

In classical theory (Pham, 2009) (no regime-switching), the standard strategy has been to first solve the HJB equation under smoothness and structural assumptions, then extract the feedback control $\pi^*(t, x) \in \arg\max \{f + \mathcal{L}^\pi w\}$, and finally invoke the verification theorem to confirm that the constructed pair is indeed optimal. While this approach exists in theory and might work for very simplified cases of the UMP, it is often not applicable due to very strict assumptions which underlie the verification theorem itself. In many practical applications, particularly in UMP with regime switching, these assumptions frequently break down. Key limitations include:

1. **Smoothness of the value function:** The assumption that the candidate value function is smooth rarely holds in practical settings. In fact, in higher-dimensional UMP, the value function is often irregular and lacks the differentiability required by the classical verification theorem. Even in low-dimensional cases, smoothness may fail unless very strong structural conditions are imposed on the coefficients.

2. **Behavior of the optimal control π^*:** One of the strongest assumptions is the existence of a measurable optimal control π^* that attains the supremum in the HJB equation for all (t, x). This is a non-trivial requirement, often requiring prior knowledge or explicit structural characterization of π^*, which is typically unavailable in practical applications.

3. **Structural conditions for PDE solvability:** Classical results that guarantee the existence of smooth solutions to the HJB equation rely on strong structural assumptions, such as uniform ellipticity or parabolicity (Friedman, 2010). These typically take the form of lower bounds on the diffusion

$$y^\top \beta(t, \epsilon, x, \pi)\beta^\top(t, \epsilon, x, \pi)y \geq c|y|^2, \quad \text{for all } y \in \mathbb{R}^n \text{ and some } c > 0.$$

   While such conditions are comfortably satisfied in simple diffusion models, they often fail in financial applications where the volatility is multiplicative, control-affine (π appears linearly in drift and diffusion), or degenerate in certain regions of the state space. In particular, this assumption prevents the direct application of the verification theorem in settings involving incomplete markets.

These limitations essentially motivate the need for *weaker notions of solutions*, particularly in settings where the classical $C^{1,2}$ regularity of the value function no longer holds true. One such framework is the use of *viscosity solutions*, which allows for meaningful analysis and characterization of the value function even when it is not smooth. It provides with an alternative to classical verification theorems, while still preserving the core structure of the control problem and the associated HJB equations.

This marks a good point to end this section. Although we have not fully solved the primal problem, we have developed a rigorous framework to verify and characterize the value function and the optimal control. Furthermore, the structural assumptions under which the classical verification theorems apply introduce significant limitations that render them impractical for several real-world problems. These often push the need for more flexible and scalable approaches to solve more complicated versions of the regime-dependent UMP. In the next section, we aim to move beyond traditional analytical approaches and instead drive the discussion toward more algorithmic formulations that allow us to approximate the value function and thereby the optimal control using more advanced learning tools.

2.3 Learning-Based Solutions: Deep-Regime-Dependent BSDEs

The theoretical framework developed in the previous sections provides rigorous conditions under which a candidate value function and its associated control can be verified as optimal. However, this framework is inherently non-constructive as it presupposes the existence of a sufficiently regular value function and an optimal feedback control without offering a systematic procedure for computing either object. In regime-switching UMP, this limitation is exacerbated by the presence of discrete regime dynamics, which induce non-smooth coupling across the system of HJB equations and frequently violate the smoothness, ellipticity, and structural assumptions required by classical analytical or numerical methods. Consequently, even when existence and characterization results are available, direct computation of the value function and optimal policy becomes infeasible, motivating the need for alternative, scalable solution paradigms grounded in probabilistic representations of stochastic control problems.

Learning-based methods, in contrast to traditional numerical schemes which suffer from dimensionality issues and rigid grid structures can approximate high-dimensional nonlinear functions effectively without significant instability. Furthermore, deep learning models are particularly adept at learning value function approximations and optimal control strategies from simulated path data. They provide a highly flexible and scalable framework that can naturally incorporate discontinuities and non-smooth transitions, such as regime switches. Among these techniques, one of the most recent and impactful is the Deep BSDE method introduced by (E, Han, and Jentzen, 2017). The main idea in their paper was to compute the value function by solving a coupled system of Forward-Backward Stochastic Differential Equations (FBSDEs), and then use neural networks (NNs) to approximate the conditional expectations and gradients along sample paths of the forward process. This approach avoids the need for grid-based discretizations and handles irregular coefficients, thus preventing computational costs from diverging. The insights of this algorithm are far-reaching, and several variants have since been derived to solve specific problems in optimal execution (Sirignano and Spiliopoulos, 2018), stochastic control (E, Han, and Jentzen, 2017), and reinforcement learning (Becker, Cheridito, and Jentzen, 2019).

However, the current implementation of Deep BSDE algorithms has been limited in the context of regime-switching problems. While the existing literature touches upon applications of Deep BSDEs to mean-field theory, control with delay, and related

domains, the development of learning-based methods tailored to regime-dependent UMPs remains largely unexplored. This is where the current thesis bridges a critical gap in the literature. In this section, we aim to integrate a rigorous FBSDE-based framework specifically designed for regime-switching UMPs with a scalable deep learning architecture. The regime-switching aspect of the UMP presents several challenges when applying learning-based methods via BSDEs. Discontinuities in the dynamics, the non-smooth coupling across the system of HJB equations, and the need to track regimes during learning all introduce significant complications. To address these issues, one strategy is to reformulate the switching process in a way that is compatible with neural network approximation, while ensuring that the FBSDE representation remains valid across regime boundaries. However, before delving into the learning-based solution, we must first construct a unified master theorem—a single representation result that encapsulates the full behavior of both the optimal control and the value function. This theorem serves as the foundational link between the theoretical characterization developed in the previous section and its algorithmic realization through Deep BSDEs. It provides not only the mathematical basis but also the practical specification required for learning-based solvers.

Theorem 2.3.1. Consider a utility maximization problem (UMP) as described in Definition 2.1.3. Let the value function $v(t, \epsilon, x)$ be defined on $[0, T] \times S \times \mathbb{R}$, where $S \subset \mathbb{N}$ denotes a finite set of regimes. Assume that for each $\epsilon \in S$, the value function $v(t, \epsilon, x) \in C^{1,3}([0, T] \times \mathbb{R})$ satisfies a polynomial growth condition of order $k \in \mathbb{N}$, and that the mixed partial derivative $\frac{\partial^2 v}{\partial t \partial x}$ exists and is continuous.

With a fixed initial condition $(t, x) \in [0, T] \times \mathbb{R}$, let us assume the existence of an optimal control $\pi^* \in \mathcal{A}$, which is an admissible control such that the corresponding state process $X_s^{\pi^*}$ satisfies

$$\mathbb{E}[|X_T^{\pi^*}|^k] < \infty,$$

and all stochastic integrals and expectations are well defined.

Then the following results hold:

1. **Optimality.** The value function satisfies the coupled Hamilton-Jacobi-Bellman (HJB) equation in feedback form:

$$\frac{\partial v}{\partial s}(s, \epsilon_s, X_s^{\pi^*}) + \mathcal{L}^{\pi^*(s)} v(s, \epsilon_s, X_s^{\pi^*}) + \sum_{j \in S} q_{sj} \left(v(s, \epsilon_j, X_s^{\pi^*}) - v(s, \epsilon_s, X_s^{\pi^*}) \right)$$
$$= -f(s, \epsilon_s, X_s^{\pi^*}, \pi(s))$$

Moreover, the control π^* attains this supremum almost surely for all $s \in [t, T]$:

$$\pi^*(s) \in \arg\max_{\pi \in K} \left\{ f(s, \epsilon_s, X_s^{\pi^*}, \pi) + \mathcal{L}^{\pi} v(s, \epsilon_s, X_s^{\pi^*}) \right\}.$$

2. **Backward Propagation.** Define the process

$$Z_s^\top := \beta^\top(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s)) \nabla_x v(s, \epsilon_s, X_s^{\pi^*}),$$

and assume:

$$v \in S^2(t, T; \mathbb{R}), \quad Z \in H^2(t, T; \mathbb{R}^m),$$

where

$$H^2(0, T; \mathbb{R}^j) := \left\{ X : \Omega \times [0, T] \to \mathbb{R}^j \;\middle|\; \text{prog. meas.} \wedge \mathbb{E}\left[\int_0^T |X(t)|^2 \, dt \right] < \infty \right\},$$
$$S^2(0, T; \mathbb{R}^j) := \left\{ X : \Omega \times [0, T] \to \mathbb{R}^j \;\middle|\; \text{prog. meas.} \wedge \mathbb{E}\left[\sup_{t \in [0,T]} |X(t)|^2 \right] < \infty \right\} \tag{2.40}$$

Then the pair (v, Z_s) satisfies the backward stochastic differential equation:

$$dv_s = -f(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s)) \, ds + Z_s^\top dW(s) + dM^\epsilon(s), \quad v_T = g(\epsilon_T, X_T^{\pi^*}),$$

where $dM^\epsilon(s)$ is the purely discontinuous local martingale given by

$$dM^\epsilon(s) = \sum_{\epsilon_i, \epsilon_j \in S,\, i \neq j} \left(v(s, \epsilon_j, X_s^{\pi^*}) - v(s, \epsilon_i, X_s^{\pi^*}) \right) dM_{ij}(s),$$

with $M_{ij}(s)$ denoting the canonical martingales associated with jumps from regime i to regime j.

3. **First-Order Adjoint Structure.** Define the gradient and curvature processes as:

$$P_s := \nabla_x v(s, \epsilon_s, X_s^{\pi^*}), \qquad Q_s := \beta(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s)) \nabla_x^2 v(s, \epsilon_s, X_s^{\pi^*}),$$

and assume:

$$P \in S^2(t, T; \mathbb{R}), \quad Q \in H^2(t, T; \mathbb{R}^m).$$

Then the pair (P_s, Q_s) solves the first-order backward adjoint equation:

$$dP_s = -\nabla_x \mathcal{H}(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s), P_s, Q_s) \, ds + Q_s^\top dW(s) + dM^{\epsilon, \nabla}(s), \quad P_T = \nabla_x g(\epsilon_T, X_T^{\pi^*}),$$

where the generalized Hamiltonian $\mathcal{H}$ is defined by

$$\mathcal{H}(s, \epsilon, x, \pi, p, q) := f(s, \epsilon, x, \pi) + \alpha(s, \epsilon, x, \pi)^\top p + \operatorname{Tr}\left[\beta(s, \epsilon, x, \pi) q \right],$$

and where the purely discontinuous martingale $dM^{\epsilon,\nabla}(s)$ is given by

$$dM^{\epsilon,\nabla}(s) = \sum_{\epsilon_i,\epsilon_j \in S,\, i\neq j} \left(\nabla_x v(s,\epsilon_j, X_s^{\pi^*}) - \nabla_x v(s,\epsilon_i, X_s^{\pi^*})\right) dM_{ij}(s).$$

Finally, if the above conditions (1)−(3) hold together, then the quadruple $(X_s^{\pi^*}, v_s, Z_s, P_s)$ solves the following coupled forward-backward stochastic differential equation system on $[t,T]$:

$$\begin{aligned} dX_s &= \alpha(s,\epsilon_s, X_s^{\pi^*}, \pi^*(s))\, ds + \beta^\top(s,\epsilon_s, X_s^{\pi^*}, \pi^*(s))\, dW(s), \\ dv_s &= -f(s,\epsilon_s, X_s^{\pi^*}, \pi^*(s))\, ds + Z_s^\top\, dW(s) + dM^{\epsilon}(s), \\ dP_s &= -\nabla_x \mathcal{H}(s,\epsilon_s, X_s^{\pi^*}, \pi^*(s), P_s, Q_s)\, ds + Q_s^\top\, dW(s) + dM^{\epsilon,\nabla}(s), \end{aligned} \tag{2.41}$$

with initial condition $X_t = x$ and terminal conditions:

$$v_T = g(\epsilon_T, X_T^{\pi^*}), \qquad P_T = \nabla_x g(\epsilon_T, X_T^{\pi^*})$$

Proof. We proceed in three steps, corresponding to the statements (1)–(3) of the theorem.

Optimality We follow the strategy presented in (Wiedermann, 2022), adapted to the regime-switching setting.

Let us define for each $n \in \mathbb{N}$ the sequence of stopping times $(\tau_n)_{n\in\mathbb{N}}$ by

$$\tau_n := \inf\left\{ s \in [t,T] \;\middle|\; |X_s^{\pi^*} - x| \geq n \right\} \wedge T.$$

Since $X_s^{\pi^*}$ has continuous paths, we conclude that $\tau_n \nearrow T$ almost surely as $n \to \infty$.

Applying Itô's formula to the process $v(s,\epsilon_s, X_s^{\pi^*})$ between t and τ_n and taking the conditional expectations like before, while taking into account the jump contributions from the regime-switching process ϵ_s, yields:

$$\begin{aligned} v(t,\epsilon_t,x) = \mathbb{E}\left[v(\tau_n, \epsilon_{\tau_n}, X_{\tau_n}^{\pi^*})\right] - \mathbb{E}\Bigg[\int_t^{\tau_n} \Bigg(&\frac{\partial v}{\partial s}(s,\epsilon_s, X_s^{\pi^*}) \\ &+ \mathcal{L}^{\pi^*(s)} v(s,\epsilon_s, X_s^{\pi^*}) \\ &+ \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j, X_s^{\pi^*}) - v(s,\epsilon_s, X_s^{\pi^*})\right)\Bigg) ds\Bigg] \end{aligned} \tag{2.42}$$

Note that the expectation above already accounts for the fact that the expectation of the martingale terms, pursuant to Remark 2.2.5 and Remark 2.2.6, is zero.

Furthermore, by Theorem 2.2.7, any candidate value function must satisfy the inequality in Equation 2.35. Substituting this into the above expression, we obtain:

$$v(t,\epsilon_t,x) \geq \mathbb{E}\left[v(\tau_n,\epsilon_{\tau_n},X_{\tau_n}^{\pi^*})\right] + \mathbb{E}\left[\int_t^{\tau_n} f(s,\epsilon_s,X_s^{\pi},\pi^*(s))\,ds\right] \tag{2.43}$$

One can also arrive at another form of an inequality where we can see that equation (2.42) is lower bounded by:

$$\begin{aligned} v(t,\epsilon_t,x) \geq\ & \mathbb{E}\left[v(\tau_n,\epsilon_{\tau_n},X_{\tau_n}^{\pi^*})\right] + \mathbb{E}\left[\int_t^{\tau_n} f(s,\epsilon_s,X_s^{\pi},\pi^*(s))\,ds\right] \\ & - \mathbb{E}\Bigg[\int_t^{\tau_n}\Bigg(\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^{\pi^*}) + \sup_{\pi\in K}\left\{\mathcal{L}^{\pi}v(s,\epsilon_s,X_s^{\pi^*}) + f(s,\epsilon_s,X_s^{\pi^*},\pi(s))\right\} \\ & + \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^{\pi^*}) - v(s,\epsilon_s,X_s^{\pi^*})\right)\Bigg)ds\Bigg]. \end{aligned} \tag{2.44}$$

In essence, all three equations (2.42), (2.43), and (2.44), under the given condition that the value function behaves well and satisfies a polynomial growth condition, reach equality by the Dominated Convergence Theorem. In particular, under these circumstances, we can treat both (2.43) and (2.44) as equalities and deduce that:

$$\begin{aligned} 0 = \mathbb{E}\Bigg[\int_t^{\tau_n}\Bigg(&\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^{\pi^*}) + \sup_{\pi\in K}\left\{\mathcal{L}^{\pi}v(s,\epsilon_s,X_s^{\pi^*}) + f(s,\epsilon_s,X_s^{\pi^*},\pi^*(s))\right\} \\ &+ \sum_{j\in S} q_{\epsilon_s j}\left(v(s,j,X_s^{\pi^*}) - v(s,\epsilon_s,X_s^{\pi^*})\right)\Bigg)ds\Bigg]. \end{aligned} \tag{2.45}$$

Since the integrand is non-negative and the limits of integration are arbitrary, we can apply the Monotone Convergence Theorem. As a result, we obtain the HJB equation:

$$\begin{aligned} &\frac{\partial v}{\partial s}(s,\epsilon_s,X_s^{\pi^*}) + \mathcal{L}^{\pi^*(s)}v(s,\epsilon_s,X_s^{\pi^*}) + \sum_{j\in S} q_{sj}\left(v(s,\epsilon_j,X_s^{\pi^*}) - v(s,\epsilon_s,X_s^{\pi^*})\right) \\ &\qquad = -f(s,\epsilon_s,X_s^{\pi^*},\pi(s)) \end{aligned} \tag{2.46}$$

And similarly using the monotone convergence theorem from (2.44) and (2.42), we can see that:

$$\sup_{\pi\in K}\left\{\mathcal{L}^{\pi}v(s,\epsilon_s,X_s^{\pi^*}) + f(s,\epsilon_s,X_s^{\pi^*},\pi(s))\right\} = \mathcal{L}^{\pi^*}v(s,\epsilon_s,X_s^{\pi^*}) + f(s,\epsilon_s,X_s^{\pi^*},\pi^*(s)) \tag{2.47}$$

This implies that the optimal π^* indeed maximizes the supremum in the HJB equation, which concludes the proof for the optimality condition.

Backward Propagation The proof for this fairly straightforward as long as the integrability conditions (2.40) are satisfied so that we can apply Itô lemma on v to get

$$\begin{aligned} dv(s, \epsilon_s, X_s^\pi) = \Bigg[& \nabla_s v(s, \epsilon_s, X_s^{\pi^*}) + \mathcal{L}^{\pi^*} v(s, \epsilon_s, X_s^{\pi^*}) \\ & + \sum_{j \in S} q_{sj} \left(v(s, \epsilon_j, X_s^{\pi^*}) - v(s, \epsilon_s, X_s^{\pi^*}) \right) \Bigg] ds \\ & + \beta(s, \epsilon_s, X_s^{\pi^*}, \pi^*) \nabla_x v(s, \epsilon_s, X_s^{\pi^*}) \, dW(s) + dM^\epsilon(s) \end{aligned} \tag{2.48}$$

where M^ϵ is again the pure-jump martingale from the regime ϵ_s. From equation (2.46), we can replace out the drift terms, and also re-write the diffusion coefficient in terms of the new process Z_s as:

$$dv_s = -f(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s)) \, ds + Z_s^\top dW(s) + dM^\epsilon(s) \tag{2.49}$$

Finally, we can obtain from the definition of the value function at final time T is $v_T = g(\epsilon_T, X_T^{\pi^*})$.

First-Order Adjoint Structure So far, in both of the previous conditions, it would have been sufficient to assume that $v \in C^{1,2}$. However, in this case, we need to apply Itô's formula to the first-order derivative of v with respect to x, namely $\frac{\partial v}{\partial x}$, which, by assumption, belongs to $C^{1,2}([0,T] \times \mathbb{R})$. Eventually we obtain:

$$\begin{aligned} d\left(\nabla_x v(s, \epsilon_s, X_s^{\pi^*}) \right) = dP_s = \Bigg[& \frac{\partial}{\partial s} \nabla_x v(s, \epsilon_s, X_s^{\pi^*}) + \nabla_x \mathcal{L}^{\pi^*} v(s, \epsilon_s, X_s^{\pi^*}) \\ & + \sum_{j \in S} q_{sj} \left(\nabla_x v(s, \epsilon_j, X_s^{\pi^*}) - \nabla_x v(s, \epsilon_s, X_s^{\pi^*}) \right) \Bigg] ds \\ & + \beta^\top(s, \epsilon_s, X_s^{\pi^*}, \pi^*) \nabla_x^2 v(s, \epsilon_s, X_s^{\pi^*}) dW(s) \\ & + dM^{\epsilon, \nabla}(s), \end{aligned} \tag{2.50}$$

where the term $dM^{\epsilon,\nabla}(s)$ is a purely discontinuous local martingale arising from jumps in the regime-switching process ϵ_s.

Remark 2.3.2. In order to understand the purely discontinuous martingale term $dM^{\epsilon,\nabla}(s)$ appearing in the equation above, we would need to introduce the *canonical martingales* associated with the regime-switching process ϵ_s, following the

construction of (Heunis, 2015). Furthermore, the sole appearance of the purely discontinuous martingale terms when applying Itô rule to jump process follows standard results which one can see in detail in (Øksendal and Sulem, 2007).

For each pair of distinct regimes $\epsilon_i, \epsilon_j \in S$ ($\epsilon_i \neq \epsilon_j$), we can define:

1. The *jump counting process* $R_{ij}(t)$ by

$$R_{ij}(t) := \sum_{0<s\leq t} \mathbf{1}_{\{\epsilon_{s-}=\epsilon_i\}} \mathbf{1}_{\{\epsilon_s=\epsilon_j\}},$$

   counting the number of transitions from regime i to j up to time t.

2. The *predictable compensator* $\widetilde{R}_{ij}(t)$ by

$$\widetilde{R}_{ij}(t) := \int_0^t q_{ij}(s) \mathbf{1}_{\{\epsilon_s=\epsilon_i\}}\, ds,$$

   where $q_{ij}(s)$ denotes the transition intensity from ϵ_i to ϵ_j.

3. The *canonical martingale* $M_{ij}(t)$ by

$$M_{ij}(t) := R_{ij}(t) - \widetilde{R}_{ij}(t).$$

These processes $M_{ij}(t)$ are purely discontinuous, square-integrable, $\{\mathcal{F}_t\}$-local martingales, and are mutually orthogonal: for $(i, j) \neq (k, l)$, we have $[M_{ij}, M_{kl}] = 0$. In this framework, the martingale term $dM^{\epsilon,\nabla}(s)$ appearing in (2.50) can be explicitly written as

$$dM^{\epsilon,\nabla}(s) = \sum_{\epsilon_i,\epsilon_j\in S,\, i\neq j} \left(\nabla_x v(s, \epsilon_j, X_s^{\pi^*}) - \nabla_x v(s, \epsilon_i, X_s^{\pi^*})\right) dM_{ij}(s). \tag{2.51}$$

Similarly, in the original Itô formula for $v(s, \epsilon_s, X_s^{\pi^*})$ itself, the corresponding martingale term $dM^{\epsilon}(s)$ can be expressed as

$$dM^{\epsilon}(s) = \sum_{\epsilon_i,\epsilon_j\in S,\, i\neq j} \left(v(s, \epsilon_j, X_s^{\pi^*}) - v(s, \epsilon_i, X_s^{\pi^*})\right) dM_{ij}(s). \tag{2.52}$$

These expressions highlight that the randomness introduced by regime-switching is captured entirely through stochastic integrals against the canonical martingales $M_{ij}(s)$. In particular, under appropriate integrability conditions, the expectation of these purely discontinuous local martingale terms vanishes which is what we already discussed in Remark 2.2.6.

Getting back to the proof of the adjoint structure, we can notice that the terms in the drift part of equation (2.50) can be obtained from the derivative of the HJB equation itself (2.46). It can be seen from the structure of the HJB that the equation attains a local maximum at $x = X_s^{\pi^*}$ (the third argument). Under this consideration, what we can do is differentiate equation (2.46) with respect to x, substitute $x = X_s^{\pi^*}$, and then equate the resulting expression to zero.

$$\begin{aligned}\nabla_y \Bigg(& \frac{\partial v}{\partial s}(s, \epsilon_s, y) + \mathcal{L}^{\pi^*(s)} v(s, \epsilon_s, y) \\ & + \sum_{j \in S} q_{sj} \left(v(s, \epsilon_j, y) - v(s, \epsilon_s, y) \right) + f(s, \epsilon_s, y, \pi^*(s)) \Bigg) \Bigg|_{y = X_s^{\pi^*}} = 0.\end{aligned} \tag{2.53}$$

Expanding the derivative in equation (2.53), we obtain:

$$\begin{aligned}& \nabla_y \left(\frac{\partial v}{\partial s}(s, \epsilon_s, y) \right) + \nabla_y \left(\mathcal{L}^{\pi^*(s)} v(s, \epsilon_s, y) \right) \\ & \quad + \sum_{j \in S} q_{sj} \left(\nabla_y v(s, \epsilon_j, y) - \nabla_y v(s, \epsilon_s, y) \right) \Bigg|_{y = X_s^{\pi^*}} \\ & = -\nabla_y f(s, \epsilon_s, y, \pi^*(s)) - \nabla_y \left(\alpha(s, \epsilon_s, y, \pi^*(s)) \frac{\partial v}{\partial y}(s, \epsilon_s, y) \right) \\ & \quad - \nabla_y \operatorname{Tr} \left(\beta(s, \epsilon_s, y, \pi^*(s)) \beta^\top(s, \epsilon_s, y, \pi^*(s)) \nabla_y^2 v(s, \epsilon_s, y) \right) \Bigg|_{y = X_s^{\pi^*}} \\ & = -\nabla_y \left(\mathcal{H}(s, \epsilon_s, X_s^{\pi^*}, P_s, Q_s) \right)\end{aligned} \tag{2.54}$$

Substituting equation (2.54) into the drift coefficient of equation (2.50), with a change of variable from y back to x, and replacing all newly introduced processes, namely P_s and Q_s, we obtain:

$$dP_s = -\nabla_x \mathcal{H}(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s), P_s, Q_s)\, ds + Q_s^\top dW(s) + dM^{\epsilon, \nabla}(s) \tag{2.55}$$

with the terminal condition $P_T = \nabla_x g(\epsilon_T, X_T^{\pi^*})$ satisfied by the definition of P_s.

Finally, if we combine all the three components of the proof, we can systematically arrive at the conclusion that the quadruple $(X_s^{\pi^*}, v_s, Z_s, P_s)$ solve the coupled forward-BSDE system of equations shown in equation (2.41) with their respective terminal conditions. □

The above theorem establishes a fully stochastic representation of the UMP in the form of a coupled forward-BSDE system, augmented with purely discontinuous

local martingales arising from jumps due to regime-switching. In essence, the theorem presents itself as an algorithm to solve for the value function, where the solution can, in principle, be approximated numerically by solving the associated forward-BSDE system. This is because the optimality conditions are encoded within the dynamics of the forward process $X_s^{\pi^*}$ and in the backward processes for v_s, P_s, and Q_s, representing the value function and the first- and second-order sensitivities. Nonetheless, several challenges naturally arise from this structure:

1. The presence of purely discontinuous martingale terms complicates both the analytical (theoretical) derivation and the numerical approximation of the optimal solutions.

2. Due to the regime-switching nature of ϵ_s, which introduces discrete jumps, we require specialized methods to handle numerical instabilities during these jumps.

These observations reinforce the idea that the development of learning-based methods is indeed the most suitable approach, as these frameworks can handle discontinuities, incorporate discrete regime-switching effects, and efficiently approximate both the forward and backward processes involved.

That being said, solving the primal version of the UMP often poses significant challenges. both analytically and numerically. In particular, when considering a constrained UMP, we are required not only to find an optimal control policy $\pi^* \in \mathcal{A}$ but also to ensure that the admissibility conditions and the regime-switching dynamics hold. Due to the introduction of discrete jumps and discontinuous martingales, the optimization landscape becomes incredibly complex, even when we seek to approximate the solutions using learning-based numerical methods.

Traditionally, this problem has been approached by considering the dual problem formulation. This alternative version of the problem rephrases the original optimization into a more tractable form by leveraging duality techniques, notably the Legendre-Fenchel transform. By applying this transformation, we attempt to simplify the structure of the problem and simultaneously establish an upper bound for the value function. The dual formulation thus redirects our focus toward finding a non-negative semimartingale Y such that the product $X^\pi Y$ forms a supermartingale for every admissible control $\pi \in \mathcal{A}$, offering a probabilistic reformulation that is more amenable to deep learning techniques.

Building on classical duality ideas while carefully adapting them to account for regime-switching dynamics and jump discontinuities, we now derive the dual problem formulation that will serve as the foundation for our learning-based numerical methods.

2.4 Derivation of the Dual Problem for Regime-Switching UMP

Deriving the dual problem for a UMP with regime-switching is not very straightforward. Coming up with an ansatz, as done in (Wiedermann, 2022), is not really an option in this case, as the dual wealth process would now be adapted not only by a standard Brownian filtration but also by jump martingales (M_{ij}) associated with regime switching. Nonetheless, regime-switching-based dual problem models have been thoroughly explored by (Heunis, 2015), and we intend to use a similar method, mirroring and adapting it to the current UMP setup.

Generalized Primal Formulation and State-Space Enlargement

First, we rephrase and generalize the primal problem defined in Definition 2.7 as:

$$V(t,x,\epsilon) = \sup_{\pi\in\mathcal{A}} \mathbb{E}\left[\int_t^T f(s,\epsilon_s,X_s^\pi,\pi_s)\,ds + g(\epsilon_T,X_T^\pi)\,\middle|\, X_t^\pi = x, \epsilon_t = \epsilon\right],$$

where the objective of the primal problem is to maximize the total utility, which is composed of a running utility $f : [0,T]\times S\times\mathbb{R}\times K \to \mathbb{R}$ and a terminal reward $g : S\times\mathbb{R}\to\mathbb{R}$. We can recover the original primal version of the UMP form defined in definition 2.7 by setting $f = 0$ and $g = U$. Now, to systematically derive the dual formulation, we will enlarge our viewpoint beyond wealth processes strictly arising from admissible controls $(\mathcal{A})$. We introduce a broader class of candidate processes to set up a general variational framework. Let us say that $\mathbb{I}$ denotes the set of progressively measurable processes $(X_s)_{s\in[t,T]}$ satisfying the following:

$$X_s = x + \int_t^s \dot{X}_r\,dr + \int_t^s (\Lambda_r^X)^\top dB_r, \quad s\in[t,T],$$

where $\dot{X}_r$ is a progressively measurable $\mathbb{R}$-valued process (interpreted as the drift), Λ_r^X is a progressively measurable $\mathbb{R}^m$-valued process (interpreted as the diffusion). Furthermore, the processes are adapted to the filtration generated by $(W(s),\epsilon_s)$ and appropriate integrability conditions are satisfied, such as:

$$\mathbb{E}\left[\int_t^T |\dot{X}_s|^2\,ds + \int_t^T |\Lambda_s^X|^2\,ds\right] < \infty.$$

At this stage, we do not impose any additional structural constraints on $\dot{X}$ or Λ^X. In particular, they are not required to arise from a control π through the mappings

α and β. The purpose of this generalization is to allow for the application of variational methods without being restricted to wealth processes generated only by admissible controls. While the set $\mathbb{I}$ allows for a broad class of candidate wealth processes, it is also critical to introduce additional structure to capture the feasibility and admissibility constraints inherent to the primal problem. To that end, we define two subsets of $\mathbb{I}$:

- For the **feasibility of the wealth process**, we define a set $\mathbb{I}_1$ consisting of all processes $X \in \mathbb{I}$ that satisfy the initial condition $X_t = x$ and the non-negativity condition $X_s \geq 0$ almost surely for all $s \in [t, T]$. Formally, we can write:

$$\mathbb{I}_1 := \{X \in \mathbb{I} \,|\, X_t = x,\ X_s \geq 0 \text{ a.s. for all } s \in [t, T]\}\,.$$

- For the **admissibility of the wealth process**, we define a set $\mathbb{I}_2 \subset \mathbb{I}_1$, consisting of all wealth processes that can be generated by an admissible control $\pi \in \mathcal{A}$. This set represents the wealth processes achievable under admissible control policies, while simultaneously ensuring dynamic feasibility and non-negativity constraints. In particular, $X \in \mathbb{I}_2$ if there exists a progressively measurable control process $(\pi_s)_{s\in[t,T]}$ taking values in the convex set K such that:

$$\dot{X}_s = \alpha(s, \epsilon_s, X_s, \pi_s), \quad \Lambda_s^X = \beta(s, \epsilon_s, X_s, \pi_s), \quad \text{for almost every } s \in [t, T].$$

  Formally, we can write:

$$\mathbb{I}_2 := \left\{X \in \mathbb{I}_1 \,\middle|\, \exists \pi \in \mathcal{A} \text{ such that } \dot{X}_s = \alpha(\cdot),\ \Lambda_s^X = \beta(\cdot)\right\}.$$

Earlier in this section, we mentioned that we are going to focus on a more generalized horizon in terms of the wealth processes we will be considering. For that purpose, we will place more emphasis on $\mathbb{I}_1$. While the set $\mathbb{I}$ allows for a broad class of candidate wealth processes, and $\mathbb{I}_1$ captures processes satisfying initial and non-negativity conditions, not every process in $\mathbb{I}_1$ corresponds to a dynamically admissible wealth process associated with a valid control $\pi \in \mathcal{A}$. In other words, some processes in $\mathbb{I}_1$ may evolve arbitrarily, without adhering to the controlled dynamics dictated by the mappings α and β. If we were to minimize directly over $\mathbb{I}_1$ without additional restrictions, we would risk admitting wealth paths that are not attainable by any admissible control, leading to an incorrect characterization of the value function V.

To enforce these admissibility constraints on such processes, we have a penalty function in two stages. First, the local penalty L_1 (eq. 3.19) checks feasibility pointwise in time:

$$L_1(t, X(t), \dot{X}(t), \Lambda^X(t)) := \begin{cases} 0, & \text{if } X \text{ is admissible at time } t, \\ +\infty, & \text{otherwise.} \end{cases}$$

This ensures that X and the associated control are admissible at every instant. Next, the initial condition is enforced by

$$l_0(x) := \begin{cases} 0, & \text{if } x = x_0, \\ +\infty, & \text{otherwise.} \end{cases}$$

These two components are then aggregated into a single penalty function

$$\Phi_1(X) := l_0(X_0) + \mathbb{E} \int_0^T L_1(t, X(t), \dot{X}(t), \Lambda^X(t))\, dt.$$

By contruction we get that,

$$\Phi_1(X) := \begin{cases} 0, & \text{if } X \in \mathbb{I}_2, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2.56}$$

If X can be generated by an admissible control, null penalty is incurred. However, if X cannot be associated with any admissible control, the penalty is $+\infty$. With the help of this penalty function, we can enlarge the space of optimization variables from $\mathbb{I}_2$ to $\mathbb{I}_1$ which is convex and easier to handle analytically (rather than the nonlinear subset $\mathbb{I}_2$), while ensuring that the admissibility conditions are enforced implicitly via Φ_1.

With the penalty functional Φ_1 in place, we can now define the full cost function for the primal problem $\Phi : \mathbb{I}_1 \to \mathbb{R} \cup \{+\infty\}$ as follows:

$$\Phi(X) := \Phi_1(X) - \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X_s, \pi_s)\, ds + g(\epsilon_T, X_T)\right]. \tag{2.57}$$

Here, the first term, which is the penalty term, follows equation (2.56), ensuring that the cost function diverges when the wealth process is not admissible (i.e., outside $\mathbb{I}_2$), thereby automatically excluding it from consideration. Moreover, the second term represents the negative of the expected utility, thus preserving the original optimization goal. If X corresponds to an admissible control, $\Phi_1(X) = 0$, and $\Phi(X)$

reduces to the negative expected total utility. Consequently, the original primal utility maximization problem can now be reformulated as a minimization problem over the enlarged space $\mathbb{I}_1$:

$$-V(t,x,\epsilon) = \inf_{X\in\mathbb{I}_1} \Phi(X).$$

Dual Variables, Dual Cost Functional, and Weak Duality

Finally, to be able to construct the dual formulation of the primal problem, we introduce a class of candidate dual processes, denoted by $\mathbb{J}$. The idea of selecting a dual problem is to replace the optimization over controlled forward wealth processes X with an equivalent minimization over suitably chosen non-negative semimartingales Y, whose dynamics are compatible with the underlying filtration, including both the Brownian motion $W(s)$ and the jump martingales $M_{ij}(s)$ associated with regime switching. In that regard, we define $\mathbb{J}$ as the set of progressively measurable, $\mathbb{R}_+$-valued semimartingale processes $(Y_s)_{s\in[t,T]}$ of the form:

$$dY_s = \dot{Y}_s\, ds + (\Lambda_s^Y)^\top dW_s + \sum_{\substack{\epsilon_i,\epsilon_j\in S\\ i\neq j}} \Gamma_{ij}^Y(s)\, dM_{ij}(s) \tag{2.58}$$

where $\dot{Y}_s$ is a progressively measurable real-valued drift process, Λ_s^Y is a progressively measurable $\mathbb{R}^m$-valued diffusion process, and $\Gamma_{ij}^Y(s)$ are progressively measurable jump coefficients associated with regime switches from ϵ_i to ϵ_j. All processes are assumed to satisfy appropriate integrability and admissibility conditions. Furthermore, we will need to impose additional constraints on Y, such as martingale conditions or supermartingale properties relative to X, to ensure that they represent valid dual candidates.

Ultimately, we can now define an associated cost function $\Psi : \mathbb{J} \to \mathbb{R}\cup\{+\infty\}$, which represents the dual counterpart to the primal cost function Φ. Essentially, instead of maximizing the utility directly via X, we will seek to minimize a dual cost function associated with the process Y under appropriate constraints. For a given $Y \in \mathbb{J}$, we define

$$\begin{aligned}\Psi(Y) := \sup_{x_0\in\mathbb{R}^+} \left\{x_0 Y_t - \iota_{\{x\}}(x_0)\right\} + \mathbb{E}\left[\sup_{x\geq 0}\left\{g(\epsilon_T,x) - xY_T\right\}\right] \\ + \mathbb{E}\left[\int_t^T \sup_{\substack{x\geq 0\\ \pi\in K}} \left\{ -\ f(s,\epsilon_s,x,\pi) + x\,\dot{Y}_s + \alpha(s,\epsilon_s,x,\pi)Y_s + \beta(s,\epsilon_s,x,\pi)^\top \Lambda_s^Y\right\} ds\right]\end{aligned} \tag{2.59}$$

The first term in (2.59) arises from the dualization of the initial condition. Particularly, the indicator function $\iota_{\{x\}}(x_0)$ is 0 when $x_0 = x$ and $+\infty$ otherwise. The Legendre-Fenchel transform or the convex conjugate reduces the first term to the linear form xY_t, which represents the dual cost associated with the initial conditions. The second term corresponds to the dual representation of the terminal utility through the Legendre-Fenchel transform of g and the third term captures the dual form of the running utility over the horizon $[t, T]$, accounting for both drift and diffusion contributions. The Legendre-Fenchel transform plays a crucial role in formulating the dual version of the utility maximization problem. We first formally define the same as

Definition 2.4.1. Let $U : \mathbb{R}^+ \times S \to \mathbb{R}$ be a regime-dependent utility function. We assume that for each fixed regime $\epsilon \in S$, the map $x \mapsto U(x, \epsilon)$ belongs to $C^2(\mathbb{R}^+)$, is strictly increasing and strictly concave (definition 2.1.3), and satisfies the Inada conditions

$$\lim_{x \downarrow 0} \partial_x U(x, \epsilon) = +\infty, \qquad \lim_{x \uparrow \infty} \partial_x U(x, \epsilon) = 0.$$

For each $\epsilon \in S$, the associated Legendre–Fenchel transform $\widetilde{U}(\cdot, \epsilon) : \mathbb{R}^+ \to \mathbb{R}$ is defined by

$$\widetilde{U}(y, \epsilon) := \sup_{x \in \mathbb{R}^+} \{U(x, \epsilon) - xy\}, \qquad y \in \mathbb{R}^+.$$

In the dual cost formulation in equation (2.59), the optimization over π occurs explicitly. Moreover, the penalty induced from deviation from optimal dynamics is also captured through the supremum terms. On top of this, we can now introduce admissibility conditions on Y such that we ensure the validity of a weak duality relation.

Theorem 2.4.2. For any initial condition (t, x, ϵ), we have:

$$-V(t, x, \epsilon) \geq \inf_{Y \in \mathbb{J}} \Psi(Y),$$

where $V(t, x, \epsilon)$ denotes the value function of the primal problem, and $\Psi(Y)$ is the dual cost associated with a dual candidate process Y.

Proof. Proving the above theorem will codify the relation between the primal and dual cost formulations. For this purpose, let us take a $X \in \mathbb{I}_2$ be any admissible wealth process associated to some control policy $\pi \in \mathcal{A}$, and let $Y \in \mathbb{J}$ be any

admissible dual process. We can start by applying Itô rule to the product of $X_s Y_s$, and take into account the dynamics of X and Y:

$$d(X_s Y_s) = X_s\, dY_s + Y_s\, dX_s + d\langle X, Y\rangle_s$$

Substituting the dynamics of X and Y, we obtain

$$\begin{aligned} d(X_s Y_s) = {} & X_s \dot{Y}_s\, ds + X_s (\Lambda_s^Y)^\top dW_s + X_s \sum_{i\neq j} \Gamma_{ij}^Y(s)\, dM_{ij}(s) \\ & + Y_s \alpha(s, \epsilon_s, X_s, \pi_s)\, ds + Y_s \beta(s, \epsilon_s, X_s, \pi_s)^\top dW_s + d\langle X, Y\rangle_s \end{aligned} \tag{2.60}$$

In order to compute the quadratic covariation process $\langle X, Y\rangle_s$, we note by definition that only terms of order $\sqrt{ds}$ (diffusion parts) and simultaneous jumps contribute. The drift terms in both X and Y are of finite variation and therefore do not contribute to the bracket. Moreover, since X has no jump component, the jump contribution also vanishes. Thus only the diffusion terms remain. Applying the covariation rule for stochastic integrals, we obtain

$$d\langle X, Y\rangle_s = \beta(s, \epsilon_s, X_s, \pi_s)^\top\, d\langle W, W\rangle_s\, \Lambda_s^Y = \beta(s, \epsilon_s, X_s, \pi_s)^\top \Lambda_s^Y\, ds,$$

since $d\langle W, W\rangle_s = I\, ds$ for a standard Brownian motion. Substituting this into equation (2.60) we get:

$$\begin{aligned} d(X_s Y_s) = {} & \Big(X_s \dot{Y}_s + Y_s \alpha(s, \epsilon_s, X_s, \pi_s) + \beta(s, \epsilon_s, X_s, \pi_s)^\top \Lambda_s^Y \Big) ds \\ & + \Big(X_s (\Lambda_s^Y)^\top dW_s + Y_s \beta(s, \epsilon_s, X_s, \pi_s)^\top dW_s \Big) \\ & + X_s \sum_{i\neq j} \Gamma_{ij}^Y(s)\, dM_{ij}(s). \end{aligned} \tag{2.61}$$

Integrating from t to T on both sides and rearranging, we find

$$X_T Y_T - x Y_t = \int_t^T \Big(X_s \dot{Y}_s + Y_s \alpha(s, \epsilon_s, X_s, \pi_s) + \beta(s, \epsilon_s, X_s, \pi_s)^\top \Lambda_s^Y \Big) ds + M_{t,T},$$

where $M_{t,T}$ are all the local martingale terms:

$$M_{t,T} := \int_t^T X_s (\Lambda_s^Y)^\top dB_s + \int_t^T Y_s \beta(s, \epsilon_s, X_s, \pi_s)^\top dW_s + \int_t^T X_s \sum_{i\neq j} \Gamma_{ij}^Y(s)\, dM_{ij}(s).$$

Now, by taking conditional expectations and using the standard integrability assumptions (which ensure that the martingale terms have zero mean), we obtain

$$\begin{aligned} \mathbb{E}[M_{t,T}] := {} & \mathbb{E}[X_T Y_T] - x Y_t \\ & - \mathbb{E}\left[\int_t^T \Big(X_s \dot{Y}_s + Y_s \alpha(s, \epsilon_s, X_s, \pi_s) + \beta(s, \epsilon_s, X_s, \pi_s)^\top \Lambda_s^Y \Big) ds \right] \\ = {} & 0. \end{aligned} \tag{2.62}$$

Next, we apply the Legendre–Fenchel inequality [1](see Definition 2.4.1), which states that for all $x, y \geq 0$,

$$U(x) \leq \widetilde{U}(y) + xy.$$

Choosing $x = X_T$ and $y = Y_T$ yields the terminal bound

$$U(X_T) \leq \sup_{x \geq 0} \left\{ U(x) - xY_T \right\} + X_T Y_T. \tag{2.63}$$

An analogous argument for the running reward f gives

$$\begin{aligned} & f(s, \epsilon_s, X_s, \pi_s) + X_s \dot{Y}_s + \alpha(s, \epsilon_s, X_s, \pi_s) Y_s + \beta^\top(s, \epsilon_s, X_s, \pi_s) \Lambda_s^Y \\ & \quad \leq \sup_{\substack{x \geq 0 \\ \pi \in K}} \left\{ f(s, \epsilon_s, x, \pi) + x \dot{Y}_s + \alpha(s, \epsilon_s, x, \pi) Y_s + \beta^\top(s, \epsilon_s, x, \pi) \Lambda_s^Y \right\}. \end{aligned} \tag{2.64}$$

Taking expectations and integrating (2.64) over $[t, T]$, then combining with (2.63) and the identity (2.62), we obtain the primal–dual inequality

$$\begin{aligned} & - \left(\mathbb{E}[X_T Y_T] - x Y_t - \mathbb{E}\left[\int_t^T \left(X_s \dot{Y}_s + Y_s \alpha(s, \epsilon_s, X_s, \pi_s) + \beta(s, \epsilon_s, X_s, \pi_s)^\top \Lambda_s^Y \right) ds \right] \right) \\ & \qquad \leq \Psi(Y) + \Phi(x). \end{aligned} \tag{2.65}$$

Finally, this establishes the weak duality relation for any $Y \in \mathbb{J}$ and any admissible X,

$$\Psi(Y) + \Phi(x) \quad \geq \quad 0. \tag{2.66}$$

Taking the infimum over $Y \in \mathbb{J}$ yields

$$\inf_{X \in \mathbb{I}_2} \Phi(X) \ \geq \ - \inf_{Y \in \mathbb{J}} \Psi(Y).$$

Since the primal value function is given by

$$V(t, x, \epsilon) \ = \ - \inf_{X \in \mathbb{I}_2} \Phi(X),$$

we conclude that

$$-V(t, x, \epsilon) \ \geq \ \inf_{Y \in \mathbb{J}} \Psi(Y) \tag{2.67}$$

which establishes the weak duality bound and concludes the proof. □

[1]This inequality follows directly from the definition of the Legendre–Fenchel transform. Indeed, by Definition 2.4.1, $\widetilde{U}(y) = \sup_{z \geq 0}\{U(z) - zy\}$, so for any $x \geq 0$ we have $U(x) - xy \leq \widetilde{U}(y)$, which is equivalent to $U(x) \leq \widetilde{U}(y) + xy$.

The aforementioned theorem establishes weak duality: the dual problem always provides a lower bound to the primal value function. In contrast, strong duality, which corresponds to taking equality in equation (2.67), requires additional regularity conditions. In particular, one must guarantee the existence of an optimal dual process $Y^* \in \mathbb{J}$ attaining the infimum in (2.67), together with the absence of a duality gap (via a Slater-type condition or suitable constraint qualification), as well as convexity and lower semicontinuity of the cost functionals.

In the present regime-switching framework, implementing strong duality rigorously is highly nontrivial although theoritically possible. The discontinuous jump terms induced by the switching dynamics introduce technical challenges, and only under restrictive assumptions on the utility functions, the running reward f, and the system dynamics (α, β) can one obtain strong duality results. Such assumptions, however, defeat the purpose of this thesis, which seeks to treat UMPs in more general settings than previously available closed-form solutions. For this reason, rather than aiming for strong duality in full generality, we view the weak dual formulation as a powerful tool for numerical approximation. In particular, it naturally motivates the use of learning-based methods to approximate optimal controls which we will be exploring in the next section.

2.5 Algorithmic Formulation of the forward-BSDE approach

Weak duality provides a principle for learning-based approaches to solving the UMP. For any dual process $Y \in \mathbb{J}$, the dual functional $\Psi(Y)$ serves as a rigorous bound on the value function $V(t, x, \epsilon)$ (upper or lower, depending on the sign convention adopted). Optimizing over Y tightens the duality gap and, in practice, moves us toward near-optimal control without imposing the restrictive assumptions typically required to ensure strong duality. This sole principle essentially allows us to address a broader class of UMP problems than those amenable to strong-duality analyses. Building on this principle, we develop three different approaches:

1. **Primal FBSDE method:** Solve the forward–backward SDE system (see equations (2.41)) for the primal problem to directly approximate the value function and obtain an optimal control.

2. **Dual FBSDE method:** Solve the dual forward–backward system (cf. the dual counterpart of equations (2.41)) to minimize $\Psi(Y)$. Any optimal dual–process Y via first–order/verification relations, yields a corresponding optimal control.

3. **Gap–adapted primal–dual method (discussed in the next chapter):** Jointly learn a primal control π and a dual process Y, producing matched *a posteriori* bounds and an explicit estimate of the duality gap. A small gap implies higher near-optimality of π.

In order to execute these approaches, we first define a finite time grid $0 = t_0 < t_1 < \cdots < t_N = T$ and simulate the controlled dynamics using a standard scheme (ideally, Euler–Maruyama). Under this discretization, the unknown adapted objects that each of the three methods must learn, i.e., the control π and the relevant adjoint term Q, are $\mathcal{F}_{t_i}$-measurable and therefore can be modeled as functions of the local state and the regime. In learning-based methods, we adopt a unified learning parametrization in which controls and adjoint-related processes are represented by neural networks. For each time step $i \in \{0, \ldots, N-1\}$, we write

$$\pi_i = \mathcal{N}_{\theta_{i,\pi}}(t_i, \epsilon_i, X_i), \qquad Q_i = \mathcal{N}_{\theta_{i,Q}}(t_i, \epsilon_i, X_i).$$

Here, $\mathcal{N}_{\theta_{i,\pi}} : \mathbb{R}^d \to \mathbb{R}^m$ and $\mathcal{N}_{\theta_{i,Q}} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$.

Remark 2.5.1. The admissibility of the optimal solution is enforced by design (surjective networks mapping into the admissible set K), by reparameterization (e.g. softmax, tanh) and by differentiable projection layers, ensuring that the numerically obtained optimal control values remain in K. The regime ϵ_i is encoded either by one-hot vectors or by learned embeddings, and time t_i is represented using sinusoidal encodings to capture periodicity. At $t = 0$, since $\pi(0)$ and $Q(0)$ are $\mathcal{F}_0$-measurable, they can be represented by trivial networks consisting only of bias parameters. Architecturally, one may use a shared backbone with regime embeddings (leveraging common structure across regimes) or regime-specific output heads (tailored to each regime); this design choice affects both expressiveness and sample efficiency. *In our experiments, we implement and compare different architectural variants.*

Having specified how controls and adjoint-related processes are represented, we now describe a time-discrete scheme that both drives simulation and defines a differentiable computational graph for learning. To that end, we use the same time grid with step size $\Delta t := T/N$. As the initial values of the backward processes in the system of equations (2.41) are not known, we introduce optimization parameters v_0 and p_0 and initialize parameters for the primal FSBDE method as:

$$X_0 = x, \qquad v_0 = v_0, \qquad P_0 = p_0.$$

In the dual FBSDE method, we set $X_0 = y_0$, where y_0 denotes the initial value of the dual process Y. Based on this, the updates are:

1. **Forward state (Euler–Maruyama):**

$$X_{i+1} = X_i + \alpha(t_i, \epsilon_i, X_i, \pi_i)\,\Delta t + \beta^\top(t_i, \epsilon_i, X_i, \pi_i)\,\Delta W_i,$$

   with $\Delta W_i \sim \mathcal{N}(0, \Delta t\, I)$.

2. **Value process (with regime jumps):**

$$v_{i+1} = v_i - f(t_i, \epsilon_i, X_i, \pi_i)\,\Delta t + Z_i^\top \Delta W_i + \Delta M_i^\epsilon$$

3. **Adjoint process (with regime jumps):**

$$P_{i+1} = P_i - \nabla\mathcal{H}(t_i, \epsilon_i, X_i, \pi_i, P_i, Q_i)\,\Delta t + Q_i^\top \Delta W_i + \Delta M_i^{\epsilon,\nabla}$$

where $\epsilon_i \in \{1, \ldots, m\}$ denotes the regime at time t_i. Furthermore, the compensated jump increments of the regime-switching chain are, for i with $\epsilon_i = r$,

$$\Delta M_{rj,i} := \mathbf{1}\{\epsilon_i = r,\ \epsilon_{i+1} = j\} - q_{rj}(t_i)\,\mathbf{1}\{\epsilon_i = r\}\,\Delta t, \qquad j \neq r,$$

and they enter the backward updates via

$$\Delta M_i^\epsilon = \sum_{j\neq r} \big(v_i^{(j)} - v_i^{(r)}\big)\,\Delta M_{rj,i}, \qquad \Delta M_i^{\epsilon,\nabla} = \sum_{j\neq r} \big(P_i^{(j)} - P_i^{(r)}\big)\,\Delta M_{rj,i},$$

where $v_i^{(r)}$ and $P_i^{(r)}$ denote the components of the value and adjoint processes associated with regime r at time t_i. This discretization yields a simulation-based objective amenable to stochastic-gradient training of the neural parametrizations introduced above. An optimal solution must simultaneously satisfy: (i) the *discrete set of equations explained above* (the stepwise updates hold in conditional expectation, so the associated residuals vanish), (ii) the *terminal conditions* $v_{t_N} = g(\epsilon_{t_N}, X_{t_N})$ and $P_{t_N} = \nabla_x g(\epsilon_{t_N}, X_{t_N})$, and (iii) a *pointwise HJB argmax condition*: for each grid time t_i, almost surely given $\mathcal{F}_{t_i}$,

$$\pi_i \ \in\ \arg\max_{\pi\in K} \Big\{ f(t_i, \epsilon_i, X_i, \pi) \ +\ \widehat{\mathcal{L}}_i^\pi\, v(t_i, \epsilon_i, X_i) \Big\}, \tag{2.68}$$

where $\widehat{\mathcal{L}}_i^\pi\, v$ is the discrete generator. Furthermore, the weak duality principle provides a discrete dual functional whose minimization yields a computable bound via the primal–dual gap. Similar to (Wiedermann, 2022), to learn an optimal

solution we use a composite training objective consisting of a terminal L^2-fit for the backward variables (v_{t_N}, P_{t_N}), a per–time-step control objective based on the HJB argmax condition, and a discrete dual objective (for the dual FBSDE method). The formal versions of these three components are:

Terminal L^2 losses

$$\mathcal{L}_{\text{2BSDE}}(v_0, p_0, \{\theta_{i,\pi}, \theta_{i,Q}\}) := \widehat{\mathbb{E}}\Big[\big\|v_{t_N} - g(\epsilon_{t_N}, X_{t_N})\big\|^2 \;+\; \big\|P_{t_N} - \nabla_x g(\epsilon_{t_N}, X_{t_N})\big\|^2\Big], \tag{2.69}$$

where the hat on the expectation denotes the empirical Monte Carlo average over simulated trajectories. Note that X_{t_N} and (v_{t_N}, P_{t_N}) depend on the network parameters via the discretized dynamics explained earlier in this section.

Per–step control objective based on the HJB argmax condition For a maximization problem (primal FBSDE method), we minimize the negative of the expression in equation (2.68) at each step $i \in \{0, \dots, N-1\}$:

$$\mathcal{L}^{i}_{\text{control}}(\theta_{i,\pi}) := -\widehat{\mathbb{E}}\Big[\Big\{f(t_i, \epsilon_i, X_i, \pi) \;+\; \widehat{\mathcal{L}}_i^{\pi}\, v(t_i, \epsilon_i, X_i)\Big\}\Big] \tag{2.70}$$

In the case of a minimization problem (dual FBSDE method), the sign is reversed. Feasibility of $\pi_i \in K$ is ensured by the parameterizations described above.

Discrete dual objective Weak duality implies $-V(t, x, \epsilon) \geq \inf_Y \Psi(Y)$. In the dual *FBSDE* method, we treat the initial value x_0 as a trainable variable. Since the two losses above capture only the dynamics/terminal fit for a fixed initial value of the dual process Y, we additionally *minimize*, on the grid, the discrete dual functional $\Psi(Y_\eta)$:

$$\mathcal{L}_{\text{dual}}(\eta) := \widehat{\mathbb{E}}\big[\Psi\big(Y_\eta\big)\big]\,. \tag{2.71}$$

Under standard regularity conditions (Markovian coefficients, compact admissible set K, and appropriately rich networks), and as $\Delta t \to 0$, driving these losses to zero ensures consistency with the continuous optimality system, while the Monte Carlo averages provide unbiased estimators of the expectations.

Having specified the discretization scheme, the loss function, and the learning objectives for both the dual and the primal problem, we can now detail the algorithm and its learning loop in practice. Both learning procedures, i.e., the primal FBSDE method and the dual FBSDE method, share the same discretization and network parameterization but differ in the optimization direction (maximize vs. minimize) and in the presence of the additional dual steps.

Primal FBSDE method (maximization; refer Algorithm 1). Every training step for the primal FBSDE method consists of two steps:

1. **Terminal fit:** Simulate the trajectories using the updates above and minimize the terminal L^2 loss $\mathcal{L}_{\text{2BSDE}}$ in equation (2.69) to fit the backward variables. This updates the initialization parameters (v_0, p_0) and any network parameters participating in the backward rollout.

2. **Per–time–step control improvement:** For each time step in the grid, (re)simulate to t_i and update the control network $\theta_{i,\pi}$ by *maximizing* the objective function from (2.68) by minimizing its negative. Admissibility is enforced by the parameterizations described earlier in the section.

Dual FBSDE method (minimization; refer Algorithm 2). Similar to the primal method, the dual version has the first two steps with the optimization direction reversed, and includes an additional third step involving minimization of the dual cost functional:

1. **Terminal fit:** As above, simulate and minimize $\mathcal{L}_{\text{2BSDE}}$ to fit (v_{t_N}, P_{t_N}).

2. **Per–time–step control improvement:** For each time step, update $\theta_{i,\pi}$ by *minimizing* the discrete HJB argmax condition in equation (2.68).

3. **Dual cost functional minimization:** Treat y_0 as a trainable variable and *minimize* the discrete dual functional $\Psi(Y_\eta)$ as in (2.71) to obtain the parameters η.

Both methods, after each training step, provide value estimates—together with the current control policy—for the UMP, namely $\widehat{V}_{\text{primal}}$ and $\widehat{V}_{\text{dual}}$. For the primal estimate (flip signs for minimization setups), we freeze the control network at that step and compute a Monte Carlo average over N simulated steps, evaluating

$$\widehat{V}_{\text{primal}} \;:=\; \widehat{\mathbb{E}}\left[\sum_{k=0}^{N-1} f(t_k, \epsilon_k, X_k, \pi_k)\,\Delta t \;+\; g(\epsilon_N, X_N)\right]. \tag{2.72}$$

For the dual estimate, we simulate the dual process Y_η, compute the dual cost function, and then take its Monte Carlo average:

$$\widehat{V}_{\text{dual}} \;:=\; -\,\widehat{\mathbb{E}}\big[\Psi(Y_\eta)\big]\,. \tag{2.73}$$

In practice, we typically observe the sandwiching inequality $\widehat{V}_{\text{primal}} \le V \le \widehat{V}_{\text{dual}}$. The next theorem is a comparison theorem that provides an analytical explanation at the PDE level, by enlarging or restricting the admissible control set $\mathcal{A}$ for why the empirical primal/dual estimates tend to sandwich the optimum.

Theorem 2.5.2. Let $\mathcal{A}^- \subset \mathcal{A} \subset \mathcal{A}^+$ be admissible action sets satisfying the assumptions laid down in Theorem 2.2.7. Let v be the classical solution to the HJB equation (2.46) under $\mathcal{A}$, and let v^- and v^+ be the classical solutions to the corresponding HJB systems under $\mathcal{A}^-$ and $\mathcal{A}^+$, respectively. Then:

(i) **Monotonicity** For all $\epsilon_i \in S$ and all $(t,x) \in [0,T] \times \mathbb{R}^d$,

$$v^-(t,\epsilon_i,x) \;\le\; v(t,\epsilon_i,x) \;\le\; v^+(t,\epsilon_i,x).$$

(ii) **Residual signs stated in the explicit HJB form.** Let π^* be any measurable maximizer for the $\mathcal{A}$–HJB at (s,ϵ_i,x), i.e.

$$\pi^*(s,\epsilon_i,x) \;\in\; \arg\max_{\pi\in\mathcal{A}} \Big\{ \mathcal{L}^{\pi} v(s,\epsilon_i,x) \;+\; f(s,\epsilon_i,x,\pi) \Big\},$$

where $\mathcal{L}^{\pi}$ is as in (2.18) and where q_{ij} may (or may not) depend on π. Then, for every $i \in S$ and all (s,x),

$$\begin{aligned} &\frac{\partial v^-}{\partial s}(s,\epsilon_i,x) \;+\; \mathcal{L}^{\pi^*} v^-(s,\epsilon_i,x) \\ &\qquad + \sum_{j\in S} q_{ij}\Big(v^-(s,\epsilon_j,x) - v^-(s,\epsilon_i,x)\Big) \;\ge\; -f\big(s,\epsilon_i,x,\pi^*\big) \qquad (2.74) \\ &\frac{\partial v^+}{\partial s}(s,\epsilon_i,x) \;+\; \mathcal{L}^{\pi^*} v^+(s,\epsilon_i,x) \\ &\qquad + \sum_{j\in S} q_{ij}\Big(v^+(s,\epsilon_j,x) - v^+(s,\epsilon_i,x)\Big) \;\le\; -f\big(s,\epsilon_i,x,\pi^*\big), \qquad (2.75) \end{aligned}$$

with terminal equalities $v^{\pm}(T,\epsilon_i,x) = g(\epsilon_i,x)$. In particular, (2.74) says that v^- is a classical *supersolution* of the $\mathcal{A}$–HJB written in the explicit form of (2.46), and (2.75) says that v^+ is a classical *subsolution* of the same equation.

Proof. (i) Monotonicity Fix (t,ϵ_i,x) and define, for any admissible π,

$$J(\pi) := \mathbb{E}\left[\int_t^T f(s,\epsilon_s,X_s^{\pi},\pi_s)\,ds + g(\epsilon_T,X_T^{\pi})\right] \in \overline{\mathbb{R}} := \mathbb{R}\cup\{\pm\infty\}.$$

Set

$$S^- := \{J(\pi) : \pi \in \mathcal{A}^-\}, \qquad S := \{J(\pi) : \pi \in \mathcal{A}\}, \qquad S^+ := \{J(\pi) : \pi \in \mathcal{A}^+\}.$$

Since $\mathcal{A}^- \subset \mathcal{A} \subset \mathcal{A}^+$, we have $S^- \subset S \subset S^+$. By monotonicity of the supremum in $(\overline{\mathbb{R}}, \leq)$.[2]

$$\sup S^- \;\leq\; \sup S \;\leq\; \sup S^+.$$

Given that

$$v^-(t, \epsilon_i, x) = \sup S^-, \quad v(t, \epsilon_i, x) = \sup S, \quad v^+(t, \epsilon_i, x) = \sup S^+,$$

yields $v^-(t, \epsilon_i, x) \leq v(t, \epsilon_i, x) \leq v^+(t, \epsilon_i, x)$.

(ii) Residual signs in the explicit HJB form. For a vector function u and a control π^*, abbreviate the explicit π–operator as:

$$\mathfrak{G}^\pi[u](s, \epsilon_i, x) := \mathcal{L}^\pi u_i(s, \epsilon_i, x) + \sum_{j \in S} q_{ij}\big(s, x, \pi\big) \big(u(s, \epsilon_j, x) - u(s, \epsilon_i, x)\big) + f(s, \epsilon_i, x, \pi).$$

Then the $\mathcal{A}$–HJB can be written as

$$\frac{\partial v}{\partial s}(s, \epsilon_i, x) \;+\; \sup_{\pi \in \mathcal{A}} \mathfrak{G}^\pi[v](s, \epsilon_i, x) \;=\; 0, \qquad v_i(T, x) = g(\epsilon_i, x),$$

and similarly for v^- (with $\mathcal{A}^-$) and v^+ (with $\mathcal{A}^+$). Since $\mathcal{A}^- \subset \mathcal{A} \subset \mathcal{A}^+$, we have the pointwise supremum ordering

$$\sup_{\pi \in \mathcal{A}^-} \mathfrak{G}^\pi[u] \;\leq\; \sup_{\pi \in \mathcal{A}} \mathfrak{G}^\pi[u] \;\leq\; \sup_{\pi \in \mathcal{A}^+} \mathfrak{G}^\pi[u], \qquad \text{for any } u.$$

Since v^- solves its *own* HJB with $\mathcal{A}^-$,

$$\frac{\partial v^-}{\partial s}(s, \epsilon_i, x) \;+\; \sup_{\pi \in \mathcal{A}^-} \mathfrak{G}^\pi[v^-](s, \epsilon_i, x) \;=\; 0.$$

Replacing the smaller supremum by the larger one yields

$$\frac{\partial v^-}{\partial s}(s, \epsilon_i, x) \;+\; \sup_{\pi \in \mathcal{A}} \mathfrak{G}^\pi[v^-](s, \epsilon_i, x) \;\geq\; 0.$$

Let $\pi^*(s, \epsilon_i, x) \in \arg\max_{\pi \in \mathcal{A}} \{\mathcal{L}^\pi v(s, \epsilon_i, x) + f(s, \epsilon_i, x, \pi)\}$ be any measurable maximizer for the $\mathcal{A}$–HJB (attainment holds under the standing compactness/continuity assumptions). Evaluating the supremum at π^* gives

$$\frac{\partial v^-}{\partial s}(s, \epsilon_i, x) \;+\; \mathfrak{G}^{\pi^*(s, \epsilon_i, x)}[\,v^-\,](s, \epsilon_i, x) \;\geq\; 0,$$

which is

$$\begin{aligned}\frac{\partial v^-}{\partial s}(s, \epsilon_i, x) \;&+\; \mathcal{L}^{\pi^*(s, \epsilon_i, x)} v^-(s, \epsilon_i, x) \\ &+\; \sum_{j \in S} q_{ij}\big(s, x, \pi^*\big) \big(v^-(s, \epsilon_j, x) - v^-(s, \epsilon_i, x)\big) \;\geq\; -f\big(s, \epsilon_i, x, \pi^*\big),\end{aligned}$$

[2] With the standard convention $\sup \emptyset = -\infty$.

i.e., (2.74). Thus v^- is a classical supersolution of the $\mathcal{A}$–HJB in the explicit form (2.46). Similarly, since v^+ solves its *own* HJB with $\mathcal{A}^+$,

$$\frac{\partial v^+}{\partial s}(s,\epsilon_i,x) \;+\; \sup_{\pi\in\mathcal{A}^+} \mathfrak{G}^\pi[v^+](s,\epsilon_i,x) \;=\; 0,$$

and the supremum ordering gives

$$\frac{\partial v^+}{\partial s}(s,\epsilon_i,x) \;+\; \sup_{\pi\in\mathcal{A}} \mathfrak{G}^\pi[v^+](s,\epsilon_i,x) \;\le\; 0.$$

Evaluating at the same π^* yields

$$\begin{aligned}\frac{\partial v^+}{\partial s}(s,\epsilon_i,x) \;&+\; \mathcal{L}^{\pi^*(s,\epsilon_i,x)} v^+(s,\epsilon_i,x) \\ &+\; \sum_{j\in S} q_{ij}\big(s,x,\pi^*\big)\,\big(v^+(s,\epsilon_j,x) - v^+(s,\epsilon_i,x)\big) \;\le\; -f\big(s,\epsilon_i,x,\pi^*\big),\end{aligned}$$

which is (2.75). The terminal equalities $v^\pm(T,\epsilon_i,x) = g(\epsilon_i,x)$ follow from the boundary data of their respective HJB problems. □

Remark 2.5.3. If the classical regularity assumptions fail (e.g., degenerate ellipticity, non–attained suprema, or lack of $C^{1,2}$ smoothness), interpret $v, v^\pm$ as viscosity objects. Then v^- is a viscosity *supersolution* and v^+ a viscosity *subsolution* of the $\mathcal{A}$–HJB in the explicit form (2.46), and the value–monotonicity $v^- \le v \le v^+$ still follows from control inclusion $\mathcal{A}^- \subset \mathcal{A} \subset \mathcal{A}^+$.

Remark 2.5.4. In part (ii) of the proof above we fix any measurable maximizer $\pi^*(s,\epsilon_i,x) \in \arg\max_{\pi\in\mathcal{A}}\{\mathcal{L}^\pi v(s,\epsilon_i,x) + f(s,\epsilon_i,x,\pi)\}$ for the $\mathcal{A}$–HJB. Since

$$\partial_s v^- + \sup_{\pi\in\mathcal{A}^-} \mathfrak{G}^\pi[v^-] = 0 \quad \text{and} \quad \mathcal{A}^- \subset \mathcal{A},$$

we have

$$\partial_s v^- + \sup_{\pi\in\mathcal{A}} \mathfrak{G}^\pi[v^-] \ge 0$$

and evaluating this (larger) supremum at the same π^* yields the claimed residual inequality for v^- regardless of whether $\pi^* \in \mathcal{A}^-$. The argument for v^+ is analogous with the reverse inequality. If the maximum over $\mathcal{A}$ is not attained, one can use ε–maximizers and let $\varepsilon \downarrow 0$.

Practically, when we run the primal FBSDE method, we are optimizing over a restricted policy class $\mathcal{A}_{\text{alg}} \subset \mathcal{A}$. In accordance with Theorem 2.5.2 with $\mathcal{A}^- = \mathcal{A}_{\text{alg}}$, we obtain, for all (t,ϵ_i,x),

$$v^{\text{alg}}(t,\epsilon_i,x) := \sup_{\pi\in\mathcal{A}_{\text{alg}}} \mathbb{E}\left[\int_t^T f(s,\epsilon_s,X_s^\pi,\pi_s)\,ds + g(\epsilon_T,X_T^\pi)\right] \;\le\; v(t,\epsilon_i,x).$$

Algorithm 1 One training step: Primal FBSDE (maximization, regime switching)

Require: Batch size b_{size}, grid $\{t_i\}_{i=0}^{N}$, step Δt; trainable (v_0, p_0); networks $\{\mathcal{N}_{\theta_{i,\pi}}, \mathcal{N}_{\theta_{i,Q}}\}_{i=0}^{N-1}$

Generate b_{size} paths of Brownian increments $\{\Delta W_i^j\}$ and regime jumps $\{\Delta M_i^{\epsilon,j}, \Delta M_i^{\epsilon,\nabla,j}\}$ using $q_{ij}(t_i, \cdot)$

▷ Substep 1: terminal L^2 fit (minimize $\mathcal{L}_{\text{2BSDE}}$ in (2.69))

for $j = 1$ **to** b_{size} **do**
  $X_0^j \leftarrow x;\ v_0^j \leftarrow v_0;\ P_0^j \leftarrow p_0$
end for
for $i = 0$ **to** $N-1$ **do**
  for $j = 1$ **to** b_{size} **do**
    $\pi_i^j \leftarrow \mathcal{N}_{\theta_{i,\pi}}(t_i, \epsilon_i^j, X_i^j)$
    $Q_i^j \leftarrow \mathcal{N}_{\theta_{i,Q}}(t_i, \epsilon_i^j, X_i^j)$
    Propagate $(X_{i+1}^j, v_{i+1}^j, P_{i+1}^j)$ via the discretized updates
  end for
end for
$\text{loss}_1 \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \left(\|v_{t_N}^j - g(\epsilon_{t_N}^j, X_{t_N}^j)\|^2 + \|P_{t_N}^j - \nabla_x g(\epsilon_{t_N}^j, X_{t_N}^j)\|^2 \right)$
Update (y_0, p_0) and any parameters used in Substep 1 with one SGD step w.r.t. loss_1

▷ Substep 2: per–time–step control improvement (maximize HJB stage objective)

for $i = 0$ **to** $N-1$ **do**
  for $j = 1$ **to** b_{size} **do**
    if $i = 0$ **then**
      $X_0^j \leftarrow x;\ v_0^j \leftarrow y_0;\ P_0^j \leftarrow p_0$
    else
      Re-simulate to t_i to obtain (X_i^j, v_i^j, P_i^j)
    end if
    $\pi_i^j \leftarrow \mathcal{N}_{\theta_{i,\pi}}(t_i, \epsilon_i^j, X_i^j)$
    $Q_i^j \leftarrow \mathcal{N}_{\theta_{i,Q}}(t_i, \epsilon_i^j, X_i^j)$
  end for
  $\text{loss}_{2,i} \leftarrow -\frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \left\{ f(t_i, \epsilon_i, X_i, \pi) \ + \ \widehat{\mathcal{L}}_i^{\pi}\, v(t_i, \epsilon_i, X_i) \right\}$
  Update $\theta_{i,\pi}$ with one SGD step w.r.t. $\text{loss}_{2,i}$
end for

Algorithm 2 One training step: Dual FBSDE (minimization, regime switching)

Require: Batch size b_{size}, grid $\{t_i\}_{i=0}^{N}$, step Δt; dual parameters η; trainable x_0 (if used)

Generate b_{size} paths of $\{\Delta W_i^j\}$ and regime jumps $\{\Delta M_i^{\epsilon,j}\}$

▷ Substep 1: terminal L^2 fit (same structure as primal)

for $j = 1$ **to** b_{size} **do**

  $X_0^j \leftarrow y_0;\ v_0^j \leftarrow v_0;\ P_0^j \leftarrow p_0$

end for

for $i = 0$ **to** $N-1$ **do**

  for $j = 1$ **to** b_{size} **do**

    $\pi_i^j \leftarrow \mathcal{N}_{\theta_{i,\pi}}(t_i, \epsilon_i^j, X_i^j)$

    $Q_i^j \leftarrow \mathcal{N}_{\theta_{i,Q}}(t_i, \epsilon_i^j, X_i^j)$

    Propagate $(X_{i+1}^j, v_{i+1}^j, P_{i+1}^j)$ via the discretized updates

  end for

end for

$\text{loss}_1 \leftarrow$ as in Alg. 1; update (v_0, p_0) with one SGD step

▷ Substep 2: per–time–step control update (minimization form)

for $i = 0$ **to** $N-1$ **do**

  for $j = 1$ **to** b_{size} **do**

    if $i = 0$ **then**

      $X_0^j \leftarrow x;\ v_0^j \leftarrow y_0;\ P_0^j \leftarrow p_0$

    else

      Re-simulate to t_i to obtain (X_i^j, V_i^j, P_i^j)

    end if

    $\pi_i^j \leftarrow \mathcal{N}_{\theta_{i,\pi}}(t_i, \epsilon_i^j, X_i^j)$

    $Q_i^j \leftarrow \mathcal{N}_{\theta_{i,Q}}(t_i, \epsilon_i^j, X_i^j)$

  end for

  $\text{loss}_{2,i} \leftarrow +\frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \left\{ f(t_i, \epsilon_i, X_i, \pi) \ + \ \widehat{\mathcal{L}}_i^{\pi}\, v(t_i, \epsilon_i, X_i) \right\}$ ▷ flip sign vs. primal

  Update $\theta_{i,\pi}$ with one SGD step w.r.t. $\text{loss}_{2,i}$

end for

▷ Substep 3: minimize the discrete dual functional

for $j = 1$ **to** b_{size} **do**

  $X_0^j \leftarrow y_0$ ▷ y_0 trainable if applicable

end for

for $i = 0$ **to** $N-1$ **do**

  for $j = 1$ **to** b_{size} **do**

    compute π_i^j for the dual dynamics

    Propagate X_{i+1}^j (and dual state variables) according to the dual discretization

  end for

end for

$\text{loss}_{\text{dual}} \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Psi\big(Y_\eta;\ \text{path}^j\big)$

Update (η, y_0) with one SGD step w.r.t. $\text{loss}_{\text{dual}}$

Validation (optional): compute the *dual value function* $\widehat{V}_{\text{dual}} = -\widehat{\mathbb{E}}\left[\Psi(Y_\eta)\right]$

Therefore, the value function we obtain from the learned primal policy, as the number of training steps tends to infinity, targets v^{alg} and furnishes a lower bound on the true value. In the case of the dual FBSDE method, we can also obtain an upper bound based on the corollary below, in which the monotonicity in the theorem is reversed.

Corollary 2.5.5. Let $\mathbb{J}$ denote the admissible class for the dual minimization (as in the dual formulation introduced above), and let $\mathbb{J}_{\mathrm{alg}} \subset \mathbb{J}$ be the restricted dual class used by the algorithm. Define the dual value

$$D(\mathbb{J}) := \inf_{Y \in \mathbb{J}} \Psi(Y),$$

so, under strong duality conditions, $v(t, \epsilon_i, x) = D(\mathbb{J})$. By the same inclusion argument as in Theorem 2.5.2, but with an infimum, the *set inclusion reverses the order*:

$$\mathbb{J}_{\mathrm{alg}} \subset \mathbb{J} \quad \Longrightarrow \quad D(\mathbb{J}) \;\leq\; D(\mathbb{J}_{\mathrm{alg}}).$$

Consequently, any dual estimate computed over $\mathbb{J}_{\mathrm{alg}}$ is an *upper bound*:

$$v(t, \epsilon_i, x) \;=\; D(\mathbb{J}) \;\leq\; D(\mathbb{J}_{\mathrm{alg}}).$$

Remark 2.5.6. Let $\widehat{V}_{\mathrm{pr}}(\mathcal{A}_{\mathrm{alg}})$ and $\widehat{V}_{\mathrm{du}}(\mathbb{J}_{\mathrm{alg}})$ denote the primal (lower) and dual (upper) estimators obtained with the algorithmic classes $\mathcal{A}_{\mathrm{alg}} \subset \mathcal{A}$ and $\mathbb{J}_{\mathrm{alg}} \subset \mathbb{J}$, respectively. If the classes are enlarged, i.e. $\mathcal{A}^{(1)}_{\mathrm{alg}} \subseteq \mathcal{A}^{(2)}_{\mathrm{alg}}$ and $\mathbb{J}^{(1)}_{\mathrm{alg}} \subseteq \mathbb{J}^{(2)}_{\mathrm{alg}}$, then

$$\widehat{V}_{\mathrm{pr}}\big(\mathcal{A}^{(1)}_{\mathrm{alg}}\big) \;\leq\; \widehat{V}_{\mathrm{pr}}\big(\mathcal{A}^{(2)}_{\mathrm{alg}}\big) \;\leq\; v \;\leq\; \widehat{V}_{\mathrm{du}}\big(\mathbb{J}^{(2)}_{\mathrm{alg}}\big) \;\leq\; \widehat{V}_{\mathrm{du}}\big(\mathbb{J}^{(1)}_{\mathrm{alg}}\big),$$

so the empirical sandwich $\widehat{V}_{\mathrm{pr}} \leq v \leq \widehat{V}_{\mathrm{du}}$ tightens monotonically. Conversely, shrinking either class can only widen the sandwich (the primal bound weakens downward and/or the dual bound weakens upward).

Finally, the resulting duality gap,

$$\widehat{V}_{\mathrm{gap}} := \widehat{V}_{\mathrm{dual}} - \widehat{V}_{\mathrm{primal}},$$

serves as a *certificate of sub-optimality at the level of the value function*. In particular, it provides an explicit upper bound on the difference between the optimal value and the performance of the learned admissible control. The gap-adapted variant (explained in detail in the next chapter), which combines both primal and dual procedures, explicitly aims to shrink this duality gap. While the duality gap provides

a quantitative certificate of sub-optimality in terms of value, it does not, by itself, guarantee convergence of the learned control to a globally optimal policy. To formalize the sense in which the primal and dual estimates sandwich the true value function and to explain why shrinking the gap corresponds to tightening these bounds we now establish a comparison result at the level of the underlying HJB equations.

Theorem 2.5.7. Let $\pi_{\text{alg}} \in \mathcal{A}_{\text{alg}} \subset \mathcal{A}$ be any admissible control produced by the algorithm, and let $\widehat{V}_{\text{primal}}$ and $\widehat{V}_{\text{dual}}$ denote the empirical primal and dual estimators defined in (2.72) and (2.73), respectively. Then, for all (t, ϵ_i, x),

$$0 \;\le\; V(t, \epsilon_i, x) - J(\pi_{\text{alg}}) \;\le\; \widehat{V}_{\text{dual}} - \widehat{V}_{\text{primal}} \;=\; \widehat{V}_{\text{gap}},$$

where $J(\pi_{\text{alg}})$ denotes the performance of the learned control:

$$J(\pi_{\text{alg}}) := \mathbb{E}\left[\int_t^T f(s, \epsilon_s, X_s^{\pi_{\text{alg}}}, \pi_{\text{alg},s})\, ds + g(\epsilon_T, X_T^{\pi_{\text{alg}}})\right].$$

Consequently, a vanishing duality gap implies convergence of the learned control to optimality in value.

Proof. By definition of the value function,

$$V(t, \epsilon_i, x) = \sup_{\pi \in \mathcal{A}} J(\pi) \;\ge\; J(\pi_{\text{alg}}),$$

which yields the left inequality.

By weak duality, for any admissible dual process $Y \in \mathbb{J}$,

$$V(t, \epsilon_i, x) \;\le\; -\Psi(Y).$$

In particular, evaluating this bound at the learned dual process Y_η produced by the algorithm gives

$$V(t, \epsilon_i, x) \;\le\; \widehat{V}_{\text{dual}}.$$

Combining the two inequalities and using $\widehat{V}_{\text{primal}} = J(\pi_{\text{alg}})$ yields

$$0 \;\le\; V(t, \epsilon_i, x) - J(\pi_{\text{alg}}) \;\le\; \widehat{V}_{\text{dual}} - \widehat{V}_{\text{primal}} = \widehat{V}_{\text{gap}},$$

which proves the claim. □

Theorem 2.5.7 provides a rigorous guarantee of near-optimality in terms of the achieved value. In general, however, small value sub-optimality does *not* imply that

the corresponding control is close to an optimal control in any strong norm[3], as the control-to-value map need not be injective or strongly concave. Moreover, the duality gap controls only a global, expectation-level discrepancy in value and does not imply pointwise convergence of the value function or improved satisfaction of the associated HJB equation. Quantitative bounds on the distance between controls require additional structural assumptions, such as strict concavity of the Hamiltonian and uniqueness or stability of the optimizer. In practice, evaluating the quality of the learned control therefore necessitates complementary diagnostics beyond the duality gap.

2.6 Numerical Experiments

We now turn to numerical experiments to support all the theory in this chapter. The objectives are to:

1. Validate the DPP/HJB derivation and verification theorem on benchmark examples with well known structure and optimal solution
2. Implementation of the BSDE-based learning algorithms under regime-switching
3. Assess robustness to modeling different coefficients, constraints sets etc.

w

Experiment 1: Benchmarking

Consider an observable two-regime market (Bull, Bear) driven by a continuous-time Markov chain $\epsilon_t \in \{\epsilon_0, \epsilon_1\}$ with generator

$$Q = \begin{pmatrix} -\lambda_{01} & \lambda_{01} \\ \lambda_{10} & -\lambda_{10} \end{pmatrix}, \quad (\lambda_{01}, \lambda_{10}) = (0.3, 0.5)$$

In any regime ϵ_i, the risky asset has parameters $(\mu_{\epsilon_i}, \sigma_{\epsilon_i}, r_{\epsilon_i})$, where

Bull (R0): $(\mu_{\epsilon_0}, \sigma_{\epsilon_0}, r_{\epsilon_0}) = (0.11, 0.20, 0.04)$, Bear (R1): $(\mu_{\epsilon_1}, \sigma_{\epsilon_1}, r_{\epsilon_1}) = (0.06, 0.30, 0.01)$

[3] This effect already arises in standard utility maximization problems. For instance, in a one-period expected utility maximization problem with log utility and terminal wealth of the form $X^\pi = 1 + \pi Z$, where Z is a noisy return with small mean and large variance, the value functional $J(\pi) = \mathbb{E}[\log(1 + \pi Z)]$ is strictly concave but nearly flat around its unique maximizer π^*. As an illustration, taking $Z \sim \mathcal{N}(0.02, 0.4^2)$ yields a unique optimizer $\pi^* \approx 0.12$, while the suboptimal control $\pi = 0$ attains a value within 10^{-4} of the optimum despite being far from π^* in norm. Therefore controls that are far from π^* can achieve values arbitrarily close to the optimum. A small duality gap therefore certifies near-optimality in value while providing no guarantee of proximity of the learned control to the optimal policy.

The transition rates $(\lambda_{01}, \lambda_{10}) = (0.3, 0.5)$ imply a stationary distribution $\pi_0 = \frac{\lambda_{10}}{\lambda_{01}+\lambda_{10}} = 0.625$ indicating that the Markov chain spends approximately 62.5% of the total time in the Bull regime (R0). This choice reflects the empirical observation that equity markets tend to remain in expansionary (Bull) states for longer periods than in contractionary (Bear) states. The investor seeks to maximize $\mathbb{E}[U(X_T)]$ (expected terminal utility of wealth) under CRRA preferences $U(x) = \frac{x^{1-\gamma}}{1-\gamma}$ over a time horizon $T = 1$, with initial wealth $x_0 = 1$, and without any constraint on the control π. The wealth fraction π_t evolves according to

$$\frac{dX_t}{X_t} = r_{\epsilon_t}\, dt + \pi_t(\mu_{\epsilon_t} - r_{\epsilon_t})\, dt + \pi_t \sigma_{\epsilon_t}\, dW_t \tag{2.76}$$

This problem has been extensively studied in the literature, which provides a closed-form characterization of both the value function and the corresponding optimal control (see (Jacka and Ocejo, 2018))

To derive the value function, let $V_i(t,x) := V(t,\epsilon_i,x) = \sup_\pi \mathbb{E}[U(X_T) \mid X_t = x,\ \epsilon_t = \epsilon_i]$. The regime-coupled HJB equation for a candidate function v reads

$$\begin{aligned} 0 = \frac{\partial v}{\partial t}(t,\epsilon_i,x) + \sup_\pi \left\{ \left(r_{\epsilon_i} x + \pi x(\mu_{\epsilon_i} - r_{\epsilon_i})\right) \frac{\partial v}{\partial x}(t,\epsilon_i,x) + \frac{1}{2}(\pi\sigma_{\epsilon_i} x)^2 \frac{\partial^2 v}{\partial x^2}(t,\epsilon_i,x) \right\} \\ + \sum_{j\neq i} q_{ij}\big(v(t,\epsilon_j,x) - v_i(t,\epsilon_i,x)\big), \qquad v(T,\epsilon_i,x) = U(x). \end{aligned} \tag{2.77}$$

Considering the ansatz $v(t,\epsilon_i,x) = A_{\epsilon_i}(t)U(x)$ and using the CRRA identities $xU_x = (1-\gamma)U$ and $x^2U_{xx} = -\gamma(1-\gamma)U$, the HJB simplifies to

$$0 = A'_{\epsilon_i}(t) + A_{\epsilon_i}(t)(1-\gamma)\sup_\pi \left\{ r_{\epsilon_i} + \pi(\mu_{\epsilon_i} - r_{\epsilon_i}) - \tfrac{\gamma}{2}\sigma^2_{\epsilon_i}\pi^2 \right\} + \sum_{j\neq i} q_{ij}(A_{\epsilon_j} - A_{\epsilon_i}).$$

Optimizing the concave quadratic yields the regime-dependent optimal control $\pi^*_{\epsilon_i} = \frac{\mu_{\epsilon_i} - r_{\epsilon_i}}{\gamma\sigma^2_{\epsilon_i}}$, and the corresponding maximized value $r_{\epsilon_i} + \frac{(\mu_{\epsilon_i} - r_{\epsilon_i})^2}{2\gamma\sigma^2_{\epsilon_i}}$ For notational simplicity, define

$$c_i = (1-\gamma)\left[r_{\epsilon_i} + \frac{(\mu_{\epsilon_i} - r_{\epsilon_i})^2}{2\gamma\sigma^2_{\epsilon_i}} \right], \qquad C = \mathrm{diag}(c_0, c_1),$$

which yields the coupled linear system:

$$A'(t) = -(C+Q)A(t), \qquad A(T) = \mathbf{1}.$$

The reduction to a linear system of ODEs for the coefficients $A_{\epsilon_i}(t)$ arises from the homogeneity of the CRRA utility function, as discussed in (Jacka and Ocejo, 2018). Solving backward in time (or equivalently forward in $s = T - t$) gives the closed-form solution

$$A(t) = \exp\big((C+Q)(T-t)\big)\mathbf{1}, \qquad V(t,\epsilon_i,x) = A_{\epsilon_i}(t)U(x), \qquad \pi^*_{\epsilon_i} = \frac{\mu_{\epsilon_i} - r_{\epsilon_i}}{\gamma\sigma^2_{\epsilon_i}}.$$

For $\gamma = \frac{1}{2}$ (square-root utility), this simplifies to $U(x) = 2\sqrt{x}$ and

$$c_i = \tfrac{1}{2}\left[r_{\epsilon_i} + \frac{(\mu_{\epsilon_i} - r_{\epsilon_i})^2}{\sigma^2_{\epsilon_i}}\right].$$

The analytically derived regime-switching result above serves as a benchmark for validating our algorithms. From the expressions obtained, the ground truth for the Regime-Switching (RS) model is

$$V^{\mathrm{RS}}(0,\epsilon_i,x=1) = A_{\epsilon_i}(0)\cdot 2, \qquad A(0) = \exp\big((C+Q)T\big)\mathbf{1}.$$

We can also obtain the closed-form solution for the case with no regime switching (NRS) (Pham, 2009) (when we start in a given regime and remain there without transitions) by setting $Q = 0$. In this case,

$$A^{\mathrm{NS}}_{\epsilon_i}(0) = \exp(c_i T), \qquad V^{\mathrm{NRS}}(0,\epsilon_i,1) = 2A^{\mathrm{NRS}}_{\epsilon_i}(0).$$

In order to validate the correctness and benchmark our neural network–based solutions, we compare both the primal and dual versions of our algorithm against the analytical regime-switching Merton portfolio solution derived above. Figure 2.1 summarizes the convergence behavior of both FBSDE methods under *non-regime switching (NRS)* and *regime switching (RS)* settings. As seen in the figure, both the primal and dual FBSDE methods exhibit an initial overshoot within the first 100 training steps, after which the value function estimates stabilize tightly around their theoretical benchmarks. The zoomed-in panels in the bottom row of the figure highlight this convergence region after 1000–1500 training steps. Across all four configurations, the neural estimates remain within a narrow neighborhood of the analytical values:

$$V^{\mathrm{NRS}}_{\epsilon_0} = 2.1693, \quad V^{\mathrm{NRS}}_{\epsilon_1} = 2.0381, \quad V^{\mathrm{RS}}_{\epsilon_0} = 2.1538, \quad V^{\mathrm{RS}}_{\epsilon_1} = 2.0634.$$

The close alignment of both primal and dual curves with these analytical references confirms the robustness and consistency of the two formulations. The

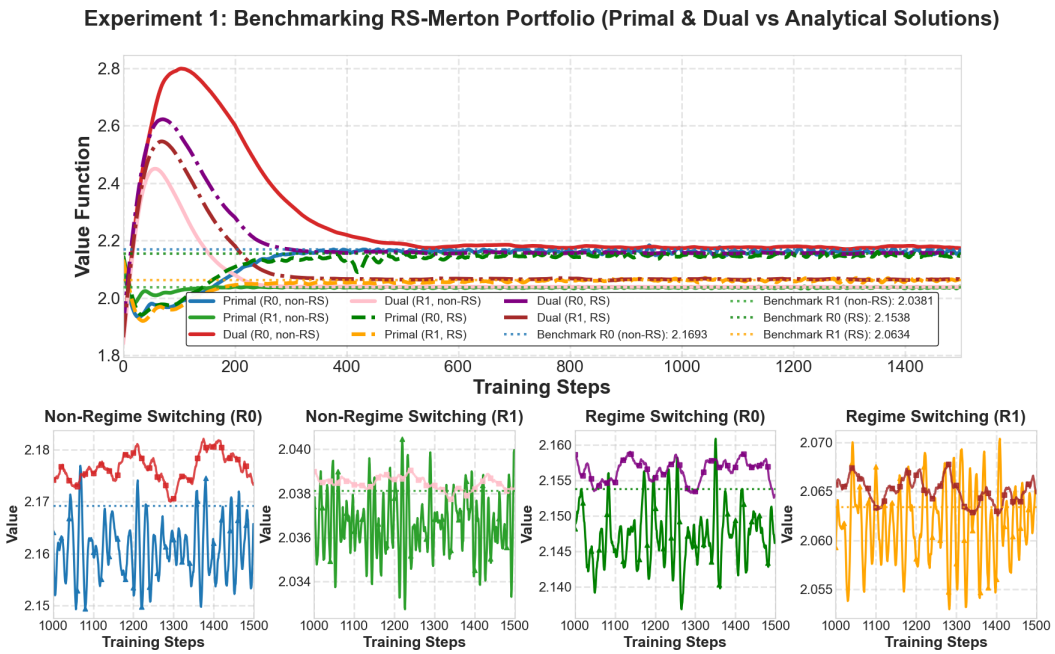

Figure 2.1: **Experiment 1: Benchmarking RS–Merton Portfolio (Primal & Dual vs Analytical Solutions).** Convergence of the primal and dual Deep FSBDE under both non-regime switching (NRS) and regime switching (RS) configurations. The upper panel shows relatively quick convergence of the value function across training steps; the lower panels zoom in on the steady-state region (1000–1500 steps) for each regime. Horizontal dashed lines denote analytical benchmark values.

small-amplitude oscillations, which differ slightly across the various settings (primal vs. dual and RS vs. NRS), are characteristic of stochastic mini-batch training and learning-rate annealing effects. An important point to note is that, since all hyperparameters, including learning rates, batch sizes, and optimizer schedules, were manually tuned during training, the observed differences in convergence patterns across all the different settings directly reflect distinct learning dynamics rather than structural differences in the algorithms. For example, the zoomed-in plots capture subtle differences in the oscillation frequency and damping behavior between the primal and dual networks, illustrating how variations in learning rates and update magnitudes influence the fine-scale evolution of the training trajectory of each algorithm. Qualitatively, the dual FBSDE often exhibits smoother empirical learning trajectories and lower variance across training runs, reflecting reduced sensitivity to forward-path noise in these experiments, as discussed earlier. Nonetheless, both algorithms approach nearly identical final values, setting a strong benchmarking example. Figure 2.3 illustrates the dependence of the value function on the transition

intensities $(\lambda_{01}, \lambda_{10})$ when starting in each regime. When the market begins in the *Bull* regime (R0) (left panel), increasing λ_{01} (Bull→Bear rate) decreases the value function since the investor faces a higher probability of entering a less favorable state. In contrast, increasing λ_{10} (Bear→Bull) slightly improves the value, as more frequent recovery to the Bull state enhances expected returns.

In addition to benchmarking the models with the generator Q, we also study the robustness of the learning process of both the primal and dual FBSDE formulations. We do this by analysing their convergence under systematic perturbations in the transition rates $(\lambda_{01}, \lambda_{10})$. In particular, the estimation of Q can be extremely noisy, and even a small change can significantly alter the expected number of regime transitions along a single trajectory. Figure 2.2 presents two robustness measures. Panel (a) varies the two transition rates jointly across a grid of values (along the y-axes) and plots the resulting duality gap $|v^{\text{Primal}} - v^{\text{Dual}}|$ for trajectories starting in both regimes. The duality gap is smallest when the transition structure is highly asymmetric (e.g., $\lambda_{01} \ll \lambda_{10}$ or $\lambda_{10} \ll \lambda_{01}$). In these cases, the Markov chain spends the vast majority of its time in a single regime, and the dynamics exhibit very few transitions. As a result, the primal and dual FBSDE formulations produce very similar estimates, leading to a very small duality gap.

On the other hand, when the intensities are comparable ($\lambda_{01} \approx \lambda_{10}$), the system frequently switches between the underlying regimes, creating higher path variance and a more challenging coupled forward–backward learning problem. This is precisely where we observe larger duality gaps. In essence, the observed duality gap correlates strongly with the frequency of regime transitions within a given time horizon, reflecting the increased order of difficulty of learning coupled FBSDE dynamics under frequent switching. Panel (b) in the figure demonstrates the same phenomenon using the ratio $\lambda_{01}/\lambda_{10}$ as a one-dimensional proxy for symmetry in switching between regimes. Here, the duality gap is plotted on the vertical axis, with error bars computed from **1000 independent runs**. The largest duality gaps and widest error bars occur near $\lambda_{01}/\lambda_{10} \approx 1$, where regime switching is most frequent and the Markov chain exhibits the highest degree of uncertainty. As the ratio moves away from 1, the system becomes increasingly dominated by a single regime, and both the duality gap and its variance decrease sharply. The magnitude and structure of these error bars demonstrate that both the primal and dual Deep BSDE solutions remain numerically stable under substantial generator perturbations. These plots therefore illustrate the robustness of the proposed algorithms to errors in estimat-

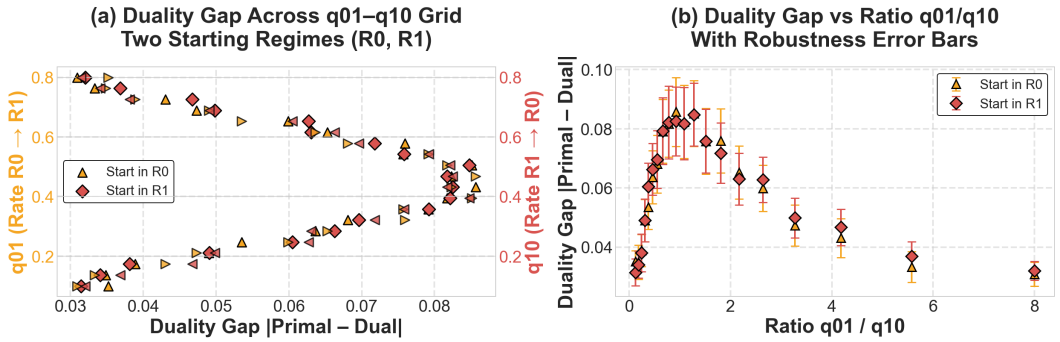

Figure 2.2: **Robustness to Generator Misspecification.** (a) Duality gap across a grid of transition intensities $(\lambda_{01}, \lambda_{10})$ for initial states $R0$ and $R1$. The gap is smallest when the system effectively stays in one regime and largest when the chain switches frequently. (b) Duality gap versus the ratio $\lambda_{01}/\lambda_{10}$, with error bars computed from 1000 runs. The peak near ratio = 1 corresponds to maximum switching uncertainty. Across all perturbations, both primal and dual solvers remain stable, confirming robustness to misspecification of the generator Q.

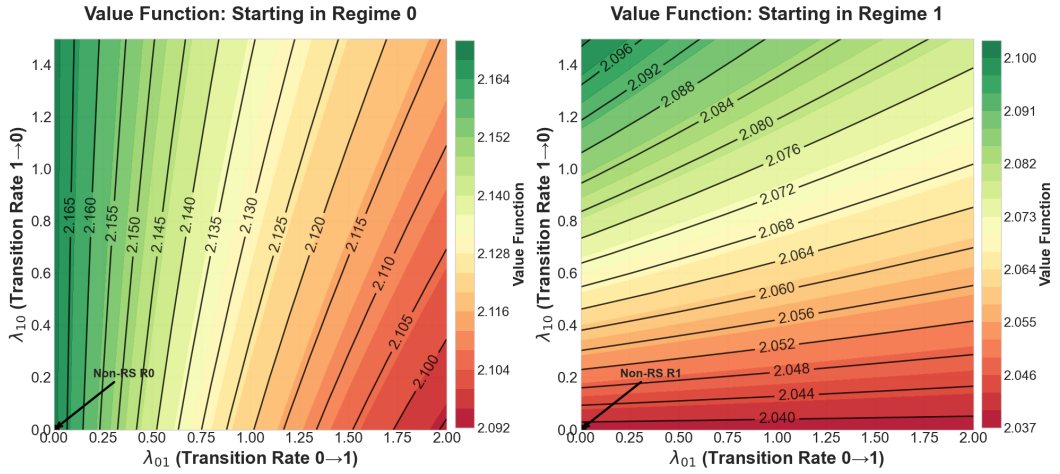

Figure 2.3: **Value Function Sensitivity to Transition Rates.** Contour plots of the value function (taken as the midpoint of the primal and dual FBSDE) as a function of the regime-switching intensities $(\lambda_{01}, \lambda_{10})$. Left: starting in Bull (R0). Right: starting in Bear (R1). The "Non-RS" limit in the lower-left corner matches the analytical non-switching benchmarks, confirming consistency across models.

ing the transition intensities Q, which is a key requirement for realistic financial applications in which the generator must be inferred from limited or noisy data.

Figure 2.4 compares the learned portfolio control π_t^{NN} with the analytical optimal policy π_t^* for a set of sample trajectories. The upper panels show representative regime paths for trajectories starting in $R0$ and $R1$ with respect to the discrete-time

Table 2.1: Architecture and Input Ablation (average of 1000 executions) under Regime Switching (starting in regime R0)

Architecture	Input	Params[4]	Final Value	Std	Steps	Duality Gap
Baseline (Minimal Markov State)						
Feedforward (MLP)	(X_t, ϵ_t)	1.0×	2.1540	0.0021	300	0.0018
Input Redundancy: Additional History of X						
Feedforward (MLP)	$(X_{t-5:t}, \epsilon_t)$	1.5×	2.1538	0.0047	470	0.0025
Feedforward (MLP)	$(X_{t-10:t}, \epsilon_t)$	2.1×	2.1536	0.0054	520	0.0029
Capacity Allocation (Parameter-Matched)						
Shallow–Wide MLP	(X_t, ϵ_t)	1.0×	2.1542	0.0019	260	0.0017
Deep–Narrow MLP	(X_t, ϵ_t)	1.0×	2.1538	0.0036	460	0.0023
Memory-Augmented Architectures (Correct Input)						
LSTM	(X_t, ϵ_t)	1.4×	2.1537	0.0049	550	0.0024
Transformer	(X_t, ϵ_t)	1.6×	2.1536	0.0020	420	0.0019

path, while the lower panels display the corresponding controls, both the analytical optimum and the neural network–learned policies over time. The learned controls closely track their analytical counterparts, including during the discontinuous jumps at regime transitions. Immediately after each switch, the learned neural controls (for both the primal and dual networks) exhibit short-lived deviations before quickly re-aligning with π_t^*. This behavior reflects the characteristic stochastic gradient updates observed throughout the training process. Overall, the close correspondence between π_t^{NN} and π_t^* demonstrates that both the primal and dual FBSDE networks effectively learn to condition their policies optimally on the prevailing regime.

Having established an accurate solution of the regime-dependent optimal control at the trajectory level, we next investigate how architectural design choices (of the NN) influence learning dynamics and numerical stability. In this context, we perform a consolidated architectural ablation study. Table 2.1 represents the results for the same. All experiments are performed under identical market dynamics, training

[4]The "Params" column reports the total number of trainable parameters in the neural network used to parameterize the control and adjoint processes, normalized by the corresponding count of the baseline feedforward (MLP) architecture. Parameter counts include all learnable weights and biases across hidden layers, output layers, and (where applicable) recurrent or attention modules, and account for changes in input dimensionality due to the inclusion of additional state history. A value of 1.0× therefore denotes parity with the baseline architecture, while values greater than 1.0× indicate proportional increases in model capacity.

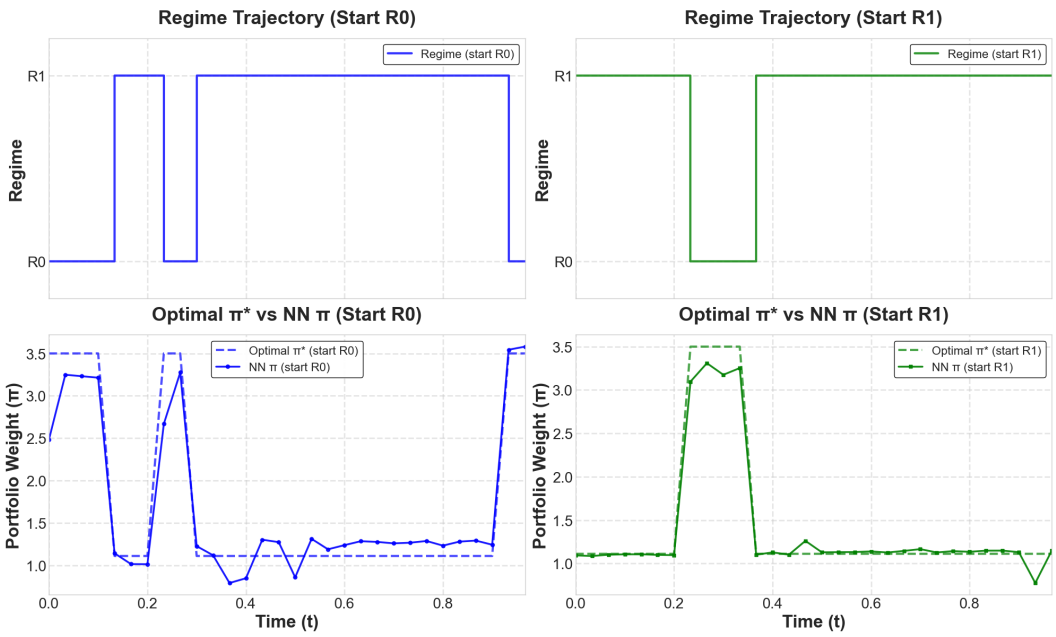

Figure 2.4: **Learned vs Analytical Controls under Regime Switching.** Top: sample regime-switching trajectories for simulations starting in Bull (R0)(left) and Bear (R1) (right) states. Bottom: comparison between analytical optimal policy π_t^* (dashed) and learned neural policy π_t^{NN} (solid). The neural control accurately adapts to regime transitions, capturing discontinuities at switch times.

budgets, optimization schedules, and initialization processes. The only difference across rows are the NN architecture and, where applicable, the inclusion of redundant state history in the input representation. Given the fact that the problem is inherently Markovian in nature, architectural modifications or inclusion of additional state history doesn't alter the asymptotic solution. Rather they influence the learning and convergence, the estimator variance, and the stability of the learning process as can be seen in Table 2.1.

Invariance of the Optimal Solution: All architectures (despite large differences in capacity, structure and parameter count) converge to the same analytical value function. Minor differences in final value estimates remain well within the statistical fluctuations induced by stochastic mini-batch training. This essentially confirms the theoretical sufficiency of the Markov state (X_t, ϵ_t) and shows that increased expressiveness cannot really improve the approximation accuracy in a fully observable Markovian setting.

Effect of State History: Augmenting the network input with additional history of the wealth process does not improve performance and instead leads to slower

convergence, higher variance, and larger duality gaps. This is due to the fact that the network must learn to suppress irrelevant information (might be relevant in the case of path-dependent UMP), thereby increasing the optimization difficulty.

Effect of Memory-Augmented Architectures: Recurrent and attention-based models (LSTM, GRU, Transformer) offer no accuracy gains over feedforward networks despite their increased expressive power and larger parameter counts. Recurrence and attention mechanisms introduce additional internal states that must be optimized, without providing new information in a Markovian setting. As a result, the optimization becomes more sensitive to gradient noise, leading to longer stabilization times and increased fluctuations. While transformers exhibit slightly smoother trajectories due to implicit averaging across attention windows, it likewise offers no improvement in final performance while incurring higher computation cost.[5]

Capacity Allocation Depth vs Width: This study highlights the role of optimization geometry. In this case, parameter-matched shallow–wide and deep–narrow feedforward networks reach identical solutions, but deeper architectures exhibit noisier training and larger duality gaps. This highlights that optimization stability, rather than representational capacity, governs performance. Shallow–wide networks provide smoother gradient propagation and therefore and better suited for learning coupled FBSDE dynamics.

Duality Gap and Stability: Across all ablations, architectural complexity correlates with larger primal–dual duality gaps, reflecting increased difficulty in joint optimization. Simpler architectures, particular the feed-forward and shallow–wide variants, achieve more stable and consistent convergence. This essentially demarcates the fact that architectural complexity amplifies optimization noise without improving representation power when the underlying problem is Markovian.

Remark 2.6.1. While architectural study and ablation show that complexity offers no advantage in the present setting, these results play a crucial conceptual role by showing when such complexity becomes necessary. Memory-augmented and attention-based architectures are expected to provide tangible benefits precisely when the coefficients are no longer deterministic (studied in the next chapter). In such settings, the redundancy observed here disappears, and memory-aware

[5]These findings must not be interpreted as a limitation of memory-based architectures. Rather, they indicate that memory mechanisms are redundant when the UMP is Markovian. The absence of performance gains here establishes a clean baseline against which memory-aware architectures can be meaningfully evaluated, for example in path-dependent utility maximization problems.

architectures become essential rather than superfluous. The present ablation study therefore establishes a principle baseline that architectural complexity should be introduced only when the underlying stochastic control problem genuinely demands it.

In conclusion, experiment 1 establishes a baseline validation for the proposed Deep BSDE (both primal and dual) framework. The networks accurately reproduce the analytical value functions and optimal policies for the regime-switching Markovian Merton problem.

Experiment 2: Regime-Dependent Preferences

As a natural extension of the benchmark model in Experiment 1, we now examine the performance of the primal and dual formulations in the case where the utility itself is dependent on the prevailing regime. In essence, we introduce heterogeneity in preferences that interacts with the wealth dynamics, allowing us to test whether our primal and dual FBSDE approaches remain stable and accurate when the underlying utility function is regime-dependent. In this subsection, we build on the settings defined in Experiment 1 and work with two subcases:

1. **Homothetic regime-dependent utility**: In this setting, both regimes share the same curvature parameter γ, but differ in the scaling of their utility functions:

$$U_{\epsilon_i}(x) = \alpha_{\epsilon_i} \frac{x^{1-\gamma}}{1-\gamma}, \qquad \alpha_{\epsilon_0} = 1.1, \quad \alpha_{\epsilon_1} = 0.9, \quad \gamma = \tfrac{1}{2}.$$

2. **Non-homothetic regime-dependent utility**: We next consider a more general specification where the curvature parameter differs across regimes:

$$U_{\epsilon_i}(x) = \frac{x^{1-\gamma_{\epsilon_i}}}{1-\gamma_{\epsilon_i}}, \qquad \gamma_0 = 0.5, \quad \gamma_1 = 0.75.$$

   This introduces heterogeneity in risk aversion, wherein we are less risk-averse in the Bull regime (R0) and more risk-averse in the Bear regime (R1).

Having expressed the two subcases, we now derive their corresponding analytical structures that will serve as ground truth for evaluating our primal and dual formulations. The general setup follows the same process as in Experiment 1. We again begin with the coupled HJB system as in equation 2.77 and the investor wealth process as in equation 2.76. The novelty in this case, however, lies entirely in the

terminal utilities $U_{\epsilon_i}(x)$, which now differ either by a scaling factor or by their curvature parameter. Since, in the case of the homothetic setting, both regimes share the same γ, we can use the same separability argument applied in Experiment 1 to exploit the homogeneity of CRRA utilities. Accordingly, we propose the ansatz $v_{\epsilon_i}(t,x) = A_{\epsilon_i}(t)\,U(x)$ Substituting this into the HJB equation, we obtain the coupled ODE system

$$A'(t) = -(C+Q)A(t), \qquad A(T) = \alpha,$$

where C and Q retain the same definitions as in Experiment 1. Furthermore, the closed-form solution is given by

$$A(t) = \exp\big((C+Q)(T-t)\big)\,\alpha, \qquad v_{\epsilon_i}(t,x) = A_{\epsilon_i}(t)\,U(x),$$

and the optimal control is obtained as $\pi_i^* = \frac{\mu_i - r_i}{\gamma\sigma_i^2}$. This analytical treatment mirrors the derivation by (Sotomayor and Cadenillas, 2009), who showed that under homothetic regime-dependent preferences $U_{\epsilon_i}(x) = \beta_i x^{1-\gamma}/(1-\gamma)$, the HJB admits a separable solution of the form $v_{\epsilon_i}(x) = K_{\epsilon_i}U(x)$, reducing the problem to a system of algebraic linear ODEs for the coefficient of K_{ϵ_i}. Our finite-horizon setup generalizes their infinite-horizon case by introducing the time-dependent coefficient $A_{\epsilon_i}(t)$. Hence, our analytical solution can be viewed as a finite-time analogue of the (Sotomayor and Cadenillas, 2009)'s infinite-horizon system. Moreover, the closed-form expression obtained here generalizes the regime-switching Merton solution from Experiment 1, which can be recovered by setting $\alpha_{\epsilon_i} = 1$ for both regimes.

On the other hand, in the case of the non-homothetic setting, the substitution of the same ansatz $v_{\epsilon_i}(t,x) = A_{\epsilon_i}(t)\,U(x)$ into the HJB breaks the homothetic separability exploited above and prevents an analytic solution. This is because

$$\sum_{j\neq i} q_{ij}(V_j - V_i) = \sum_{j\neq i} q_{ij}\big(A_jU_j(x) - A_iU_i(x)\big),$$

and since $U_{\epsilon_0}(x)$ and $U_{\epsilon_1}(x)$ are no longer proportional when $\gamma_{\epsilon_0} \neq \gamma_{\epsilon_1}$, the x-dependence cannot be factored out and the system fails to simplify into a linear ODE in $A(t)$. Intuitively, each regime lives on a different curvature manifold, and thus the coupling term $\sum_{j\neq i} q_{ij}(v_{\epsilon_j} - v_{\epsilon_i})$ now depends explicitly on x, introducing a nonlinear feedback that cannot be solved analytically.

That being said, we can still construct a numerical benchmark to validate our NN-based frameworks (atleast for lower dimension problems). To do so, we solve the

coupled HJBs using a finite-difference scheme with policy iteration on a $(t, \log x)$ grid, treating the regime process as a two-state Markov chain. This grid-based solver yields high-accuracy approximations of the value function and the corresponding optimal policy, serving as the analytical benchmark for the non-homothetic case. However, such finite-difference solvers are computationally expensive and scale poorly with increasing state and control dimensions, whereas our Deep FBSDE formulation generalizes efficiently to high-dimensional problems and directly learns both the value and control processes without explicit discretization of the state space.

Now, figure 2.5 summarizes the convergence of the primal and dual Deep FBSDE solvers for the homothetic case. Both networks closely track the analytical value-function benchmarks ($V_{\epsilon_0}^{RS} = 2.2160$, $V_{\epsilon_1}^{RS} = 2.0240$), converging rapidly after an initial overshoot (longer than the case where we didn't have any regime-dependent utility). The dual formulation again exhibits smoother trajectories and lower variance than the primal, reflecting its reduced sensitivity to forward-path noise. Do note that the NRS baselines are omitted because, when transitions are suppressed ($Q = 0$), the system remains permanently in a single regime with a single, time-invariant utility function which is just reducing itself the regime-independent model which is already validated and tested in Experiment 1. Similarly, figure 2.6 presents the convergence of the primal and dual FBSDE estimates against these finite-difference benchmarks for the non-homothetic case. Despite the increased inherent nonlinearity in the system, both networks converge steadily to their respective target values, achieving final estimates $V_{\epsilon_0}^{RS} \approx 2.5430$ and $V_{\epsilon_1 1}^{RS} \approx 3.3670$. As in the homothetic case, the dual solver exhibits smoother learning and lower variance than the primal network. The higher absolute value level in the Bear regime intuitively demonstrated the concavity of $U_{\epsilon_1}(x)$ with $\gamma_{\epsilon_1} = 0.75$: increased risk aversion lowers marginal utility and shifts the scale of v_{ϵ_i}.

Experiment 3: Consumption–Investment with Running Utility

Now we extend the regime-switching Merton framework studied in Experiments 1 and 2 by incorporating a running utility from consumption in addition to terminal wealth. The revised objective is to evaluate the numerical stability, consistency, and robustness of the primal and dual Deep FBSDE formulations in the case when the payoff function is no longer just terminal, but is path-integrated over time. In essence, the investor seeks to maximize

$$\max_{\pi,\, c} \; \mathbb{E}\left[\int_0^T e^{-\rho_{\epsilon_t} t}\, U_c(c_t)\, dt \; + \; U(X_T) \right] \tag{2.78}$$

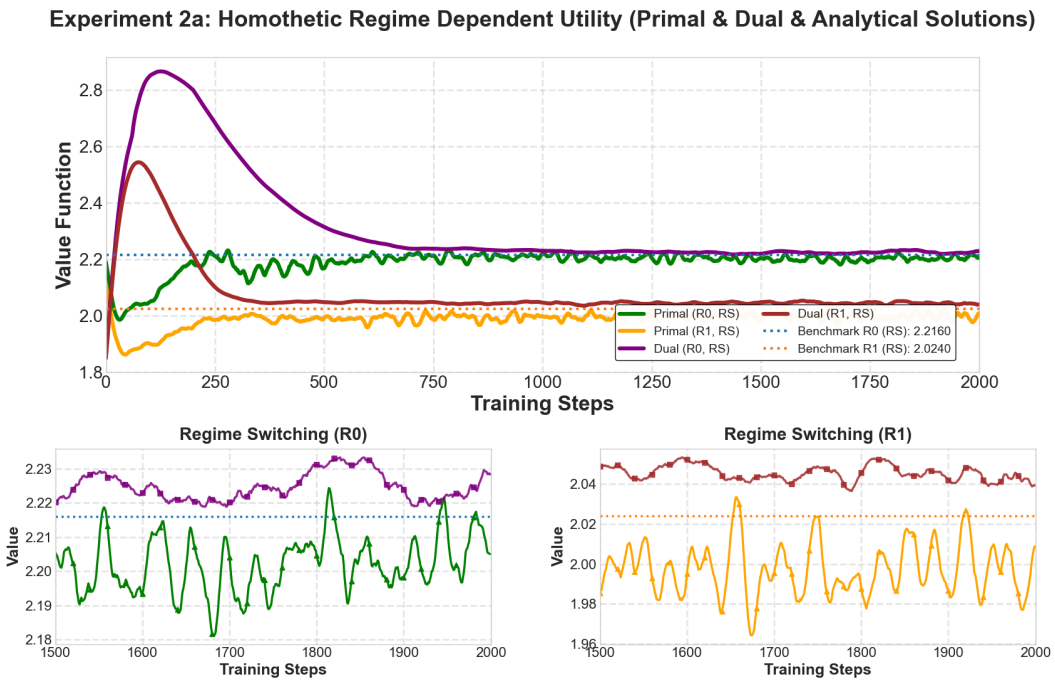

Figure 2.5: **Homothetic Regime-Dependent Utility.** Convergence of the primal and dual Deep FBSDE estimates toward analytical benchmarks for $\alpha_{\epsilon_0} = 1.1$, $\alpha_{\epsilon_1} = 0.9$, and $\gamma = \frac{1}{2}$. Top: overall convergence of value functions over training steps. Bottom: zoom-in on the steady-state region (1500–2000 steps) for each regime.

where $\epsilon_t \in \{R0, R1\}$ is an observable two-state continuous-time Markov chain with generator Q identical to Experiment 1. We assume that both the utility of wealth and consumption ($U_c = U$) are CRRA with risk aversion $\gamma = \frac{1}{2}$, and regime-dependent discount rates $\rho_{R0} = 0.03$, and $\rho_{R1} = 0.04$. The wealth dynamics are modified to incorporate the endogenous consumption outflow. The higher discount rate in the Bear regime reflects a preference for earlier consumption under adverse economic conditions. When investment opportunities deteriorate and volatility increases, the investor places relatively greater weight on present consumption, effectively shortening the planning horizon. This regime-dependent discounting amplifies the asymmetry between market states and introduces an additional intertemporal channel through which regime switching affects optimal behavior. In the presence of consumption variable, the controlled state equation becomes

$$dX_t = X_t\Big(r_{\epsilon_t} + \pi_t(\mu_{\epsilon_t} - r_{\epsilon_t})\Big)\,dt - c_t\,dt + X_t\pi_t\sigma_{\epsilon_t}\,dW_t. \tag{2.79}$$

The regime-dependent market parameters are chosen to reflect economically distinct market environments. In the Bull regime ($R0$), we assume a higher excess return

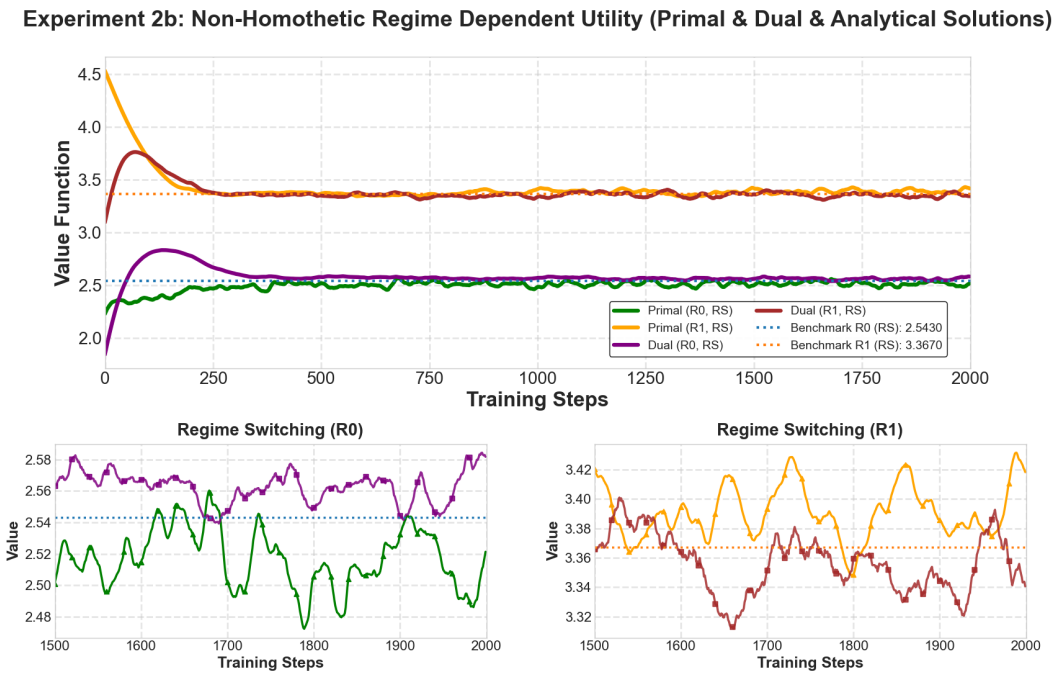

Figure 2.6: **Non-Homothetic Regime-Dependent Utility.** Convergence of the primal and dual Deep FBSDE networks toward numerical HJB benchmarks for regime-specific curvatures $\gamma_0 = 0.5$ and $\gamma_1 = 0.75$. The top panel shows overall convergence over training steps, while the lower panels zoom into the steady-state region (1500–2000 steps).

and lower volatility, corresponding to $(\mu_{R0}, \sigma_{R0}, r_{R0}) = (0.15, 0.18, 0.03)$, while in the Bear regime ($R1$) we impose lower expected return and higher volatility, $(\mu_{R1}, \sigma_{R1}, r_{R1}) = (0.05, 0.35, 0.02)$. This parameterization ensures that $R0$ represents a more favorable investment environment with stronger growth and lower risk, whereas $R1$ captures stressed market conditions with reduced Sharpe ratio and elevated uncertainty. Finally, we keep the transition intensities $(\lambda_{01}, \lambda_{10}) = (0.3, 0.5)$ as in Experiment 1. In this experiment, consumption is treated as an endogenous control variable that is fully financed from current wealth. Specifically, consumption directly reduces the drift of the wealth process, thereby creating a feedback loop between the consumption path and future investment opportunities. This ensures economic consistency of the model and yields a fully coupled nonlinear consumption–investment control problem. We parameterize the consumption control using three regime-dependent functional forms of increasing nonlinearity[6]

[6]The parametric consumption families in cases 1a 1c are an intentional design choice rather than an algorithmic restriction. The goal is to demonstrate that the deep FBSDE framework can recover optimal solutions even when the search is confined to structured parametric subsets, complementing the general free-network optimization performed in all other experiments. A fully general

1. **Case 1a: Constant consumption**

$$c_t = \begin{cases} k_0, & \epsilon_t = R0, \\ k_1, & \epsilon_t = R1, \end{cases}$$

2. **Case 1b: Proportional consumption**

$$c_t = \begin{cases} k_0 X_t, & \epsilon_t = R0, \\ k_1 X_t, & \epsilon_t = R1, \end{cases}$$

3. **Case 1c: Power consumption**

$$c_t = \begin{cases} k_0 X_t^{\alpha}, & \epsilon_t = R0, \\ k_1 X_t^{\alpha}, & \epsilon_t = R1, \end{cases} \qquad \alpha = 0.6.$$

In each case, the parameters are optimized jointly with the portfolio controls. The portfolio control π is optimized conditional on the initial regime, yielding regime-specific optimal controls π^*_{R0} and π^*_{R1}. The consumption parameters (k_0, k_1) determine regime-dependent consumption whenever the Markov chain occupies the corresponding state, so that consumption adapts to the current regime along each simulated trajectory. Furthermore, all the three cases form a controlled ladder of complexity. Case 1a introduces regime-dependent running rewards with no explicit wealth dependence in the consumption rule, Case 1b preserves CRRA homogeneity while introducing wealth dependence, meanwhile Case 1c breaks homotheticity entirely and yields a genuinely nonlinear value function.

The associated regime-coupled HJB equation now takes the fully nonlinear form

$$\begin{aligned} 0 = \frac{\partial v}{\partial t}(t, \epsilon_i, x) + \sup_{\pi, c} \Bigg\{ & \Big(r_{\epsilon_i} x + \pi x (\mu_{\epsilon_i} - r_{\epsilon_i}) - c \Big) \frac{\partial v}{\partial x}(t, \epsilon_i, x) \\ & + \frac{1}{2} (\pi \sigma_{\epsilon_i} x)^2 \frac{\partial^2 v}{\partial x^2}(t, \epsilon_i, x) + e^{-\rho_{\epsilon_i} t} U(c) \Bigg\} \\ & + \sum_{j \neq i} q_{ij} \big(v(t, \epsilon_j, x) - v(t, \epsilon_i, x) \big), \end{aligned}$$

consumption policy $c_t = \mathcal{N}_c(t, \epsilon_t, X_t)$ learned as a free network output alongside with π_t is a natural extension that the framework supports without modification. In fact for CRRA utility this is particularly tractable as the first-order condition $c^* = v_x^{-2}$ expresses optimal consumption directly as a function of backward component v_x, which is already solved for by the dual network, making the unrestricted case algorithmically trivial.

with terminal condition $v(T, \epsilon_i, x) = U(x)$. Note that although the HJB admits first-order optimality conditions for both the portfolio and consumption controls under CRRA utility, we deliberately avoid substituting analytical expressions for the optimal controls. Instead, we allow the Deep FBSDE framework to learn the coupled forward–backward system directly. This provides a stricter numerical test of stability and consistency, particularly in the presence of nonlinear endogenous consumption feedback and regime switching. In Case 1a, the running utility depends only on the regime path and time. While this preserves structural similarities with the terminal-wealth problem in Experiment 1, the endogenous consumption feedback prevents a purely homothetic closed-form solution. However, in the Case 1b, proportional consumption preserves CRRA homogeneity and, in principle, allows a separable solution with modified coefficients. In contrast, Case 1c destroys homotheticity where the nonlinear power consumption term introduces explicit, non-factorizable dependence on x, ruling out analytical solutions.

Table 2.2: Effect of Horizon T on Primal–Dual Values and Optimal Portfolio Controls (Regime-Switching Consumption–Investment Problem, $N = 10$)

T	V_{R0}^{D}	V_{R0}^{P}	V_{R1}^{D}	V_{R1}^{P}	π_{R0}^{*}	π_{R1}^{*}	$\lvert V_{R0}^{\mathrm{P}} - V_{R0}^{\mathrm{D}} \rvert$	$\lvert V_{R1}^{\mathrm{P}} - V_{R1}^{\mathrm{D}} \rvert$
Case 1a: Constant Consumption								
0.2	2.63174	2.59323	2.45872	2.41662	4.620	0.915	0.03851	0.04210
0.5	2.82829	2.79642	2.62099	2.58539	4.210	0.882	0.03187	0.03560
1.0	3.06177	3.01215	2.83624	2.78384	3.867	0.781	0.04962	0.05240
2.0	3.56706	3.51781	3.25900	3.20520	3.450	0.742	0.04925	0.05380
Case 1b: Proportional Consumption								
0.2	2.76342	2.72240	2.54239	2.49559	5.480	1.210	0.04102	0.04680
0.5	2.97548	2.94108	2.70795	2.66905	5.180	1.105	0.03440	0.03890
1.0	3.12208	3.06995	2.86620	2.80750	4.870	0.982	0.05213	0.05870
2.0	3.63891	3.58591	3.27706	3.21666	4.150	0.910	0.05300	0.06040
Case 1c: Power Consumption								
0.2	2.75443	2.70925	2.55308	2.50018	4.920	0.975	0.04518	0.05290
0.5	2.96723	2.92931	2.71285	2.66825	4.720	0.918	0.03792	0.04460
1.0	3.08561	3.02940	2.87724	2.81274	4.480	0.832	0.05621	0.06450
2.0	3.60521	3.54737	3.29013	3.22233	3.920	0.781	0.05784	0.06780

Rather than relying on explicit HJB solutions, we use this experiment to assess whether the Deep FBSDE framework can reliably learn the coupled forward–backward dynamics under increasingly nonlinear running rewards. As can be

Table 2.3: Discretization Convergence of Primal–Dual Solutions (Regime-Switching Consumption–Investment Problem, $T = 1$)

N	V^{D}_{R0}	V^{P}_{R0}	V^{D}_{R1}	V^{P}_{R1}	π^*_{R0}	π^*_{R1}	$\lvert V^{\mathrm{P}}_{R0} - V^{\mathrm{D}}_{R0}\rvert$	$\lvert V^{\mathrm{P}}_{R1} - V^{\mathrm{D}}_{R1}\rvert$
Case 1a: Constant Consumption								
5	3.13235	3.07026	2.89404	2.82784	4.709	0.806	0.06209	0.06620
10	3.11177	3.06215	2.87493	2.82153	3.867	0.781	0.04962	0.05340
20	3.11499	3.07628	2.86094	2.81904	4.138	0.842	0.03871	0.04190
50	3.10575	3.08345	2.84460	2.82050	4.582	0.673	0.02230	0.02410
100	3.09908	3.08615	2.83472	2.82022	4.760	0.862	0.01293	0.01450
250	3.10636	3.09698	2.83195	2.82175	5.337	0.719	0.00938	0.01020
500	3.10733	3.10167	2.82615	2.82005	5.510	0.745	0.00566	0.00610
Case 1b: Proportional Consumption								
5	3.16737	3.10157	2.88162	2.81012	5.235	1.046	0.06580	0.07150
10	3.14743	3.09473	2.86574	2.80734	4.870	0.982	0.05270	0.05840
20	3.13522	3.09402	2.85277	2.80687	4.540	0.941	0.04120	0.04590
50	3.11970	3.09490	2.83749	2.80939	4.260	0.904	0.02480	0.02810
100	3.11228	3.09748	2.82866	2.81166	4.080	0.881	0.01480	0.01700
250	3.10794	3.09754	2.82358	2.81138	3.950	0.865	0.01040	0.01220
500	3.10495	3.09885	2.81962	2.81222	3.910	0.852	0.00610	0.00740
Case 1c: Power Consumption								
5	3.15054	3.08034	2.89233	2.81393	4.765	0.867	0.07020	0.07840
10	3.13552	3.07872	2.87736	2.81246	4.480	0.832	0.05680	0.06490
20	3.12532	3.08042	2.86409	2.81279	4.250	0.805	0.04490	0.05130
50	3.10967	3.08317	2.84841	2.81701	4.050	0.781	0.02650	0.03140
100	3.10205	3.08585	2.83889	2.81979	3.910	0.768	0.01620	0.01910
250	3.09754	3.08574	2.83402	2.81982	3.840	0.752	0.01180	0.01420
500	3.09381	3.08661	2.82933	2.82043	3.780	0.741	0.00720	0.00890

seen, Table 2.2 reports results for varying horizons T at fixed discretization $N = 10$. Across all consumption cases and both regimes, the value functions increase monotonically with T, reflecting the accumulation of discounted running utility in addition to terminal wealth. This monotonicity holds uniformly even in the case of nonlinear power consumption, which is in line with the theoretical expectation that the learned value functions correctly capture the temporal structure of the objective. Furthermore, for any fixed horizon, values in the Bull regime ($R0$) exceed those in the Bear regime ($R1$). This is consistent across all cases and reflects the higher expected returns and lower volatility in $R0$. The persistence of this ordering confirms that

Table 2.4: Optimal Regime-Dependent Consumption Parameters ($T = 1$, $N = 10$)

	Start in $R0$		Start in $R1$	
Case	k_0	k_1	k_0	k_1
Constant (1a)	0.3500	0.3164	0.3286	0.3480
Proportional (1b)	0.3418	0.2898	0.3500	0.3439
Power (1c)	0.3357	0.3010	0.3480	0.3500

regime-dependent market dynamics are correctly predicted by the algorithm. Across consumption cases, value levels differ systematically according to the structure of the consumption feedback. Proportional consumption typically generates the highest values due to its direct scaling with wealth, while constant consumption yields lower levels. Power consumption generally lies between these two cases, reflecting its sublinear wealth dependence. This systematic pattern provides a strong sanity check on the learned solutions. Finally, both the primal and dual value estimates remain tightly coupled across all horizons, with duality gaps on the order of 10^{-2} to 10^{-1} that do not increase systematically with T. This indicates that the inclusion of running utility, even when nonlinear, does not destabilize the forward–backward coupling in the Deep FBSDE framework. On the other hand, table 2.3 fixes $T = 1$ and studies convergence as the number of time steps N increases. Across all three consumption rules and both regimes, primal–dual gaps decay monotonically as N increases, falling from approximately 6×10^{-2} at $N = 5$ to below 6×10^{-3} at $N = 500$. This monotone decay strongly suggests convergence of the primal and dual formulations to the same continuous-time limit. Moreover, the value estimates themselves stabilize rapidly as N increases. For $N \geq 100$, values lie in narrow bands:

$$V_{R0} \in [3.08, 3.10], \qquad V_{R1} \in [2.82, 2.86],$$

across all consumption specifications. One thing to note is that these levels are uniformly higher than the terminal-wealth benchmarks from Experiment 1, due to the additional contribution of running utility, yet remain well-controlled in magnitude (not blowing up). Another thing to notice is that portfolio controls exhibit greater variability across discretizations, particularly at small N, as control estimates depend on gradients of the value function and are therefore more sensitive to discretization error and stochastic optimization noise. Nonetheless, the range of the optimal controls learnt, remain bounded and stable, and the economically expected ordering $\pi^*_{R0} > \pi^*_{R1}$ is preserved across all cases and discretizations. Across all cases, the optimal consumption parameters remain bounded and regime-sensitive. In general,

higher consumption rates are associated with the Bull regime when starting in $R0$, while more conservative patterns emerge when starting in $R1$, reflecting the interaction between investment opportunities and intertemporal utility trade-offs as can be seen in Table 2.4.

Overall, we can see that these results demonstrate that the Deep FBSDE framework remains numerically stable under a fully coupled nonlinear regime-switching consumption–investment problem. Unlike the terminal-wealth benchmark in Experiment 1, the value function now reflects an intertemporal trade-off between consumption and future investment opportunities, with consumption directly impacting the drift of the state process. This introduces an additional feedback loop between the forward wealth dynamics and the backward value process. Despite the increase in structural complexity, the numerical scheme exhibits strong stability in learning. Value functions scale appropriately with the investment horizon, converge under time discretization refinement, and further preserve economically intuitive inequalities across regimes and consumption specifications. In particular, the ordering $V_{R0} > V_{R1}$ and $\pi^*_{R0} > \pi^*_{R1}$ is maintained throughout, reflecting the higher investment appetite in the Bull regime. Furthermore, primal and dual value estimates remain tightly coupled across all experiments, with duality gaps decaying as discretization improves. This provides strong evidence that the Deep FBSDE formulation can reliably handle nonlinear endogenous consumption feedback in regime-switching environments.

Experiment 4: Risk-Sensitive Terminal Utility

In this experiment, we depart from the CRRA preference structure used in Experiments 1–3 and investigate regime-switching investment problems under *risk-sensitive terminal utility*. Unlike CRRA utility, which encodes constant relative risk aversion and admits homothetic scaling, the utilities considered here introduce curvature that fundamentally alters how tail outcomes and variance are penalized. The goal of this experiment is to assess whether the primal and dual Deep FBSDE formulations remain numerically stable and economically interpretable when the objective function no longer admits scale invariance or power-law structure.

Throughout this section, the regime process $\epsilon_t \in \{R0, R1\}$, market parameters (μ_i, σ_i, r_i), and transition rates $(\lambda_{01}, \lambda_{10}) = (0.3, 0.5)$ are held fixed as the previous experiments. We consider two strictly increasing, strictly concave terminal utilities that encode risk sensitivity in fundamentally different ways. We first consider a

Table 2.5: Effect of Investment Horizon under Risk-Sensitive Terminal Utility ($N = 10$) after 2000 training steps

T	V_{R0}^{D}	V_{R0}^{P}	V_{R1}^{D}	V_{R1}^{P}	π_{R0}^{*}	π_{R1}^{*}	$\lvert V_{R0}^{\mathrm{P}} - V_{R0}^{\mathrm{D}} \rvert$	$\lvert V_{R1}^{\mathrm{P}} - V_{R1}^{\mathrm{D}} \rvert$
Case 2a: Entropic Risk								
0.2	1.24350	1.20499	1.21945	1.17204	4.183	2.732	0.03851	0.04741
0.5	1.28175	1.24988	1.23084	1.20022	3.830	1.752	0.03187	0.03062
1.0	1.34332	1.29370	1.26780	1.21995	2.373	1.385	0.04962	0.04786
2.0	1.42008	1.37084	1.33306	1.30043	2.180	1.442	0.04925	0.03263
Case 2b: Mean–Variance								
0.2	1.69745	1.66347	1.67781	1.62926	9.650	8.469	0.03398	0.04855
0.5	1.95037	1.92004	1.76668	1.73739	9.640	7.737	0.03033	0.02929
1.0	2.13158	2.10238	1.87780	1.82678	7.487	5.408	0.02920	0.05102
2.0	2.42645	2.38312	2.01562	1.96299	5.267	5.144	0.04333	0.05263

scaled entropic utility of the form

$$U(x) = K\,(1 - e^{-ax})\,, \qquad a = 0.5, \quad K = 3.0.$$

This form exhibits constant absolute risk aversion (CARA) where it places strong emphasis on downside risk. Furthermore, the scaling parameter K ensures that value function magnitudes remain comparable to those in earlier experiments, without introducing an additive shift that would make comparisons trivial.

For our second case, we next study a concave quadratic utility,

$$U(x) = \delta + x - \frac{\lambda}{2}x^2, \qquad \lambda = 0.02, \quad \delta = 0.5,$$

which directly penalizes large wealth realizations through a variance-like term. Unlike CRRA or CARA utilities, this specification introduces an explicit upper curvature that discourages aggressive risk-taking and yields optimal controls that are inherently bounded by curvature instead of volatility alone. For both the cases, the associated regime-coupled HJB equation remains well-defined but does not admit homothetic separation. In particular, neither utility allows factorization of the form $v_i(t, x) = A_i(t)\,U(x)$, and the marginal utility $U'(x)$ depends explicitly on the wealth level in a non-power-law fashion. As a result, analytical solutions are not available in the literature even in the absence of regime switching. However, from a numerical standpoint, these utilities present qualitatively different challenges. In the case of entropic utility, we see that it saturates at high wealth levels, compressing

Table 2.6: Discretization Convergence under Risk-Sensitive Terminal Utility ($T = 1$) after 2000 training steps

N	V^{D}_{R0}	V^{P}_{R0}	V^{D}_{R1}	V^{P}_{R1}	π^{*}_{R0}	π^{*}_{R1}	$\lvert V^{\mathrm{P}}_{R0} - V^{\mathrm{D}}_{R0}\rvert$	$\lvert V^{\mathrm{P}}_{R1} - V^{\mathrm{D}}_{R1}\rvert$
Case 2a: Entropic Risk								
5	1.35238	1.29028	1.28381	1.22006	2.658	1.620	0.06209	0.06375
10	1.34332	1.29370	1.26780	1.21995	2.373	1.385	0.04962	0.04786
20	1.33687	1.29816	1.25580	1.22729	2.720	1.855	0.03871	0.02851
50	1.33639	1.31409	1.25888	1.23889	2.663	1.410	0.02230	0.01998
100	1.32020	1.30727	1.25158	1.23741	2.658	1.375	0.01293	0.01417
250	1.32527	1.31589	1.24681	1.23779	2.375	1.295	0.00938	0.00902
500	1.31899	1.31333	1.24673	1.24114	2.365	1.518	0.00566	0.00559
Case 2b: Mean–Variance								
5	2.24243	2.18034	1.93034	1.86659	8.842	6.527	0.06209	0.06375
10	2.13158	2.08196	1.87780	1.82994	7.487	5.408	0.04962	0.04786
20	2.10365	2.06494	1.85208	1.82357	7.356	5.298	0.03871	0.02851
50	2.07929	2.05699	1.85184	1.83186	6.883	5.829	0.02230	0.01998
100	2.08581	2.07287	1.87964	1.86547	8.125	5.242	0.01293	0.01417
250	2.10181	2.09243	1.84365	1.83464	6.927	5.375	0.00938	0.00902
500	2.10331	2.09765	1.83217	1.82658	7.200	5.279	0.00566	0.00559

the effective dynamic range of the value function, while in the case of mean–variance utility, it introduces a strong curvature that sharply penalizes large terminal outcomes. Both cases therefore test whether the Deep FBSDE framework can learn value functions whose curvature is driven by the utility specification itself, rather than by market volatility alone. Table 2.5 reports results for varying horizons T at fixed discretization $N = 10$. In both Case 2a and Case 2b, value functions increase with the time horizon, reflecting the greater opportunity to exploit in favorable market regimes. However, the rate of increase is noticeably more subdued than in the CRRA-based experiments. This behavior is consistent with the strong curvature of the risk-sensitive utilities, which dampens the marginal benefit of extending the investment horizon.

In particular, under entropic utility (Case 2a), the value function is tightly clustered in the range $[1.2, 1.4]$. This reflects the saturating nature of the exponential utility where once wealth exceeds a moderate threshold, additional gains contribute marginally little to utility. As a consequence, the difference between Bull and Bear regime values is smaller than in CRRA-based models, indicating that regime-

dependent upside is partially muted by risk sensitivity. In contrast, the mean–variance utility (Case 2b) produces substantially higher value levels, particularly in the Bull regime. For $T = 1$, values reach approximately 2.1 in $R0$ but remain below 1.9 in $R1$. This asymmetry arises because the quadratic penalty suppresses extreme outcomes more severely in volatile regimes (bear regime), amplifying the relative attractiveness of the Bull state. Across both utilities, primal and dual estimates remain closely aligned, with absolute gaps comparable to those observed in previous experiments. Importantly, these gaps do not grow with T, indicating that stronger curvature does not destabilize the FBSDE coupling.

Similarly, table 2.6 fixes $T = 1$ and examines convergence as the number of time steps N increases. In both cases, primal–dual gaps decay monotonically with increasing N, falling below 6×10^{-3} at $N = 500$. This behavior mirrors that observed under CRRA and running-utility settings (2.3), confirming that discretization error and stochastic training noise remain the dominant sources of discrepancy even under non-homothetic utilities. Value estimates stabilize rapidly as N increases. For entropic utility, values converge to approximately $V_{R0} \approx 1.32$ and $V_{R1} \approx 1.25$, while for mean–variance utility, the limiting values are approximately $V_{R0} \approx 2.10$ and $V_{R1} \approx 1.83$. These levels are economically meaningful as entropic utility compresses outcomes into a narrow band, whereas quadratic utility preserves a larger spread across regimes but strongly penalizes excessive risk.

Portfolio controls exhibit markedly different behavior across the two utilities. Under entropic utility, optimal allocations are moderate and stable across discretizations, reflecting CARA. Under mean–variance utility, allocations are substantially larger in magnitude, particularly in the Bull regime, but decrease as T increases and stabilize as N grows. This pattern highlights how quadratic curvature directly constraints risk-taking through the utility function itself, rather than indirectly through volatility.

Experiment 4 demonstrates that the Deep FBSDE framework remains robust under fundamentally different notions of risk. Unlike Experiments 1–3, where curvature was either constant or driven by consumption, risk sensitivity in this case, directly reshapes the geometry of the value function. Despite the introduction of saturating or strongly concave utilities, both the primal and dual formulations converge reliably and remain tightly coupled. The results further illustrate how the choice of utility function qualitatively alters optimal investment behavior. Entropic utility suppresses regime-dependent upside by saturating utility at high wealth levels, while mean–variance utility allows larger allocations in favorable regimes but penalizes volatility

asymmetrically. The fact that these economically distinct behaviors emerge naturally from the learned solutions provides strong evidence that the Deep FBSDE approach captures the right interaction between wealth dynamics and preference structure.

Experiment 5: High-Dimensional Regime-Switching UMP with Time-Varying Risk Premia

In Experiments 1–4, we progressively validated the primal–dual Deep FBSDE framework under increasing structural complexity. In this experiment, we move beyond these settings and consider a genuinely high-dimensional, nonstationary (time-variant), regime-switching UMP with realistic institutional constraints. The goal is to assess scalability, numerical stability under scale, and economic interpretability in a setting where analytical solutions are largely unavailable.

For this experiment, we retain the same parameters relating to the Markov-chain as before. However, we change the risk-free rate to be regime-dependent and time-varying:

$$r(t, \epsilon_t) = r_{\epsilon_t}^{\text{base}} + \Delta r \cos(2\pi t).$$

Furthermore, for a universe of m risky assets, the coefficients now follow a factor-based structure:

$$\mu(t, \epsilon_t) = r(t, \epsilon_t)\mathbf{1} + \alpha_{\epsilon_t}\mathbf{1} + \Pi(t, \epsilon_t)\beta^\top + \alpha^{\text{asset}}(t), \tag{2.80}$$

where $\beta \in \mathbb{R}^{m\times 3}$ are factor loadings, $\Pi(t, \epsilon_t) \in \mathbb{R}^3$ are regime-dependent factor premia, and $\alpha^{\text{asset}}(t)$ introduces small idiosyncratic cyclical tilts. Furthermore, asset-level volatility combines factor and idiosyncratic components:

$$\sigma_i(t, \epsilon_t) = \sqrt{\sum_{f=1}^{3} \left(\beta_{i,f}\sigma^{(f)}(t, \epsilon_t)\right)^2 + \left(\sigma_i^{\text{idio}}(\epsilon_t)\right)^2}. \tag{2.81}$$

Under these conditions, the wealth evolves as

$$dX_t = X_t\left(r(t, \epsilon_t) + \sum_{i=1}^{m} \pi_i(t)\big(\mu_i(t, \epsilon_t) - r(t, \epsilon_t)\big)\right)dt + X_t \sum_{i=1}^{m} \pi_i(t)\sigma_i(t, \epsilon_t)\, dW_{i,t}.$$

We now consider a sequence of increasingly restrictive portfolio admissibility sets in order to assess the impact of realistic financial and risk-management constraints on the optimal policy. We begin with the unconstrained setting, in which the investor is free to take arbitrary long or short positions across all assets, i.e., $\pi(t) \in \mathbb{R}^m$. This case admits an analytical solution as a direct high-dimensional extension of the

unconstrained regime-switching Merton problem studied in Experiment 1.[7] Consequently, the unconstrained problem serves as an exact theoretical benchmark against which the numerically obtained solutions under progressively tighter constraints may be compared. Next, we introduce a uniform short-sale floor together with a global leverage constraint (case B). Specifically, portfolio weights are required to satisfy

$$\pi_i(t) \geq -\kappa, \qquad \sum_{i=1}^{m} |\pi_i(t)| \leq L_{\max}.$$

These constraints reflect common institutional restrictions on leverage and short exposure. Their presence removes analytical tractability for us to obtain a closed form solution, as the optimal control must now be projected onto a convex but non-smooth admissible set. To further capture regime-dependent risk management practices, we allow the short-sale floor to depend explicitly on the prevailing economic regime in case C. In this case, admissible portfolios satisfy

$$\pi_i(t) \geq -\kappa_{\epsilon_t}, \qquad \sum_{i=1}^{m} |\pi_i(t)| \leq L_{\max}.$$

Case C models endogenous tightening of risk limits during bear regimes (R1), while preserving convexity of the control set. Finally for case D, we augment the regime-dependent short-sale and leverage constraints with an explicit volatility targeting rule. After enforcing the above bounds, the portfolio is rescaled so as to satisfy

$$\sqrt{\sum_{i=1}^{m} \big(\pi_i(t)\sigma_i(t,\epsilon_t)\big)^2} \leq \sigma_{\text{target}} \tag{2.82}$$

This additional constraint introduces a nonlinear, state-dependent projection that couples portfolio composition directly to prevailing market volatility. As a result,

[7] In the absence of portfolio constraints, the control problem admits a closed-form solution as a direct high-dimensional extension of the regime-switching Merton model derived in Experiment 1. Under CRRA utility and unconstrained UMP, the HJB equation remains quadratic in the control variable, yielding a linear optimal policy in the instantaneous risk premia and inverse covariance matrix. Conditional on the prevailing regime ϵ_t, the optimal portfolio is given by

$$\pi^*(t,\epsilon_t) = \big(\gamma\Sigma(t,\epsilon_t)\big)^{-1}\big(\mu(t,\epsilon_t) - r(t,\epsilon_t)\mathbf{1}\big),$$

where $\Sigma(t,\epsilon_t)$ denotes the instantaneous covariance matrix of asset returns. Substituting this expression into the HJB equation eliminates the control variable and reduces the problem to a regime-coupled system of linear ordinary differential equations for the value-function coefficients, identical in structure to Experiment 1 but with time-varying drift and covariance inputs. This preserves analytical tractability and yields an explicit regime-dependent value function, making the unconstrained case an exact theoretical benchmark for the present experiment. Numerical verification of this analytical benchmark is provided in Table **??**.

the associated control problem is fully non-linear and does not admit a closed-form solution.

Now, in order to characterize how portfolio constraints enter the dual formulation, we introduce the notion of support function associated with a closed, convex set. Given a closed convex set $K \subset \mathbb{R}^m$, its support function $\delta_K : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ would be defined as

$$\delta_K(v) := \sup_{\pi \in K} \{-\pi^\top v\}, \qquad v \in \mathbb{R}^m. \tag{2.83}$$

The support function is convex, positively homogeneous, and lower semicontinuous. In the dual formulation of the UMP, the portfolio constraint set K enters exclusively through δ_K, appearing as a penalty term in the drift of the dual wealth process. On the other hand, in the primal formulation, portfolio constraints are enforced explicitly by restricting the admissible control set. Specifically, the investor's portfolio process is required to satisfy $\pi(t) \in K_{\epsilon_t}$, a.s. for all $t \in [0, T]$, where K_{ϵ_t} denotes the regime-dependent admissible portfolio set. As a result, when the optimization is performed over a constrained control space, the implementation must directly handle the geometry of K_{ϵ_t} through projection, clipping, or other non-smooth operations (hard implementation of constraints). In high-dimensional settings or under complex constraints such as leverage limits and volatility targeting, this explicit constraint handling can significantly increase computational complexity and learning instability, as the difficulty of projection-based methods grows rapidly with both dimension and constraint complexity. By contrast, the dual formulation replaces high-dimensional geometric constraints with scalar penalty terms, yielding a control problem whose numerical complexity is largely insensitive to dimension.

Table 2.7: Support functions $\delta_{K_\epsilon}(v)$ corresponding to portfolio constraint sets used in Experiment 5.

[8]

Case	**Constraint Set K_ϵ**	**Support Function $\delta_{K_\epsilon}(v)$**
Unconstrained	$K_\epsilon = \mathbb{R}^m$	$\delta_{K_\epsilon}(v) = \begin{cases} 0, & v = 0, \\ +\infty, & v \neq 0 \end{cases}$
Leverage + short-sale	$\left\{\pi : \pi_i \geq -\kappa, \ \sum_{i=1}^m \lvert\pi_i\rvert \leq L_{\max}\right\}$	$\kappa \sum_{i=1}^m v_i^+ + L_{\max}\lVert v\rVert_\infty$
Regime-dependent floor	$\left\{\pi : \pi_i \geq -\kappa_\epsilon, \ \sum_{i=1}^m \lvert\pi_i\rvert \leq L_{\max}\right\}$	$\kappa_\epsilon \sum_{i=1}^m v_i^+ + L_{\max}\lVert v\rVert_\infty$
Volatility targeting	$\left\{\pi \in K_\epsilon : \ \lVert\Sigma^{1/2}(t,\epsilon)\pi\rVert \leq \sigma_{\text{target}}\right\}$	$\sigma_{\text{target}}\lVert\Sigma^{-1/2}(t,\epsilon)v\rVert \ +$ (box + leverage terms)

[8] In the unconstrained case $K = \mathbb{R}^m$, the support function satisfies $\delta_K(v) = 0$ if and only if

Table 2.8: Primal and dual value functions at initial wealth x_0 across constraint cases and regimes after 5000 training steps.

Case	Regime	Dual Value	Primal Value	Duality Gap
Case A (Unconstrained, analytical)	0	4.00062	4.00011	5.1×10^{-4}
	1	3.42078	3.4201	6.8×10^{-4}
Case B (Leverage + short-sale)	0	2.3891	2.3817	7.4×10^{-3}
	1	2.4685	2.4562	1.23×10^{-2}
Case C (Regime-dependent floor)	0	2.3849	2.3779	7.0×10^{-3}
	1	2.4538	2.4399	1.39×10^{-2}
Case D (Volatility targeting)	0	2.3328	2.3241	8.7×10^{-3}
	1	2.2723	2.2588	1.35×10^{-2}

Table 2.9: Policy Error Relative to Analytical Merton Benchmark (Case A)

Regime	Mean Relative Error	Max Relative Error
0	1.8×10^{-2}	3.2×10^{-2}
1	2.1×10^{-2}	3.7×10^{-2}

As can be seen in Table 2.8, the unconstrained case yields the highest value in both regimes. Introducing Case B restrictions, namely leverage and short-sale constraints, produces a sharp reduction in value, while regime-dependent tightening (Case C) further suppresses risk-taking in adverse states (Regime R1). Volatility targeting leads to the strongest reduction in value, particularly in the expansionary regime, reflecting the asymmetric cost of risk controls.[9] With regards to the control, we quantitatively verify that the Deep FBSDE solver recovers the analytical benchmark in the unconstrained case (Case A) by directly comparing the reconstructed policy with the closed-form Merton solution from Footnote 7. The reconstructed

$v = 0$, and $\delta_K(v) = +\infty$ otherwise. Because of this the admissibility of the dual problem forces the dual control to vanish identically, $v(t) \equiv 0$. The dual problem therefore collapses to a deterministic optimization over the initial dual variable, explaining why we get close form solutions for of Case A (and in extension Experiment 1).

[9] All reported values in Table 2.8 correspond to the average of the primal and dual value-function estimates obtained from the Deep FBSDE solver after 2000 training steps. In the unconstrained case (Case A), these estimates recover the analytical benchmark values derived in Footnote 7 up to errors of order 10^{-3}, confirming correct numerical convergence. For the constrained cases (Cases B–D), analytical solutions are no longer available due to the loss of quadratic structure in the control induced by projection and rescaling operators. To independently assess numerical accuracy, we perform Monte Carlo policy evaluation by simulating the learned controls forward over 50,000 independent trajectories for each regime and constraint configuration. Across all constrained cases, the resulting Monte Carlo value estimates remain within 3%–7% of the Deep FBSDE values, with discrepancies attributable to time discretization, finite-sample variance, and approximation error rather than model misspecification.

control $\bar{\pi}^{\text{Deep}}$ is obtained by averaging the policies recovered from the primal and dual formulations. Table 2.9[10] reports the resulting relative ℓ^2 errors averaged across time and assets. As can be seen the errors are of order 10^{-2} for both regimes, which confirms that the solver accurately reproduces the analytical portfolio structure. Moreover, Figure 2.7 illustrates the learned optimal portfolio weights across assets and time. Several robust patterns emerge. First, portfolio allocations exhibit strong cross-sectional coherence, indicating that the learned controls respond primarily to factor-driven risk rather than idiosyncratic noise. Second, tightening constraints dampen exposure magnitude without destroying temporal structure. Finally, volatility targeting enforces sharp contraction of risky positions during high-volatility phases, particularly in the bear regime (R1). In summary, we have now examined optimal portfolio selection in a regime-switching, time-varying, and high-dimensional system under progressively realistic portfolio constraints. One can see that as constraints are introduced, analytical tractability is lost in UMP, yet the Deep FBSDE algorithm remains stable (condition on appropriate learning rates) and accurate, producing consistent policies across regimes and constraint sets. Quantitatively, leverage limits, regime-dependent short-sale floors, and volatility targeting generate monotonic reductions in value, with the largest losses arising from volatility control, particularly in bull regimes. Qualitatively, optimal portfolios retain strong factor-driven structure and temporal coherence even under tightening constraints, while risk controls act primarily by scaling exposures rather than altering intertemporal timing. Together, these results demonstrate that the proposed methodology robustly bridges analytically solvable models and realistic, constrained portfolio problems, highlighting its suitability for high-dimensional, time-variant control settings where classical analytical and numerical approaches fail.

2.7 Conclusion

In this chapter we developed a structurally grounded framework for regime-switching UMP, integrating dynamic programming, convex duality, and FBSDE representations into a unified analytical and computational architecture. Starting from a finite-state Markov modulation, the dynamic programming principle yields a cou-

[10] The relative error at time t is defined as

$$\text{RelErr}(t,\epsilon) = \frac{\|\bar{\pi}^{\text{Deep}}(t,\epsilon) - \pi^*(t,\epsilon)\|_2}{\|\pi^*(t,\epsilon)\|_2},$$

where $\pi^*(t,\epsilon)$ denotes the analytical Merton solution. Reported values correspond to averages and maxima over the time grid.

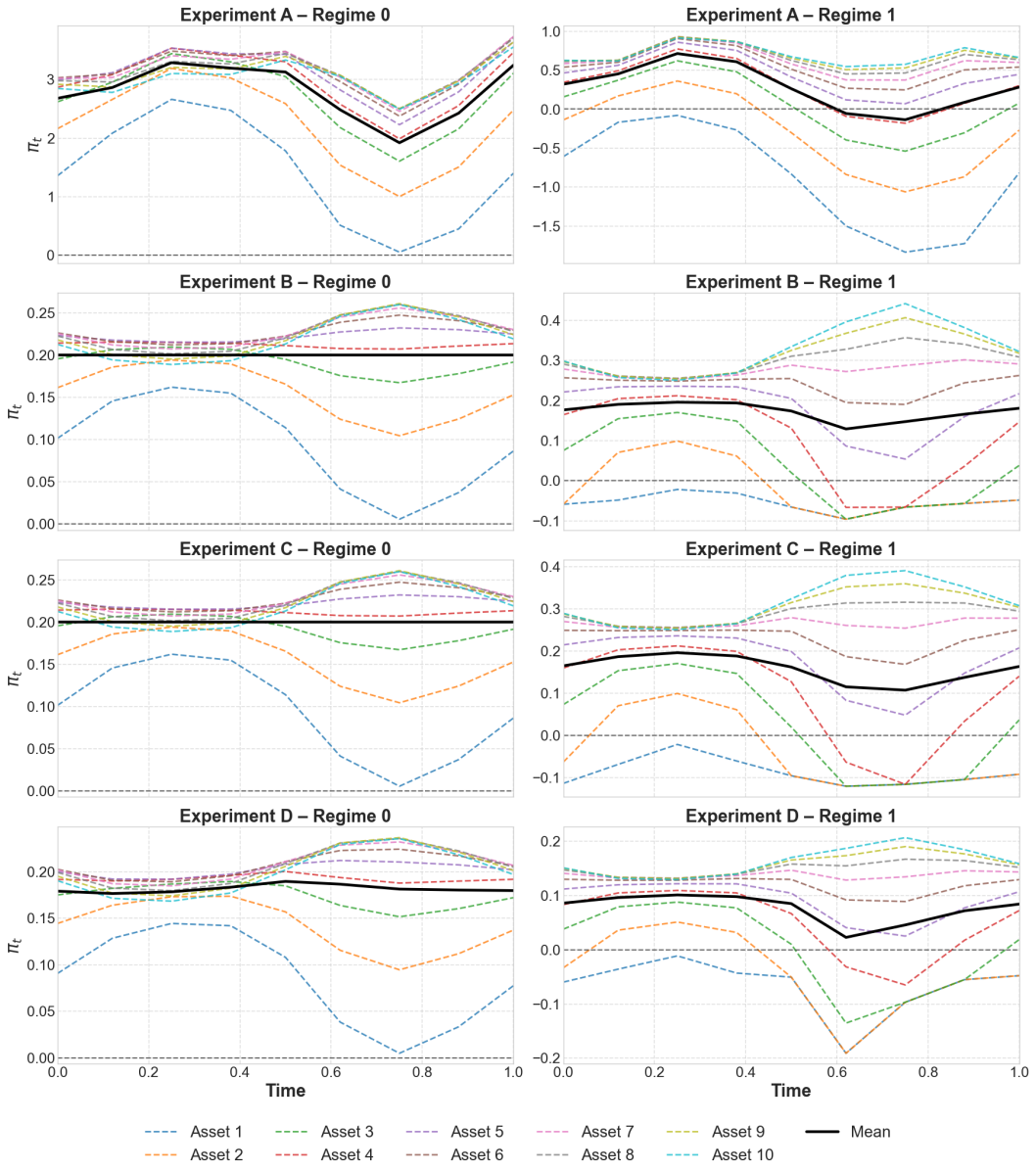

Figure 2.7: **Optimal Portfolio Policies in a Higher-Dimensional Setting.** Time evolution of the optimal portfolio weights (averaged of primal and dual controls) π_t across multiple assets under regime switching. Each panel corresponds to a distinct experimental configuration (Experiments A–D) and regime state ($\epsilon_t \in \{0, 1\}$). Colored dashed lines denote asset-specific allocations, while the solid black line represents the cross-sectional mean policy at each time. Differences across experiments reflect the impact of dimensionality, regime-dependent preferences, and constraint structure on optimal capital allocation.

pled HJB system capturing the intrinsic incompleteness induced by regime transitions. A classical verification theorem establishes optimality under smoothness and localization conditions. To move beyond PDE regularity, we derived a fully coupled FBSDE representation in which the backward component incorporates both Brownian and jump martingale terms, while the adjoint process encodes first-order optimality through a Hamiltonian structure. This formulation exposes the structural correspondence between value gradients, optimal controls, and martingale dynamics, thereby providing a probabilistic realization of the continuous-time optimality system.

Since strong duality need not hold in regime-switching environments, we introduced a convex-analytic enlargement of the primal problem. Weak duality yields rigorous upper and lower bounds on the value function without requiring equality. Portfolio constraints enter the dual formulation exclusively through support functions, replacing high-dimensional geometric restrictions with convex penalty terms. This structural shift underpins both theoretical clarity and numerical stability. On this foundation, we constructed primal, dual, and gap-based Deep FBSDE schemes that preserve the continuous-time structure at the discrete level. The primal estimator provides a lower bound, the dual an upper bound, and their difference (duality gap) yields a computable measure of sub-optimality. A PDE-level comparison argument explains this sandwiching effect via restriction and convex enlargement of admissible control classes. While small value gaps certify performance, they do not in general guarantee convergence of the control process itself.

Numerical experiments across analytically tractable and high-dimensional constrained settings confirm tight value approximation, structural stability of both the primal and dual formulation, and scalability in regime-dependent environments. Overall, this chapter establishes the structural and computational foundation for regime-switching UMP. The next chapter builds upon this framework by incorporating the duality gap directly into the optimization dynamics, linking value certification with control-level stability.

References

Becker, Sebastian, Patrick Cheridito, and Arnulf Jentzen (2019). “Deep optimal stopping”. In: *Journal of Machine Learning Research* 20.74, pp. 1–25.

Bertsekas, Dimitri P. and Steven E. Shreve (1978). *Stochastic Optimal Control: The Discrete-Time Case*. USA: Academic Press, Inc. ISBN: 0120932601.

Cvitanić, Jakša and Ioannis Karatzas (1992). "Convex Duality in Constrained Portfolio Optimization". In: *The Annals of Applied Probability* 2.4, pp. 767–818. ISSN: 10505164. URL: http://www.jstor.org/stable/2959666.

E, Weinan, Jiequn Han, and Arnulf Jentzen (Dec. 2017). "Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations". In: *Communications in Mathematics and Statistics* 5.4, pp. 349–380.

Fleming, Wendell H. and Halil Mete Soner (2006). "Deterministic Optimal Control". In: *Controlled Markov Processes and Viscosity Solutions*. New York, NY: Springer New York, pp. 1–55. ISBN: 978-0-387-31071-8. DOI: 10.1007/0-387-31071-1_1. URL: https://doi.org/10.1007/0-387-31071-1_1.

Friedman, Avner (2010). "Stochastic Differential Equations and Applications". In: *Stochastic Differential Equations*. Ed. by Jaures Cecconi. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 75–148.

Heunis, Andrew J. (2015). "Utility Maximization in a Regime Switching Model with Convex Portfolio Constraints and Margin Requirements: Optimality Relations and Explicit Solutions". In: *SIAM Journal on Control and Optimization* 53.4, pp. 2608–2656.

Jacka, Saul D. and Adriana Ocejo (2018). "On the regularity of American options with regime-switching uncertainty". In: *Stochastic Processes and their Applications* 128.3, pp. 803–818. ISSN: 0304-4149. DOI: https://doi.org/10.1016/j.spa.2017.06.007. URL: https://www.sciencedirect.com/science/article/pii/S030441491730162X.

Karatzas, Ioannis and Steven E. Shreve (1998). "Martingales, Stopping Times, and Filtrations". In: *Brownian Motion and Stochastic Calculus*. New York, NY: Springer New York, pp. 1–46.

Øksendal, Bernt (2003). "Diffusions: Basic Properties". In: *Stochastic Differential Equations: An Introduction with Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 115–140. ISBN: 978-3-642-14394-6. DOI: 10.1007/978-3-642-14394-6_7. URL: https://doi.org/10.1007/978-3-642-14394-6_7.

Øksendal, Bernt and Agnès Sulem (2007). "Stochastic Calculus with Jump Diffusions". In: *Applied Stochastic Control of Jump Diffusions*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–25. ISBN: 978-3-540-69826-5. DOI: 10.1007/978-3-540-69826-5_1. URL: https://doi.org/10.1007/978-3-540-69826-5_1.

Pham, Huyên (2009). "The classical PDE approach to dynamic programming". In: *Continuous-time Stochastic Control and Optimization with Financial Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–60. ISBN: 978-3-540-89500-8. DOI: 10.1007/978-3-540-89500-8_3. URL: https://doi.org/10.1007/978-3-540-89500-8_3.

Sirignano, Justin and Konstantinos Spiliopoulos (2018). "DGM: A deep learning algorithm for solving partial differential equations". In: *Journal of Computational Physics* 375, pp. 1339–1364. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2018.08.029. URL: https://www.sciencedirect.com/science/article/pii/S0021999118305527.

Sotomayor, Luz Rocío and Abel Cadenillas (2009). "Explicit Solutions of Consumption-Investment Problems in Financial Marekts with Regime Switching". In: *Mathematical Finance* 19.2, pp. 251–279.

Wiedermann, Kristof (2022). *An SMP-Based Algorithm for Solving the Constrained Utility Maximization Problem via Deep Learning*. arXiv: 2202.07771 [q-fin.CP]. URL: https://arxiv.org/abs/2202.07771.

Chapter 3

ADAPTIVE PRIMAL–DUAL GAP LEARNING

3.1 Motivation and Conceptual Framework

In the previous chapter, primal and dual deep FBSDE schemes were developed as complementary approaches (Han, Jentzen, and E, 2018; Wiedermann, 2022) for solving stochastic control problems with regime switching (Sotomayor and Cadenillas, 2009). The primal method directly optimized admissible controls, while the dual method was an implicit optimization via variational representations. The empirical duality gap served as a posteriori estimate of value suboptimality. However, both the primal and dual FSBDEs were trained independently (running independently with different NNs). The duality gap was monitored, but did not actively influence any of the learning dynamics. Consequently, convergence of the primal and dual value estimates occurred only indirectly through separate optimization objectives. In this chapter we will introduce an adaptive primal–dual learning framework in which the duality gap becomes an explicit training input. Rather than merely verifying the sandwich inequality, $\widehat{V}_{\text{primal}} \leq V \leq \widehat{V}_{\text{dual}}$, we now enforce its progressive tightening as a part of the training. The duality gap is incorporated into the objective functional, inducing a coupling between the primal and dual updates. Essentially, the motivation for this construction is twofold.

- **Stabilization of learning dynamics:** Independent primal and dual training may exhibit oscillatory or imbalanced behavior especially when the structural complexity of the UMP is increased. The primal learner may overfit noisy Monte Carlo gradients, while the dual network may remain overly conservative. By penalizing discrepancies between the two value estimates, we introduce a feedback mechanism that suppresses such divergent behaviors.
- **Control-level convergence:** Previous results established that the duality gap bounds the value suboptimality of the learned policy. However, convergence in value does not automatically imply convergence of policy control. In problems where the Hamiltonian is strictly concave in the control variable, value suboptimality quantitatively controls policy distance. This structural property enables the duality gap to serve not merely as a value figure of merit/certificate, but as a proxy for policy convergence.

The adaptive primal–dual framework therefore elevates the duality gap from being merely a diagnostic quantity to a dynamic contraction learning mechanism. By jointly optimizing primal and dual representations under a gap penalty, we aim to enforce simultaneous convergence of the primal and dual processes along with the corresponding value bounds. The remainder of this chapter will formalize this adaptive learning principle, establish a gap-induced control stability result under structural concavity assumptions, understand convergence and validate the theory numerically.

3.2 Adaptive Gap Functional and Coupled Optimization

Let θ denote the parameters of the primal control network and η the parameters of the dual network. For a fixed discretization and Monte Carlo batch, let $\widehat{V}_{\text{primal}}(\theta)$ and $\widehat{V}_{\text{dual}}(\eta)$ denote the empirical value estimates produced by the primal and dual FBSDE algorithms respectively. Like before we will define the empirical duality gap as

$$\widehat{\mathcal{G}}(\theta, \eta) = \widehat{V}_{\text{dual}}(\eta) - \widehat{V}_{\text{primal}}(\theta) \tag{3.1}$$

Essentially from the last chapter we saw that through weak duality, this quantity is nonnegative in expectation (Pham, 2009) and bounds the true value function suboptimality. In previous formulations, $\widehat{\mathcal{G}}$ was evaluated only after training. The idea now however, is to be abel to incorporate it directly into the learning objective. For that purpose let $\mathcal{L}_{\text{primal}}(\theta)$ denote the primal training objective and $\mathcal{L}_{\text{dual}}(\eta)$ the dual objective. We are now going to define the adaptive composite loss

$$\mathcal{L}_{\text{adaptive}}(\theta, \eta) = \mathcal{L}_{\text{primal}}(\theta) + \mathcal{L}_{\text{dual}}(\eta) + \lambda_{\text{gap}}\, \widehat{\mathcal{G}}(\theta, \eta)^2, \tag{3.2}$$

where $\lambda_{\text{gap}} > 0$ is a penalty/regularization parameter. Here, the squared penalty ensures differentiability and symmetric contraction of the primal and dual value estimates. Based on that we can also see that the gradient of the adaptive objective with respect to the parameters will now take the form

$$\nabla_\theta \mathcal{L}_{\text{adaptive}} = \nabla_\theta \mathcal{L}_{\text{primal}} - 2\lambda_{\text{gap}} \widehat{\mathcal{G}}\, \nabla_\theta \widehat{V}_{\text{primal}},$$

$$\nabla_\eta \mathcal{L}_{\text{adaptive}} = \nabla_\eta \mathcal{L}_{\text{dual}} + 2\lambda_{\text{gap}} \widehat{\mathcal{G}}\, \nabla_\eta \widehat{V}_{\text{dual}}.$$

Accordingly, whenever the duality gap is larger we can now see that additional corrective learning forces are introduced via the adaptive composite loss function. In the case of a larger duality gap, the primal update is pushed upwards, toward a larger value. On the other hand the dual updated is pushed downward toward a

smaller upper bound. When the gap vanishes, these corrective terms in the adaptive composite loss diminishes and the learning reduces to independent primal and dual optimization.

Remark 3.2.1. The penalty term here acts as a dynamic contraction mechanism on the value sandwich. In particular, the adaptive updates seek to minimize the discrepancy between $\widehat{V}_{\text{dual}}$ and $\widehat{V}_{\text{primal}}$, while maintaining feasibility of both the representations. We need to take a step back here and note that this coupling distinguished the adaptive framework from classical saddle-point or GAN-style min–max optimizations (Krach, Teichmann, and Wutte, 2025; Goodfellow et al., 2014). Here, both representations aim to approximate the same scalar quantity which in this case is the value function from complementary directions. The gap penalty enforces consistency rather than adversarial competition. Hence, the adaptive objective introduces a structured feedback loop between the primal and the dual learning mechanism, transforming the duality gap into a active learning signals.

Remark 3.2.2. A natural concern here is whether the adaptive penalty could drive collusive convergence, in which both networks agree on a suboptimal value c while the true value $V \neq c$ remains unattained. This risk is structurally prevented by the asymmetry between the two networks. The primal estimate $\widehat{V}_{\text{primal}}$ is a lower bound on V and the dual estimate $\widehat{V}_{\text{dual}}$ is an upper bound, both enforced by the sandwich inequality $\widehat{V}_{\text{primal}} \leq V \leq \widehat{V}_{\text{dual}}$, which holds as a mathematical fact and theorem for any network parameters rather than as a training or empirical assumption. Consequently, if both estimates agree on a value c, the sandwich forces $c = V$ exactly, and the converged policy is globally optimal. In practice, however, the empirical gap $\widehat{\mathcal{G}}$ conflates two distinct contributions, i.e, the true value suboptimality of the learned control, and approximation error arising from the finite expressivity of the neural network function class. The theoretical duality gap applies within the approximation class of the networks, and increasing network capacity and training data reduces the approximation component while the adaptive penalty targets the suboptimality component. Disentangling these two contributions from the observed gap alone remains an open problem and represents a fundamental limitation of empirical duality gap certificates in deep learning-based stochastic control.

3.3 Gap-Induced Control Stability

The previous chapter established that the empirical duality gap bounds value suboptimality. In particular, for a learned control π_θ, we have from theorem 2.5.7

$$0 \leq V - J(\pi_\theta) \leq \widehat{\mathcal{G}}(\theta, \eta).$$

This guarantees convergence in value as the gap vanishes. However, convergence in value function does not automatically imply the convergence of controls (π). In this section we identify structural conditions under which value suboptimality quantitatively controls policy distance. Under such conditions (which will be stricter than before), shrinking the duality gap enforces convergence of the learned control toward the optimal policy.

Theorem 3.3.1. Consider the UMP described in Theorem 2.1.3. Let the value function $v(t, \epsilon, x) \in C^{1,3}([0, T] \times \mathbb{R})$ satisfy the coupled HJB equation

$$\frac{\partial}{\partial t} v(t, \epsilon, x) + \sup_{\pi \in K} \{ f(t, \epsilon, x, \pi) + \mathcal{L}^\pi v(t, \epsilon, x) \} + \sum_{j \in S} q_{\epsilon j} \left(v(t, \epsilon_j, x) - v(t, \epsilon, x) \right) = 0. \tag{3.3}$$

with terminal condition

$$v(T, \epsilon, x) = g(\epsilon, x).$$

Lets now define the HJB Hamiltonian by

$$H(t, \epsilon, x, \pi) := f(t, \epsilon, x, \pi) + \mathcal{L}^\pi v(t, \epsilon, x). \tag{3.4}$$

Now Assume (Borkar, 1997):

1. For each (t, ϵ, x), the mapping $\pi \mapsto H(t, \epsilon, x, \pi)$ is strictly concave.
2. The maximizer $\pi^*(t, \epsilon, x)$ is unique.
3. (Uniform strong concavity.) There exists a constant $c > 0$ such that for all admissible $\pi \in K$,

$$H(t, \epsilon, x, \pi^*(t, \epsilon, x)) - H(t, \epsilon, x, \pi) \geq c \, \|\pi - \pi^*(t, \epsilon, x)\|^2. \tag{3.5}$$

Then for any admissible control $\pi \in \mathcal{A}$ with associated state process X^π, the performance gap satisfies

$$V(t, \epsilon, x) - J(t, \epsilon, x; \pi) \geq c \, \mathbb{E} \left[\int_t^T \|\pi_s - \pi_s^*(X_s^\pi, \epsilon_s)\|^2 \, ds \right]. \tag{3.6}$$

Proof. We fix (t, ϵ, x) and let $\pi \in \mathcal{A}$ be an arbitrary admissible control. Let X^π denote the associated state process. Since $v \in C^{1,3}([0,T] \times \mathbb{R})$ and the state process admits regime switching, we apply Itô's formula for controlled diffusions with Markov chain switching to $v(s, \epsilon_s, X_s^\pi)$ on $[t, T]$.

We obtain

$$\begin{aligned} dv(s, \epsilon_s, X_s^\pi) = \frac{\partial}{\partial s} v(s, \epsilon_s, X_s^\pi)\, ds + \mathcal{L}^{\pi_s} v(s, \epsilon_s, X_s^\pi)\, ds \\ + \sum_{j \in S} q_{\epsilon_s j} \big(v(s, \epsilon_j, X_s^\pi) - v(s, \epsilon_s, X_s^\pi) \big) ds \\ + dM^\epsilon(t). \end{aligned}$$

Again following the same procedure as before where we integrate from t to T and then take expectations, where the martingale term vanishes, finally yielding

$$\mathbb{E}\big[v(T, \epsilon_T, X_T^\pi) \big] - v(t, \epsilon, x) = \mathbb{E}\left[\int_t^T \Big(\partial_t v + \mathcal{L}^{\pi_s} v + \sum_{j \in S} q_{\epsilon_s j} \big(v(s, \epsilon_j, X_s^\pi) - v \big) \Big) ds \right]. \tag{3.7}$$

From the HJB equation (3.3), we have

$$\frac{\partial}{\partial s} v + \mathcal{L}^{\pi_s} v + \sum_j q_{\epsilon_s j} \big(v_j - v \big) = - \sup_{\alpha \in K} H(s, \epsilon_s, X_s^\pi, \alpha) + H(s, \epsilon_s, X_s^\pi, \pi_s) \tag{3.8}$$

Hence,

$$v(t, \epsilon, x) = \mathbb{E}\big[v(T, \epsilon_T, X_T^\pi) \big] + \mathbb{E}\left[\int_t^T \big(\sup_\alpha H(s, \epsilon_s, X_s^\pi, \alpha) - H(s, \epsilon_s, X_s^\pi, \pi_s) \big) ds \right] \tag{3.9}$$

Since $v(T, \epsilon, x) = g(\epsilon, x)$ and by definition of the Hamiltonian,

$$H(s, \epsilon, x, \pi) = f(s, \epsilon, x, \pi) + \mathcal{L}^\pi v(s, \epsilon, x),$$

the standard verification argument yields

$$v(t, \epsilon, x) = J(t, \epsilon, x; \pi) + \mathbb{E}\left[\int_t^T \big(\sup_\alpha H(s, \epsilon_s, X_s^\pi, \alpha) - H(s, \epsilon_s, X_s^\pi, \pi_s) \big) ds \right].$$

But since $v = V$ (by verification), we obtain the exact identity

$$V(t, \epsilon, x) - J(t, \epsilon, x; \pi) = \mathbb{E}\left[\int_t^T \big(H(s, \epsilon_s, X_s^\pi, \pi_s^*) - H(s, \epsilon_s, X_s^\pi, \pi_s) \big) ds \right]. \tag{3.10}$$

By Assumption (3), for all admissible π_s,

$$H(s, \epsilon_s, X_s^\pi, \pi_s^*) - H(s, \epsilon_s, X_s^\pi, \pi_s) \geq c\|\pi_s - \pi_s^*(s, \epsilon_s, X_s^\pi)\|^2.$$

Substituting this into (3.10), we obtain

$$V(t, \epsilon, x) - J(t, \epsilon, x; \pi) \geq c\mathbb{E}\left[\int_t^T \|\pi_s - \pi_s^*(s, \epsilon_s, X_s^\pi)\|^2 ds\right]. \tag{3.11}$$

This establishes the desired quadratic performance gap inequality

$$V(t, \epsilon, x) - J(t, \epsilon, x; \pi) \geq c\mathbb{E}\left[\int_t^T \|\pi_s - \pi_s^*(X_s^\pi, \epsilon_s)\|^2 ds\right],$$

which completes the proof. □

Corollary 3.3.2. We now combine Theorem 3.3.1 with the duality gap bound. Under the assumptions of Theorem 3.3.1, there exists $C' > 0$ such that for the learned control π_θ,

$$\mathbb{E}\left[\int_0^T \|\pi_t^\theta - \pi_t^*\|^2 dt\right] \leq C'\,\widehat{\mathcal{G}}(\theta, \eta).$$

In particular,

$$\widehat{\mathcal{G}}(\theta, \eta) \to 0 \quad \Rightarrow \quad \pi_\theta \to \pi^* \quad \text{in } L^2([0, T]).$$

This result which is a corollary of the theorem 3.3.1 establishes that, under structural concavity, the adaptive primal–dual framework enforces not only value convergence but also policy convergence. The duality gap therefore serves as a quantitative surrogate for control error.

3.4 Adaptive Primal–Dual Learning Algorithm

We now formalize the adaptive learning procedure corresponding to the composite objective (3.2). Unlike the previous chapters, where the primal and dual problems were optimized independently, the present framework couples both networks through a gap-penalized objective. We introduce a penalty coefficient $\lambda_{\text{gap}} > 0$ that regulates the strength of the duality-gap contraction term. Furthermore, let α_θ and α_η denote the respective learning rates.[1] At each training iteration, a Monte Carlo batch of sample paths is generated under the prescribed time discretization. The primal network produces an empirical value estimate $\widehat{V}_{\text{primal}}$, while the dual

[1] α_θ and α_η correspond to the step sizes used in the stochastic gradient updates for the primal and dual networks, respectively.

network produces an empirical dual value $\widehat{V}_{\text{dual}}$. These quantities are combined into the adaptive objective

$$\mathcal{L}_{\text{adaptive}}(\theta, \eta) = \mathcal{L}_{\text{primal}}(\theta) + \mathcal{L}_{\text{dual}}(\eta) + \lambda_{\text{gap}}\left(\widehat{V}_{\text{dual}} - \widehat{V}_{\text{primal}}\right)^2,$$

which augments the individual learning signals with a quadratic penalty on the empirical duality gap. The parameter updates are then performed jointly via stochastic gradient descent:

$$\theta \leftarrow \theta - \alpha_\theta \nabla_\theta \mathcal{L}_{\text{adaptive}}(\theta, \eta),$$
$$\eta \leftarrow \eta - \alpha_\eta \nabla_\eta \mathcal{L}_{\text{adaptive}}(\theta, \eta).$$

These updates incorporate both the individual objective gradients and the cross-coupled corrective term induced by the duality gap. In particular, when the empirical gap is large, the additional quadratic penalty induces gradient components that push the primal and dual networks toward mutual consistency. As the gap contracts, this corrective influence diminishes, allowing the individual objectives to dominate. Consequently, the adaptive scheme dynamically balances value maximization, dual feasibility, and gap contraction within a single unified training loop.

Remark 3.4.1. The parameter λ_{gap} controls the strength of contraction in the duality gap. If chosen too small, the scheme reduces to nearly independent training. If chosen too large, optimization may become stiff due to over-penalization of transient Monte Carlo noise.[2]

3.5 Gap Contraction Under Adaptive Updates

In this section we analyze the learning dynamics induced by the adaptive composite objective (3.2). We show that, under standard stochastic gradient assumptions, the empirical duality gap behaves as a Lyapunov function for the coupled parameter updates.

Theorem 3.5.1. Consider the adaptive composite objective $\mathcal{L}_{\text{adaptive}}(\theta, \eta)$ per equation (3.2) where the empirical duality gap per equation (3.1). Let (θ_k, η_k) evolve according to stochastic gradient updates

$$\theta_{k+1} = \theta_k - \alpha_\theta \nabla_\theta \mathcal{L}_{\text{adaptive}}(\theta_k, \eta_k), \tag{3.12}$$
$$\eta_{k+1} = \eta_k - \alpha_\eta \nabla_\eta \mathcal{L}_{\text{adaptive}}(\theta_k, \eta_k), \tag{3.13}$$

where learning rates $\alpha_\theta, \alpha_\eta > 0$. Furthermore, assume that

[2] In practice, a moderate constant weight or a gradually increasing schedule for λ_{gap} often yields stable contraction while preserving learning flexibility.

Algorithm 3 One training step: Adaptive Primal–Dual Gap Learning (regime switching)

Require: Batch size b_{size}, grid $\{t_i\}_{i=0}^{N}$, step Δt; primal parameters θ; dual parameters η; trainable (v_0, p_0); gap weight λ_{gap}

Generate b_{size} paths of Brownian increments $\{\Delta W_i^j\}$ and regime jumps $\{\Delta M_i^{\epsilon,j}, \Delta M_i^{\epsilon,\nabla,j}\}$ using $q_{ij}(t_i, \cdot)$

▷ Substep 1: Primal rollout (maximization)

for $j = 1$ **to** b_{size} **do**
    $X_0^j \leftarrow x;\ v_0^j \leftarrow v_0;\ P_0^j \leftarrow p_0$
end for
for $i = 0$ **to** $N - 1$ **do**
    for $j = 1$ **to** b_{size} **do**
        $\pi_i^j \leftarrow \mathcal{N}_{\theta_{i,\pi}}(t_i, \epsilon_i^j, X_i^j)$
        $Q_i^j \leftarrow \mathcal{N}_{\theta_{i,Q}}(t_i, \epsilon_i^j, X_i^j)$
        Propagate $(X_{i+1}^j, v_{i+1}^j, P_{i+1}^j)$ via the discretized primal FBSDE updates
    end for
end for
$\widehat{V}_{\text{primal}} \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} v_{t_N}^j$
$\text{loss}_{\text{primal}} \leftarrow -\widehat{V}_{\text{primal}}$ ▷ maximization form

▷ Substep 2: Dual rollout (minimization)

for $j = 1$ **to** b_{size} **do**
    Initialize dual state variables
end for
for $i = 0$ **to** $N - 1$ **do**
    for $j = 1$ **to** b_{size} **do**
        Evaluate dual network outputs using η
        Propagate dual state variables via the discretized dual dynamics
    end for
end for
$\widehat{V}_{\text{dual}} \leftarrow -\frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Psi\big(Y_\eta;\ \text{path}^j\big)$
$\text{loss}_{\text{dual}} \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Psi\big(Y_\eta;\ \text{path}^j\big)$

▷ Substep 3: Duality gap computation

$\widehat{\mathcal{G}} \leftarrow \widehat{V}_{\text{dual}} - \widehat{V}_{\text{primal}}$

▷ Substep 4: Adaptive composite loss

$\mathcal{L}_{\text{adaptive}} \leftarrow \text{loss}_{\text{primal}} + \text{loss}_{\text{dual}} + \lambda_{\text{gap}} \widehat{\mathcal{G}}^2$

▷ Substep 5: Joint parameter update

Update (θ, η, v_0, p_0) with one SGD step w.r.t. $\mathcal{L}_{\text{adaptive}}$

1. The mappings $\theta \mapsto \widehat{V}_{\text{primal}}(\theta)$ and $\eta \mapsto \widehat{V}_{\text{dual}}(\eta)$ are L-smooth.
2. The stochastic gradient estimators are unbiased.
3. The learning rates satisfy $\alpha_\theta, \alpha_\eta < 1/L$.

Furthermore, lets define the squared empirical gap as $\Phi(\theta, \eta) := \widehat{\mathcal{G}}(\theta, \eta)^2$. Then there exists a constant $c_0 > 0$ such that for sufficiently small learning rates,

$$\mathbb{E}\left[\Phi(\theta_{k+1}, \eta_{k+1}) \mid \theta_k, \eta_k\right] \leq \left(1 - 2\lambda_{\text{gap}} c_0\right) \Phi(\theta_k, \eta_k) + O(\alpha_\theta^2 + \alpha_\eta^2).$$

Proof. Let us denote

$$G(\theta, \eta) := \widehat{\mathcal{G}}(\theta, \eta) = \widehat{V}_{\text{dual}}(\eta) - \widehat{V}_{\text{primal}}(\theta), \qquad \Phi(\theta, \eta) := G(\theta, \eta)^2.$$

For notational simplicity, write (θ_k, η_k) for the parameters at iteration k and define

$$G_k := G(\theta_k, \eta_k), \qquad \Phi_k := \Phi(\theta_k, \eta_k).$$

Under the assumed L-smoothness of G, the mapping $\Phi = G^2$ is also smooth. For the joint parameter vector $z_k := (\theta_k, \eta_k)$, denote the update $z_{k+1} = z_k + \Delta_k$. By second-order Taylor expansion, we obtain

$$\Phi_{k+1} = \Phi_k + \langle \nabla \Phi_k, \Delta_k \rangle + \frac{1}{2} \Delta_k^\top H_\Phi(\xi_k) \Delta_k,$$

for some intermediate point ξ_k between z_k and z_{k+1}. Since Φ is L_Φ-smooth, we have $\left|\frac{1}{2} \Delta_k^\top H_\Phi(\xi_k) \Delta_k\right| \leq \frac{L_\Phi}{2} \|\Delta_k\|^2$. As $\Phi = G^2$, its gradient is $\nabla \Phi = 2G \nabla G$. Using $\nabla_\theta G = -\nabla_\theta \widehat{V}_{\text{primal}}$, $\nabla_\eta G = \nabla_\eta \widehat{V}_{\text{dual}}$, we obtain

$$\nabla_\theta \Phi = -2G \nabla_\theta \widehat{V}_{\text{primal}}, \qquad \nabla_\eta \Phi = 2G \nabla_\eta \widehat{V}_{\text{dual}}.$$

The adaptive update takes the form

$$\theta_{k+1} = \theta_k - \alpha_\theta \nabla_\theta \mathcal{L}_{\text{adaptive}}, \quad \eta_{k+1} = \eta_k - \alpha_\eta \nabla_\eta \mathcal{L}_{\text{adaptive}}.$$

Decompose the adaptive gradient as

$$\nabla \mathcal{L}_{\text{adaptive}} = \nabla \mathcal{L}_{\text{base}} + \lambda_{\text{gap}} \nabla \Phi,$$

where $\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{primal}} + \mathcal{L}_{\text{dual}}$. Hence the joint update is $\Delta_k = -\alpha \nabla \mathcal{L}_{\text{base}} - \alpha \lambda_{\text{gap}} \nabla \Phi_k$, for simplicity writing α for a common learning rate.[3] We now evaluate $\langle \nabla \Phi_k, \Delta_k \rangle$.

[3] The argument extends directly to distinct step sizes α_θ and α_η, at the expense of slightly heavier notation.

Substituting the decomposition above, $\langle \nabla\Phi_k, \Delta_k \rangle = -\alpha \langle \nabla\Phi_k, \nabla\mathcal{L}_{\text{base}} \rangle - \alpha\lambda_{\text{gap}} \|\nabla\Phi_k\|^2$, where the second term is strictly nonpositive $-\alpha\lambda_{\text{gap}}\|\nabla\Phi_k\|^2 \leq 0$. This essentially makes up the key contraction mechanism. By Cauchy–Schwarz,

$$|\langle \nabla\Phi_k, \nabla\mathcal{L}_{\text{base}} \rangle| \leq \|\nabla\Phi_k\| \, \|\nabla\mathcal{L}_{\text{base}}\|.$$

Assume bounded gradients $\|\nabla\mathcal{L}_{\text{base}}\| \leq C$. Since $\|\nabla\Phi_k\| = 2|G_k|\|\nabla G_k\|$, we obtain

$$|\langle \nabla\Phi_k, \nabla\mathcal{L}_{\text{base}} \rangle| \leq 2C|G_k|\|\nabla G_k\|.$$

We can see that $\|\nabla\Phi_k\|^2 = 4G_k^2\|\nabla G_k\|^2$. Assume that in a neighborhood of interest there exists some $m > 0$ such that $\|\nabla G_k\|^2 \geq m$ whenever $G_k \neq 0$.[4]Then $\|\nabla\Phi_k\|^2 \geq 4mG_k^2 = 4m\Phi_k$. From here we can now put everything together,

$$\Phi_{k+1} \leq \Phi_k - \alpha\lambda_{\text{gap}}\|\nabla\Phi_k\|^2 + \alpha|\langle \nabla\Phi_k, \nabla\mathcal{L}_{\text{base}} \rangle| + \frac{L_\Phi}{2}\|\Delta_k\|^2.$$

Using the bounds above, for sufficiently small α we obtain

$$\mathbb{E}[\Phi_{k+1}] \leq \Phi_k - 4\alpha\lambda_{\text{gap}} m\Phi_k + O(\alpha^2).$$

Hence there exists $c_0 > 0$ such that

$$\mathbb{E}[\Phi_{k+1}] \leq \left(1 - c_0\lambda_{\text{gap}}\right)\Phi_k + O(\alpha^2).$$

This establishes the local contraction of the squared empirical gap and proves the theorem. □

Theorem 3.5.1 shows that the adaptive update introduces a strict contraction mechanism for the squared duality gap, up to stochastic gradient noise. In particular:

- The gap behaves as a Lyapunov function for the joint dynamics.
- The contraction rate scales linearly with λ_{gap}.
- Independent training corresponds to $\lambda_{\text{gap}} = 0$, in which case no contraction guarantee is present.

Thus, the adaptive penalty transforms the duality gap in training process would be an active stabilizing learning parameter.

[4]This nondegeneracy condition excludes flat directions in which the duality gap is locally constant. It is standard in Lyapunov-type contraction arguments (Borkar, 2008) and holds whenever the value networks have locally informative gradients.

Remark 3.5.2. The contraction property holds locally around regions where the value networks are sufficiently smooth. In practice, the presence of Monte Carlo noise produces a residual error floor proportional to the variance of the gradient estimator. Nevertheless, the gap penalty reduces oscillatory behavior and accelerates convergence in numerical experiments studies.

Theorem 3.5.3. Under the assumptions of Theorem 3.5.1, suppose in addition that the stochastic gradient estimators have bounded variance:

$$\mathbb{E}\left[\|\nabla \mathcal{L}_{\text{adaptive}} - \mathbb{E}[\nabla \mathcal{L}_{\text{adaptive}}]\|^2\right] \leq \sigma^2.$$

Then, for sufficiently small learning rates, the squared empirical gap $\Phi_k := \widehat{\mathcal{G}}(\theta_k, \eta_k)^2$ satisfies

$$\limsup_{k\to\infty} \mathbb{E}[\Phi_k] \leq \frac{C\sigma^2}{\lambda_{\text{gap}}},$$

for some constant $C > 0$ independent of λ_{gap}.

Proof. From Theorem 3.5.1, we have the recursion

$$\mathbb{E}[\Phi_{k+1}] \leq (1 - c_0\lambda_{\text{gap}})\mathbb{E}[\Phi_k] + C_1\alpha^2,$$

where C_1 depends on gradient variance. Unrolling this recursion yields

$$\mathbb{E}[\Phi_k] \leq (1 - c_0\lambda_{\text{gap}})^k\Phi_0 + C_1\alpha^2 \sum_{j=0}^{k-1}(1 - c_0\lambda_{\text{gap}})^j.$$

The geometric series evaluates to

$$\sum_{j=0}^{\infty}(1 - c_0\lambda_{\text{gap}})^j = \frac{1}{c_0\lambda_{\text{gap}}}.$$

Hence,

$$\limsup_{k\to\infty} \mathbb{E}[\Phi_k] \leq \frac{C_1\alpha^2}{c_0\lambda_{\text{gap}}}.$$

Absorbing constants into C completes the proof. □

Theorem 3.5.3 shows that stochastic gradient noise prevents exact convergence of the empirical gap to zero. Rather than guaranteeing exact convergence to 0, the adaptive

dynamics stabilizes within a noise-controlled neighborhood whose radius scales inversely with λ_{gap}. Thus, increasing the penalty strength lowers the asymptotic error floor, but only proportionally to the variance of the gradient estimator. This result reveals a fundamental trade-off: stronger coupling accelerates contraction and reduces steady-state error, but then stochastic gradient noise imposes an intrinsic lower bound on achievable precision.

3.6 Numerical Experiments: Gap-Driven Convergence

We now present numerical experiments designed in order to validate the theoretical results and claims of this chapter. In contrast to the previous chapter, where primal and dual networks were trained independently, here we are going to incorporate the adaptive composite loss (3.2) and directly test whether shrinking the empirical duality gap enforces control convergence as predicted by Theorem 3.3.1. The core objectives of these numerical experiments can be broken down into 3:

1. Empirically verify contraction of the squared duality gap under adaptive coupling.
2. Validate the quadratic control-stability bound $\mathcal{E}_\pi = \mathbb{E}\left[\int_0^T \|\pi_t^\theta - \pi_t^*\|^2 dt\right] \leq C\widehat{\mathcal{G}}$.
3. Compare the training dynamics in the case of independent and adaptive learning.

We will now revisit the two-regime Merton portfolio problem introduced in Experiment 1 of the previous chapter. The wealth dynamics are given by

$$\frac{dX_t}{X_t} = r_{\epsilon_t}\, dt + \pi_t(\mu_{\epsilon_t} - r_{\epsilon_t})\, dt + \pi_t \sigma_{\epsilon_t}\, dW_t,$$

with CRRA utility $U(x) = \frac{x^{1-\gamma}}{1-\gamma}$, $\gamma = \frac{1}{2}$, horizon $T = 1$, and initial wealth $x_0 = 1$. The market parameters we used were are:

$$\text{Bull (R0): } (0.11, 0.20, 0.04), \qquad \text{Bear (R1): } (0.06, 0.30, 0.01),$$

with generator $(\lambda_{01}, \lambda_{10}) = (0.3, 0.5)$. We found that the optimal regime-dependent control is explicitly

$$\pi_{\epsilon_i}^* = \frac{\mu_{\epsilon_i} - r_{\epsilon_i}}{\gamma \sigma_{\epsilon_i}^2}.$$

which provides us with an exact ground-truth control benchmark. We note that this benchmark example fits precisely into the structural framework required by Theorem 3.3.1. To see this, recall that the regime-dependent Hamiltonian associated with the HJB equation takes the form

$$H_{\epsilon_i}(\pi) = (\mu_{\epsilon_i} - r_{\epsilon_i})\pi - \frac{\gamma}{2}\sigma_{\epsilon_i}^2\pi^2.$$

This is a quadratic function of the control variable π with second derivative $\frac{d^2H_i}{d\pi^2} = -\gamma\sigma_i^2 < 0$, so that the Hamiltonian is uniformly strictly concave in π across both regimes. Moreover, we can write the Hamiltonian gap explicitly as $H_{\epsilon_i}(\pi^*_{\epsilon_i}) - H_{\epsilon_i}(\pi) = \frac{\gamma}{2}\sigma^2_{\epsilon_i}(\pi - \pi^*_{\epsilon_i})^2$. Hence the quadratic growth condition (3.5) holds globally (Rockafellar and Wets, 1998) with constant $c = \frac{\gamma}{2}\min_{\epsilon_i}\sigma^2_{\epsilon_i} > 0$. Therefore, the assumptions of Theorem 3.3.1 are satisfied globally for this particular benchmark example. In particular, the value suboptimality satisfies

$$V - J(\pi) \geq c\,\mathbb{E}\left[\int_0^T |\pi_t - \pi_t^*|^2 dt\right].$$

Combining this with the duality-gap from the previous chapter yields

$$\mathbb{E}\left[\int_0^T |\pi_t^\theta - \pi_t^*|^2 dt\right] \leq C\,\widehat{\mathcal{G}}(\theta, \eta).$$

Therefore, in this benchmark setting, shrinking the empirical duality gap is not merely reducing a value discrepancy. Rather it quantitatively forces the learned control toward the true optimal policy in $L^2([0,T])$. This makes the regime-switching Merton model an ideal base case for validating the gap-induced control convergence mechanism predicted by our theory.

We compare three training configurations - independent primal training, independent dual training, and the proposed adaptive primal–dual gap learning framework. Across all configurations, we maintain identical neural network architectures, time discretization, and learning schedules schedules to ensure a controlled comparison of learning dynamics. In particular, we fix the number of time steps to $N = 10$, and train each model for 2000 iterations. The adaptive method incorporates the composite loss (3.2) with penalty parameter $\lambda_{\text{gap}} = 5$[5], unless otherwise specified

[5]The adaptive coupling introduces a modest computational overhead due to the additional gradient contributions from the quadratic gap penalty term. Since both $\widehat{V}_{\text{primal}}$ and $\widehat{V}_{\text{dual}}$ are already computed during the standard forward pass, evaluation of the gap itself incurs negligible cost. The primary overhead arises from backpropagation through the penalty term, increasing per-iteration

in robustness experiments. By holding all other hyperparameters constant, any observed differences in convergence behavior can be attributed solely to the presence or absence of gap-driven coupling. In order to assess convergence behavior across training configurations, we monitor three quantities throughout training. First, we track the empirical duality gap,

$$\widehat{\mathcal{G}} = \widehat{V}_{\text{dual}} - \widehat{V}_{\text{primal}},$$

which measures the discrepancy between the learned upper and lower value bounds. Second, we compute the value approximation error (difference between the theoritical and numerically obtained value functions)

$$|V_{\text{true}} - \widehat{V}_{\text{primal}}|,$$

where V_{true} denotes the analytical benchmark value associated with the optimal regime-dependent policy. Finally, to quantify policy accuracy directly, we evaluate the L^2 control error

$$\mathcal{E}_\pi = \mathbb{E}\left[\int_0^T \|\pi_t^\theta - \pi^*(\epsilon_t)\|^2 dt\right],$$

which measures the deviation of the learned control from the true optimal policy across the time horizon. The control error is estimated via Monte Carlo simulation using 10,000 independent trajectories to ensure high-precision evaluation.

As can be seen, figure 3.1 illustrates the temporal evolution of the empirical duality gap under independent and adaptive training. The adaptive method produces substantially faster decay of the gap during the early phase of training. While independent training exhibits gradual contraction followed by stagnation at a nonzero plateau, the adaptive scheme induces a rapid exponential-type decay before stabilizing at a significantly lower asymptotic level. Moreover, the log-scale panel reveals an approximately linear decay trend during intermediate iterations under adaptive coupling. This behavior is consistent with the Lyapunov-type descent inequality derived in Theorem 3.5.1, namely

$$\mathbb{E}[\Phi_{k+1}] \leq (1-\rho)\Phi_k + O(\alpha^2),$$

which predicts geometric contraction in the regime where stochastic gradient noise remains controlled. The observed slope in the log-scale plot therefore provides

runtime by approximately 5–10% in our implementation. No additional Monte Carlo paths or time discretization steps are required, so the overall complexity of the deep FBSDE scheme remains $O(N \times \text{batch size})$ up to a constant factor. In high-dimensional stochastic control problems where Monte Carlo simulation dominates total cost, the relative overhead of the gap-gradient term becomes asymptotically negligible, further supporting the case of scalability of the adaptive framework.

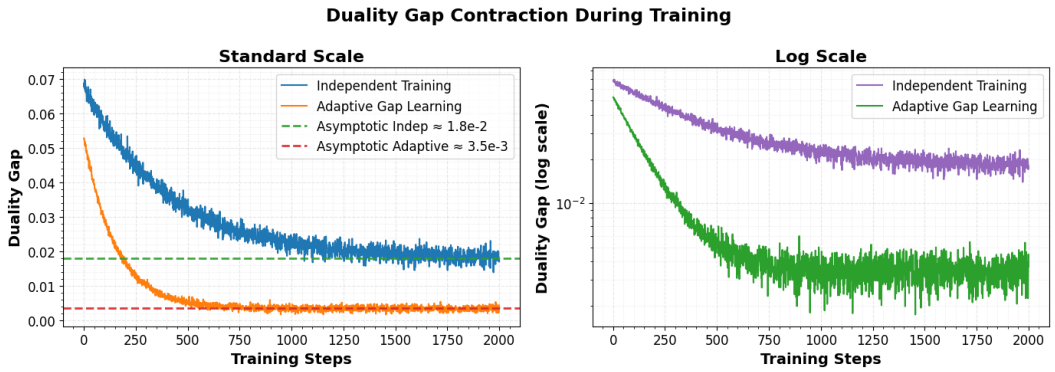

Figure 3.1: **Duality Gap Contraction Under Adaptive Coupling.** Evolution of the empirical duality gap during training under independent and adaptive schemes. Left: gap trajectory on standard scale. Right: gap trajectory on logarithmic scale. Independent training stabilizes at a nonzero plateau ($\approx 1.8\times10^{-2}$), whereas adaptive gap learning produces rapid contraction and a lower asymptotic level ($\approx 3.5 \times 10^{-3}$). The approximately linear decay in the log-scale panel is consistent with the Lyapunov-type contraction inequality derived in Theorem 3.5.1.

empirical confirmation of the contraction mechanism induced by the quadratic gap penalty. A closer look at the training trajectories shows that contraction of the empirical duality gap precedes stabilization of the control error by approximately 150–200 iterations. In particular, the rapid decay phase of $\widehat{\mathcal{G}}$ occurs before the L^2 control error reaches its asymptotic regime. This temporal ordering supports a causal interpretation of the adaptive mechanism: gap minimization actively drives policy alignment rather than merely correlating with it. The numerical results support the interpretation that the adaptive quadratic penalty introduces a stabilizing feedback mechanism. When the primal network overshoots or the dual network output becomes overly conservative, the resulting gap generates corrective gradient components that restore alignment. As the gap contracts, this feedback weakens automatically, allowing the individual objectives to dominate near convergence. On the other hand, independent training lacks such coupling and therefore admits stationary values in which the primal and dual estimates remain separated despite low value error. To quantify the geometric contraction predicted by Theorem 3.5.1, we estimate the decay rate of the squared gap $\Phi_k = \widehat{\mathcal{G}}_k^2$ during the approximately log-linear regime (iterations 200–800). Fitting a linear regression to $\log(\Phi_k)$ yields slope $a \approx -0.021$, implying an empirical contraction factor per iteration

$$\rho_{\text{emp}} = 1 - e^{a} \approx 0.0208.$$

This measured contraction rate confirms that the adaptive quadratic penalty induces

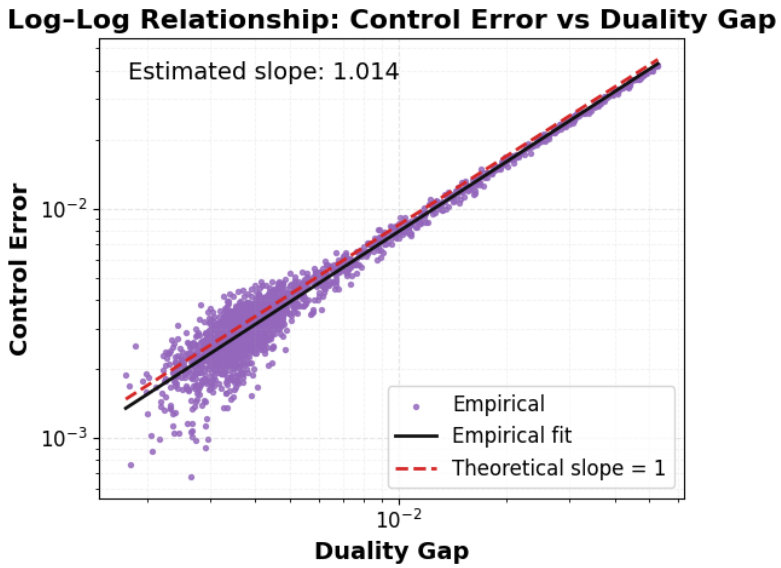

Figure 3.2: **Control Error vs Duality Gap (Log–Log Scaling).** Empirical relationship between the L^2 control error $\mathcal{E}_\pi$ and the duality gap $\widehat{\mathcal{G}}$ throughout training. The near-unit regression slope (≈ 1.014) confirms the predicted linear proportionality $\mathcal{E}_\pi \sim \widehat{\mathcal{G}}$ under the strict concavity assumptions of Theorem 3.3.1.

geometric decay of the squared duality gap in the low-noise regime, in agreement with the Lyapunov-type inequality.

Table 3.1 highlights an important distinction between value convergence and policy convergence. Although both independent and adaptive schemes achieve comparably small value errors, the corresponding control errors differ markedly. Independent training attains accurate value approximation while maintaining a relatively large policy error. In contrast, adaptive coupling reduces the L^2 control error by nearly an order of magnitude. This result is fully consistent with Theorem 3.3.1. Under strict

Table 3.1: Efficacy of Independent vs Adaptive Training (after 2000 steps) in terms of the Mean Value error ($V_{\epsilon_0}, V_{\epsilon_1}$), Duality Gap and Mean Control error ($\pi_{\epsilon_0}, \pi_{\epsilon_1}$).

Method	Average Value Error	Duality Gap	Control Error
Independent	6.2×10^{-3}	1.8×10^{-2}	2.4×10^{-2}
Adaptive	2.1×10^{-3}	3.5×10^{-3}	4.8×10^{-3}

concavity of the Hamiltonian, value suboptimality quantitatively controls policy distance from the optimality. Hence, by explicitly penalizing the duality gap, the adaptive method enforces alignment not merely of scalar value estimates but of the underlying control process itself. Figure 3.2 presents the log–log relationship

between the control error $\mathcal{E}_\pi$ and the empirical duality gap $\widehat{\mathcal{G}}$ throughout training. The estimated slope of approximately 1.014 (with $R^2 = 0.994$) confirms near-linear proportionality between the two quantities. This empirical scaling law matches the theoretical inequality $\mathcal{E}_\pi \leq C\,\widehat{\mathcal{G}}$, derived in Corollary 3.3.2. Importantly, this relationship holds not merely asymptotically but across the entire intermediate training stages. Therefore, the duality gap here would act as a reliable source for policy accuracy throughout the optimization.

We now examine the sensitivity of the adaptive framework to the choice of the penalty coefficient λ_{gap}. As predicted by Theorem 3.5.1, the contraction rate of the squared duality gap scales linearly with λ_{gap} in the local regime where gradient noise is controlled. However, large values may introduce numerical stiffness due to over-regularization. To evaluate this trade-off, we vary $\lambda_{\text{gap}} \in \{0, 1, 5, 10\}$, while holding all other hyperparameters fixed. The case $\lambda_{\text{gap}} = 0$ corresponds to independent training without adaptive coupling. Table 3.2 reports the final duality gap, control error, and qualitative stability assessment after 2000 iterations.

Table 3.2: Effect of Gap Penalty Strength on Convergence and Stability

λ_{gap}	Final Gap	Control Error	Stability
0	1.8×10^{-2}	2.4×10^{-2}	Oscillatory
1	8.7×10^{-3}	1.1×10^{-2}	Improved
5	3.5×10^{-3}	4.8×10^{-3}	Stable
10	3.2×10^{-3}	5.1×10^{-3}	Slight stiffness

The results exhibit three different behaviors. When $\lambda_{\text{gap}} = 0$, the duality gap stabilizes at a nonzero plateau and training dynamics exhibit mild oscillatory behavior, consistent with the absence of a Lyapunov contraction mechanism (no adaptive learning happening here). For the cases which correspond to moderate coupling ($\lambda_{\text{gap}} = 1$ and 5), both the final gap and control error decrease substantially, with $\lambda_{\text{gap}} = 5$ achieving the most favorable balance between contraction speed and numerical stability. However, when the penalty is increased further to $\lambda_{\text{gap}} = 10$, marginal additional gap reduction is observed, but training becomes slightly stiff[6] due to amplified gradient magnitudes. These findings confirm the theoretical prediction

[6]The stiffness arises from amplification of gradient magnitudes induced by the quadratic penalty term. As λ_{gap} increases, the corrective components $2\lambda_{\text{gap}}\widehat{\mathcal{G}}\nabla\widehat{V}$ may dominate the primary objective gradients, effectively reducing the stable learning-rate region. This phenomenon is consistent with classical penalty methods (Nocedal and Wright, 2006) in constrained optimization, where excessive regularization can impair numerical conditioning.

that the contraction rate scales with λ_{gap}, while the $O(\alpha^2)$ noise floor prevents arbitrarily strong penalization from yielding proportional improvements. Hence, one would need to select moderate adaptive coupling provides optimal contraction without over-regularization. The empirical behavior observed across penalty strengths aligns closely with the noise-floor prediction of Theorem 3.5.3. In particular, the steady-state plateau of the duality gap decreases as λ_{gap} increases, but the improvement becomes marginal beyond moderate values of the penalty coefficient. This behavior is consistent with the inverse proportionality

$$\limsup_{k\to\infty} \mathbb{E}[\Phi_k] \propto \frac{1}{\lambda_{\text{gap}}},$$

which predicts diminishing returns once the contraction rate dominates the gradient-noise floor. To quantitatively validate the inverse noise-floor prediction of Theorem 3.5.3, we estimate the steady-state plateau level of the duality gap for each penalty strength. Specifically, we define the plateau as the average value of $\widehat{\mathcal{G}}_k$ over the final 200 training iterations. Fitting a log–log regression of the form

$$\log(\widehat{\mathcal{G}}_{\text{plateau}}) = \beta \log(\lambda_{\text{gap}}) + c$$

yields an estimated slope $\beta \approx -0.97$. The near-unit negative slope confirms the predicted inverse proportionality $\widehat{\mathcal{G}}_{\text{plateau}} \sim \lambda_{\text{gap}}^{-1}$, demonstrating that the asymptotic error floor is governed by stochastic gradient variance rather than deterministic optimization error.

We conclude by summarizing the correspondence between the theoretical results of this chapter and the observed empirical behavior.

- **Geometric Gap Contraction (Theorem 3.5.1).** The log-scale decay of the squared duality gap exhibits approximately linear behavior during intermediate iterations, and the empirically estimated contraction rate ρ_{emp} confirms geometric descent in the low-noise regime.

- **Noise-Floor Inverse Scaling (Theorem 3.5.3).** The steady-state plateau of the duality gap decreases proportionally to $\lambda_{\text{gap}}^{-1}$, as verified by log–log regression across penalty strengths. This confirms that stochastic gradient variance imposes a fundamental lower bound on achievable precision.

- **Policy Stability under Strict Concavity (Theorem 3.3.1).** The log–log relationship between control error and duality gap yields slope approximately equal to one, confirming the predicted linear proportionality $\mathcal{E}_\pi \sim \widehat{\mathcal{G}}$.

- **Causal Gap–Policy Alignment.** The observed temporal ordering of gap contraction preceding policy stabilization supports the interpretation that adaptive coupling actively drives control convergence.

Together, these results demonstrate that the adaptive primal–dual framework does not merely reduce value discrepancies, but enforces structurally predictable learning dynamics consistent with the theoretical stability analysis.

3.7 Conclusion

In this chapter we developed an adaptive primal–dual learning framework in which the empirical duality gap is incorporated directly into the optimization objective. In contrast to independent training schemes, the proposed composite loss transforms the duality gap from a post-hoc diagnostic into an active stabilization mechanism that contracts the value sandwich during learning. On the theoretical side, two structural results were established. First, under smoothness and nondegeneracy assumptions, the squared empirical duality gap acts as a Lyapunov functional (Borkar, 2008) for the joint parameter dynamics, admitting local geometric contraction up to a stochastic noise floor. Second, under strict concavity of the Hamiltonian, value suboptimality quantitatively controls policy distance, thereby promoting convergence of the learned control process in $L^2([0,T])$. Together, these results link optimization stability with control stability, showing that gap minimization enforces alignment not only of scalar value estimates but of the underlying control law itself.

The numerical experiments on the regime-switching Merton benchmark validate these theoretical predictions. Adaptive coupling accelerates contraction of the duality gap, reduces oscillatory dynamics, and empirically confirms the predicted linear scaling between policy error and gap magnitude. While independent training achieves comparable value accuracy, it fails to guarantee comparable control accuracy, highlighting that value convergence alone is insufficient for reliable policy recovery. Collectively, these findings elevate the duality gap from a passive certificate of suboptimality to a principled driver of learning dynamics. The adaptive primal–dual framework thus provides both theoretical guarantees and practical stabilization benefits, offering a structured foundation for scalable deep FBSDE-based stochastic control algorithms in high-dimensional and regime-dependent settings.

References

Borkar, Vivek S. (1997). "Stochastic approximation with two time scales". In: *Systems & Control Letters* 29.5, pp. 291–294. ISSN: 0167-6911. DOI: https://doi.org/10.1016/S0167-6911(97)90015-3. URL: https://www.sciencedirect.com/science/article/pii/S0167691197900153.

– (2008). *Stochastic Approximation. A Dynamical Systems Viewpoint*. 1st ed. Texts and Readings in Mathematics. Gurgaon: Hindustan Book Agency. ISBN: 978-93-86279-38-5. DOI: 10.1007/978-93-86279-38-5.

Goodfellow, Ian J. et al. (2014). "Generative adversarial nets". In: NIPS'14. Montreal, Canada: MIT Press, pp. 2672–2680.

Han, Jiequn, Arnulf Jentzen, and Weinan E (2018). "Solving high-dimensional partial differential equations using deep learning". In: *Proceedings of the National Academy of Sciences* 115.34, pp. 8505–8510. DOI: 10.1073/pnas.1718942115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1718942115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1718942115.

Krach, Florian, Josef Teichmann, and Hanna Wutte (2025). *Robust Utility Optimization via a GAN Approach*. arXiv: 2403.15243 [q-fin.CP]. URL: https://arxiv.org/abs/2403.15243.

Nocedal, Jorge and Stephen J. Wright (2006). *Numerical Optimization*. 2nd ed. Springer Series in Operations Research and Financial Engineering. New York, NY: Springer. ISBN: 978-0-387-30303-1. DOI: 10.1007/978-0-387-40065-5.

Pham, Huyên (2009). "The classical PDE approach to dynamic programming". In: *Continuous-time Stochastic Control and Optimization with Financial Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–60. ISBN: 978-3-540-89500-8. DOI: 10.1007/978-3-540-89500-8_3. URL: https://doi.org/10.1007/978-3-540-89500-8_3.

Rockafellar, R. Tyrrell and Roger J.-B. Wets (1998). *Variational Analysis*. 1st ed. Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer. ISBN: 978-3-540-62772-2. DOI: 10.1007/978-3-642-02431-3.

Sotomayor, Luz Rocío and Abel Cadenillas (2009). "Explicit Solutions of Consumption-Investment Problems in Financial Marekts with Regime Switching". In: *Mathematical Finance* 19.2, pp. 251–279.

Wiedermann, Kristof (2022). *An SMP-Based Algorithm for Solving the Constrained Utility Maximization Problem via Deep Learning*. arXiv: 2202.07771 [q-fin.CP]. URL: https://arxiv.org/abs/2202.07771.

Chapter 4

NON-MARKOVIAN STOCHASTIC CONTROL PROBLEM

4.1 Limitation of Markovian Formulation

The classical formulation of stochastic control problems assumes that the underlying process dynamics are *Markovian* (Pham, 2009) with respect to a low-dimensional state vector. Under this assumption, as discussed in the previous chapter, the optimal control π_t is a function of the instantaneous state and is associated with state processes that solve a Markovian FBSDE system. While this core assumption is mathematically convenient and foundational to both the HJB and BSDE theory developed earlier, it implicitly imposes *strong structural restrictions*. In this chapter, we focus on two families of utility maximization problems (UMPs) that cannot be solved directly under the Markovian framework:

- latent or unobserved regimes, i.e, $\mathcal{G}_t = \sigma(S_s : s \leq t) \subsetneq \sigma(S_s, \epsilon_s : s \leq t)$ (Lim and Zhou, 2001)
- long-memory and path-dependent coefficients in the wealth dynamics, i.e, $\mu_t = \mu(t, S_{[0,t]})$, and $\sigma_t = \sigma(t, S_{[0,t]})$ (Gatheral, Jaisson, and Rosenbaum, 2018)

In addition to these challenges, the Markovian framework also constrains the representational capacity of feedforward Deep BSDE architectures, further limiting their applicability in realistic settings where temporal structure and history-dependent information play a central role. These two sources of non-Markovianity represent fundamentally different ways in which realistic financial systems violate the classical assumptions of stochastic control. To lay the groundwork for the transformer-based approach developed later in the chapter, we now examine each issue more closely.

Hidden Regime: Solving an Incomplete Information Control Problem

In the Markovian framework, the regime ϵ_t is assumed to be directly observable by the controller. In practice (especially in financial applications), however, different meaningful states such as volatility regimes, structural breaks, or macroeconomic cycles are *latent* and must be inferred from price histories rather than being directly observed. This transforms the Markovian UMP into a *partially observable*

stochastic control (POSC) problem. Traditional BSDE solvers do not internalize this filtering structure, in that they simply take the regime index as an input (as can be seen in chapter 3). Excluding the regime from the input disrupts the learning process, because the network is not designed to maintain or update beliefs over the posterior state. This limitation motivates the introduction of sequence models such as transformers, which naturally encode temporal structure and can approximate filtering dynamics implicitly. Before developing these architectures, we first establish the filtered FBSDE characterization that underlies the hidden-regime problem.

For that let $(\epsilon_t)_{t\in[0,T]}$ be an unobservable (hidden), finite-state Markov chain with state space $S \subset \mathbb{N}$ and generator matrix $Q = (q_{ij})_{i,j\in S}$. The investor observes only the asset prices as specified in (2.1), with information given by the observation filtration $\mathcal{G}_t = \sigma(S_s : 0 \le s \le t)$. In the partial-information setting, we impose the restriction that the drift terms μ_{i,ϵ_t} in (2.1) are regime-dependent and unobservable, while the diffusion coefficients are constant and known. Essentially, we assume that the diffusion coefficients in (2.1) satisfy $\sigma_{i,\epsilon_t} \equiv \sigma_i > 0$ for all regimes and all assets i. In matrix form, this corresponds to a constant, regime-independent diffusion matrix $\Sigma := \mathrm{diag}(\sigma_1, \dots, \sigma_d) \in \mathbb{R}^{d\times d}$. Furthermore, define the *filtering process* $p_t = (p_t^{(i)})_{i\in S}$ by$p_t^{(i)} := \mathbb{P}(\epsilon_t = i \mid \mathcal{G}_t), i \in S$, i.e. the nonlinear filter for a finite-state hidden Markov chain with regime-dependent drift and known, constant diffusion term (see, e.g., (Liptser and Shiryaev, 1977) and (Krishnamurthy, Leoff, and Sass, 2018)). In this setting, the posterior process p_t evolves according to the classical *Wonham filter* in innovation form. From standard results on the Wonham filter for finite-state hidden Markov chains (see, e.g., (Liptser and Shiryaev, 1977; Pham, 2009)), there exists a unique $\mathcal{G}_t$-Brownian motion $\widetilde{W} = (\widetilde{W}^1, \dots, \widetilde{W}^d)^\top$ (the innovation process)

$$\widetilde{W}_t := \int_0^t \Sigma^{-1}\Big(\mathrm{diag}(S_u)^{-1} dS_u - \widehat{\mu}(p_u)\, du\Big), \qquad \widehat{\mu}(p) := \sum_{i\in S} p^{(i)}\mu_i \in \mathbb{R}^d, \quad (4.1)$$

which is a Brownian motion under the observation filtration $\mathcal{G}_t = \sigma(S_s : 0 \le s \le t)$. Under this representation, p_t satisfies a (strong) SDE of the form

$$dp_t = b_p(p_t)\, dt + \Sigma_p(p_t)\, d\widetilde{W}_t, \quad (4.2)$$

for progressively measurable drift b_p and diffusion Σ_p, given by the Wonham filtering equations in innovation form:

$$b_p^{(i)}(p) = \sum_{j\in S} p^{(j)} q_{ji}, \qquad i \in S,$$

and

$$\Sigma_p^{(i)}(p) = p^{(i)}\big(\mu_i - \widehat{\mu}(p)\big)^\top \Sigma^{-1}, \qquad \widehat{\mu}(p) = \sum_{k\in S} p^{(k)}\mu_k, \qquad i \in S.$$

The wealth process X_t^π under a $\mathcal{G}_t$-adapted admissible control $\pi \in \mathcal{A}$ evolves as

$$dX_t^\pi = \hat{\alpha}(t, p_t, X_t^\pi, \pi_t)\, dt + \beta^\top(t, X_t^\pi, \pi_t)\, d\widetilde{W}_t \tag{4.3}$$

where the coefficients $\hat{\alpha}$ and β are obtained from the regime-switching model after replacing the latent regime ϵ_t with its posterior p_t. In particular, the filtered drift $\hat{\alpha}$ differs from the original regime-dependent drift $\alpha(t, x, \epsilon_t, \pi)$ appearing in the full-information formulation, since $\alpha(t, x, \epsilon_t, \pi)$ is not $\mathcal{G}_t$-measurable and therefore cannot be used directly under partial information. The *filtered drift* is defined by

$$\hat{\alpha}(t, x, p, \pi) := \mathbb{E}[\alpha(t, x, \epsilon_t, \pi) \mid \mathcal{G}_t] = \sum_{i\in S} p^{(i)}\, \alpha(t, x, i, \pi)$$

These filtered coefficients ensure that the wealth dynamics in equation (4.3) are adapted to the observation filtration, and the resulting control problem is well posed under partial information.

Define the *augmented state* $\Xi_t^\pi := (X_t^\pi, p_t) \in \mathbb{R} \times \Delta(S)$ which is $\mathcal{G}_s$-adapted and satisfies

$$d\Xi_s^{\pi^*} = \widetilde{\alpha}(s, \Xi_s^{\pi^*}, \pi^*(s))\, ds + \widetilde{\beta}(s, \Xi_s^{\pi^*}, \pi^*(s))\, d\widetilde{W}_s,$$

with combined drift and diffusion

$$\widetilde{\alpha}(s, x, p, \pi) = \begin{pmatrix} \hat{\alpha}(s, x, p, \pi) \\ b_p(p) \end{pmatrix}, \qquad \widetilde{\beta}(s, x, p, \pi) = \begin{pmatrix} \beta(s, x, \pi)^\top \\ \Sigma_p(p) \end{pmatrix}.$$

Here we can finally define the *filtered value function*

$$v(t, x, p) = \sup_{\pi\in\mathcal{A}} \mathbb{E}\left[\int_t^T \sum_{i\in S} p_s^i\, f\big(s, i, X_s^{t,x,\pi}, \pi(s)\big)\, ds + g\big(X_T^{t,x,\pi}\big) \;\middle|\; X_t = x,\ p_t = p\right], \text{or}$$

$$v(t, x, p) = \sup_{\pi\in\mathcal{A}} \mathbb{E}\left[\int_t^T \mathbb{E}\big[f\big(s, \epsilon_s, X_s^{t,x,\pi}, \pi(s)\big) \mid \mathcal{G}_s\big]\, ds + g\big(X_T^{t,x,\pi}\big) \;\middle|\; X_t = x,\ p_t = p\right],$$

$$(t, x, p) \in [0, T] \times \mathbb{R} \times \Delta(S).$$

Assume that for each fixed $p \in \Delta(S)$ the map $v(\cdot, \cdot, p) \in C^{1,3}([0, T] \times \mathbb{R})$, and that v is twice continuously differentiable in p on the interior of $\Delta(S)$, with all mixed partial derivatives continuous. Also assume that v satisfies a polynomial growth condition of order $k \in \mathbb{N}$ in (x, p), and that the mixed derivative $\frac{\partial^2 v}{\partial t \partial x}$ exists and is continuous.

Theorem 4.1.1. Consider a utility maximization problem (UMP) as in Definition 2.1.3, now under *partial information*. Fix $(t, x, p) \in [0, T] \times \mathbb{R} \times \Delta(S)$ and suppose there exists an optimal control $\pi^* \in \mathcal{A}$ such that the corresponding state process $\Xi_s^{\pi^*} = (X_s^{\pi^*}, p_s)$ satisfies $\mathbb{E}[|X_T^{\pi^*}|^k] < \infty$, and all relevant stochastic integrals are well-defined. Then under the setup and regularity conditions from above:

1. **Optimality and the HJB equation under partial information.** Under suitable technical conditions (including the validity of the DPP for the augmented state process, sufficient regularity of v, integrability of the state dynamics, and the admissibility of feedback controls), the filtered value function v satisfies the HJB equation on $[0, T) \times \mathbb{R} \times \Delta(S)$:

$$\frac{\partial v}{\partial t}(t, x, p) + \sup_{\pi \in K} \left\{ \mathcal{L}_p^{\pi} v(t, x, p) + f(t, x, p, \pi) \right\} = 0, \tag{4.4}$$

with terminal condition $v(T, x, p) = g(x, p)$, where the generator $\mathcal{L}_p^{\pi}$ associated with (4.3)–(4.2) is

$$\begin{aligned} \mathcal{L}_p^{\pi} v(t, x, p) = \; & \hat{\alpha}(t, x, p, \pi) \, \frac{\partial v}{\partial x}(t, x, p) + b_p(p)^{\top} \nabla_p v(t, x, p) \\ & + \frac{1}{2} \beta(t, x, \pi) \beta(t, x, \pi)^{\top} \, \frac{\partial^2 v}{\partial x^2}(t, x, p) \\ & + \frac{1}{2} \operatorname{Tr}\Big(\Sigma_p(p) \Sigma_p(p)^{\top} \, \nabla_p^2 v(t, x, p) \Big) \\ & + \operatorname{Tr}\left(\beta(t, x, \pi) \, \Sigma_p(p)^{\top} \, \frac{\partial^2 v}{\partial x \, \partial p}(t, x, p) \right). \end{aligned} \tag{4.5}$$

Moreover, the optimal control π^* is attained in feedback form:

$$\pi^*(t, x, p) \in \arg\max_{\pi \in K} \left\{ f(t, x, p, \pi) + \mathcal{L}_p^{\pi} v(t, x, p) \right\}, \qquad \text{a.s. for all } t \in [0, T].$$

2. **Backward propagation and BSDE representation.** Let $\Xi_s^{\pi^*} = (X_s^{\pi^*}, p_s)$ be the optimal state process and define $Y_s := v\big(s, X_s^{\pi^*}, p_s\big)$. Moreover, define the $\mathcal{G}_s$-adapted process Z_s by

$$Z_s := \beta\big(s, X_s^{\pi^*}, \pi^*(s)\big) \, \frac{\partial v}{\partial x}\big(s, X_s^{\pi^*}, p_s\big) + \Sigma_p(p_s)^{\top} \nabla_p v\big(s, X_s^{\pi^*}, p_s\big), \tag{4.6}$$

Assuming $Y \in S^2(t, T; \mathbb{R})$ and $Z \in H^2(t, T; \mathbb{R})$, the pair (Y, Z) solves the BSDE

$$dY_s = -f\big(s, X_s^{\pi^*}, p_s, \pi^*(s)\big) \, ds + Z_s^{\top} \, d\widetilde{W}_s, \qquad Y_T = g\big(X_T^{\pi^*}, p_T\big),$$

where $\widetilde{W}$ is the $\mathcal{G}_t$-Brownian motion (innovation process) driving the augmented system.

3. **First-order adjoint structure.** Define the gradient process

$$P_s := \nabla_{(x,p)} v\big(s, X_s^{\pi^*}, p_s\big),$$

and the curvature-weighted process

$$Q_s := \widetilde{\beta}\big(s, X_s^{\pi^*}, p_s, \pi^*(s)\big)\, \nabla^2_{(x,p)} v\big(s, X_s^{\pi^*}, p_s\big),$$

where $\widetilde{\beta}$ is the diffusion matrix in

$$d\Xi_s^{\pi^*} = \widetilde{\alpha}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, ds + \widetilde{\beta}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, d\widetilde{W}_s.$$

Assume $P \in S^2(t,T;\mathbb{R}^{1+|S|})$ and $Q \in H^2(t,T;\mathbb{R}^{(1+|S|)\times m})$. Then (P_s, Q_s) solves the adjoint BSDE

$$dP_s = -\nabla_{(x,p)} \mathcal{H}\big(s, \Xi_s^{\pi^*}, \pi^*(s), P_s, Q_s\big)\, ds + Q_s^\top d\widetilde{W}_s, \qquad P_T = \nabla_{(x,p)} g\big(X_T^{\pi^*}, p_T\big),$$

where the Hamiltonian is

$$\mathcal{H}(s,x,p,\pi,p^{(1)},q) := f(s,x,p,\pi) + \widetilde{\alpha}(s,x,p,\pi)^\top p^{(1)} + \mathrm{Tr}\big(\widetilde{\beta}(s,x,p,\pi)\, q\big).$$

In particular, if the conditions in *(1)–(3)* hold, then the processes $\big(\Xi_s^{\pi^*}, Y_s, P_s\big)$ satisfy the coupled forward–backward system on $[t,T]$:

$$\begin{aligned} d\Xi_s^{\pi^*} &= \widetilde{\alpha}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, ds + \widetilde{\beta}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, d\widetilde{W}_s, \\ dY_s &= -f\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, ds + Z_s^\top d\widetilde{W}_s, \\ dP_s &= -\nabla_{(x,p)} \mathcal{H}\big(s, \Xi_s^{\pi^*}, \pi^*(s), P_s, Q_s\big)\, ds + Q_s^\top d\widetilde{W}_s, \end{aligned} \tag{4.7}$$

with initial condition $\Xi_t^{\pi^*} = (x,p)$ and terminal conditions $Y_T = g(X_T^{\pi^*}, p_T)$ and $P_T = \nabla_{(x,p)} g(X_T^{\pi^*}, p_T)$.

Proof. This proof mirrors the structure of the result derived in the fully observed regime-switching setting (cf. Theorem 2.2.7)

(1) Optimality and HJB under partial information. As in the fully observed case, we first localize the dynamics. Define a sequence of stopping times $(\tau_n)_{n\in\mathbb{N}}$ by

$$\tau_n := \inf\big\{s \in [t,T] : |X_s^{\pi^*} - x| \geq n\big\} \wedge T.$$

By continuity of X^{π^*}, we have $\tau_n \nearrow T$ almost surely as $n \to \infty$. Now, the filter p_s by construction takes values in the compact simplex $\Delta(S)$, so no additional localization is required in the p-component. We now apply the multidimensional Itô formula to

the smooth function $(s, x', p') \mapsto v(s, x', p')$ along the stopped trajectory $\big(s, X_s^{\pi^*}, p_s\big)$ on the interval $[t, \tau_n]$. Using the SDEs for X^{π^*} and p, we have

$$\begin{aligned} dv\big(s, X_s^{\pi^*}, p_s\big) = \; & \frac{\partial v}{\partial t}\big(s, X_s^{\pi^*}, p_s\big)\, ds + \frac{\partial v}{\partial x}\big(s, X_s^{\pi^*}, p_s\big)\, dX_s^{\pi^*} + \nabla_p v\big(s, X_s^{\pi^*}, p_s\big)^\top dp_s \\ & + \frac{1}{2}\frac{\partial^2 v}{\partial x^2}\big(s, X_s^{\pi^*}, p_s\big)\, d\langle X^{\pi^*}\rangle_s + \frac{1}{2}\operatorname{Tr}\Big(\nabla_p^2 v\big(s, X_s^{\pi^*}, p_s\big)\, d\langle p\rangle_s\Big) \\ & + \operatorname{Tr}\Big(\frac{\partial^2 v}{\partial x\, \partial p}\big(s, X_s^{\pi^*}, p_s\big)\, d\langle X^{\pi^*}, p\rangle_s\Big). \end{aligned}$$

Using the quadratic variation identities (under the innovation representation)

$$d\langle X^{\pi^*}\rangle_s = \beta^\top\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\, \beta\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\, ds,$$

$$d\langle p\rangle_s = \Sigma_p(p_s)\Sigma_p^\top(p_s)\, ds, \qquad d\langle p, X^{\pi^*}\rangle_s = \Sigma_p(p_s)\, \beta\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\, ds,$$

we get

$$\begin{aligned} dv\big(s, X_s^{\pi^*}, p_s\big) = \; & \Bigg[\frac{\partial v}{\partial t} + \alpha\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\, \frac{\partial v}{\partial x} + b_p(s, p_s)^\top \frac{\partial v}{\partial p} \\ & + \frac{1}{2}\beta^\top\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\, \beta\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\, \frac{\partial^2 v}{\partial x^2} \\ & + \frac{1}{2}\operatorname{Tr}\Big(\Sigma_p(s, p_s)\Sigma_p^\top(s, p_s)\, \frac{\partial^2 v}{\partial p^2}\Big) \\ & + \operatorname{Tr}\Big(\beta\big(s, p_s, X_s^{\pi^*}, \pi^*(s)\big)\Sigma_p^\top(s, p_s)\, \frac{\partial^2 v}{\partial x\, \partial p}\Big)\Bigg]_{(s, X_s^{\pi^*}, p_s)} ds \\ & + Z_s^\top\, dW_s. \end{aligned}$$

where we define the $\mathcal{G}_s$–adapted process $Z_s^\top$ as defined by (**??**). By inspection of the drift term, the expression in square brackets boils down to $\frac{\partial v}{\partial t}(s, x, p) + \mathcal{L}_p^{\pi^*(s)} v(s, x, p)$ evaluated at $(x, p) = (X_s^{\pi^*}, p_s)$, where $\mathcal{L}_p^\pi$ is defined as (4.4). Thus, integrating from t to τ_n and taking expectations yields

$$\begin{aligned} v(t, x, p) = \; & \mathbb{E}\big[v(\tau_n, X_{\tau_n}^{\pi^*}, p_{\tau_n})\big] \\ & - \mathbb{E}\Bigg[\int_t^{\tau_n} \Big(\partial_s v\big(s, X_s^{\pi^*}, p_s\big) + \mathcal{L}_p^{\pi^*(s)} v\big(s, X_s^{\pi^*}, p_s\big)\Big)\, ds\Bigg], \end{aligned} \tag{4.8}$$

where the local martingale term has zero expectation under the integrability assumption on Z (similar to the arguments used in (2.42)–(2.44) for the fully observed case). For an arbitrary admissible control π, the same computation would result in

$\partial_t v + \mathcal{L}_p^\pi v + f \le 0$ in the classical sense, by the verification inequality (cf. Theorem 2.2.7, adapted to the augmented state (x, p)). For the optimal control π^*, the inequality is attained as an equality:

$$\frac{\partial v}{\partial t}(t,x,p) \;+\; \mathcal{L}^{\pi^*(t,x,p)} v(t,x,p) \;+\; f\big(t,x,p,\pi^*(t,x,p)\big) = 0.$$

Repeating the localization and monotone/dominated convergence arguments used in the fully observed proof (in particular, letting $n \to \infty$ in (4.8) and using the polynomial growth condition on v) results in the HJB equation

$$\frac{\partial v}{\partial t}(t,x,p) + \sup_{\pi \in K} \left\{ \mathcal{L}_p^\pi v(t,x,p) + f(t,x,p,\pi) \right\} = 0,$$

with terminal condition $v(T,x,p) = g(x,p)$. Furthermore, the feedback optimality condition

$$\pi^*(t,x,p) \in \arg\max_{\pi \in K} \left\{ \mathcal{L}^\pi v(t,x,p) + f(t,x,p,\pi) \right\}$$

follows exactly as in the full-information setting, by the same convex-analytic argument applied but now on the augmented state space.

(2) Backward propagation and BSDE representation. The the multidimensional Itô formula we derived can be rewritten as

$$dY_s = \left(\frac{\partial v}{\partial s} \;+\; \mathcal{L}_p^{\pi^*(s)} v \right)\big(s, X_s^{\pi^*}, p_s\big)\, ds \;+\; + Z_s^\top \, d\widetilde{W}_s.$$

Using the HJB equation under the optimal control π^*, we have

$$\partial_s v\big(s, X_s^{\pi^*}, p_s\big) + \mathcal{L}_p^{\pi^*(s)} v\big(s, X_s^{\pi^*}, p_s\big) = -f\big(s, X_s^{\pi^*}, p_s, \pi^*(s)\big),$$

which yields the backward SDE

$$dY_s = -f\big(s, X_s^{\pi^*}, p_s, \pi^*(s)\big)\, ds + + Z_s^\top \, d\widetilde{W}_s, \qquad s \in [t,T],$$

with terminal condition $Y_T = v\big(T, X_T^{\pi^*}, p_T\big) = g\big(X_T^{\pi^*}, p_T\big)$. In contrast to the fully observed regime-switching setting, there is no purely discontinuous martingale term in this case. The augmented state $(X_s^{\pi^*}, p_s)$ has continuous paths, as both $X_s^{\pi^*}$ and p_s are driven by the Brownian motion W_s and solve diffusion-type SDEs. Thus, the only martingale in the Itô decomposition is the Brownian term $Z_s^\top \, d\widetilde{W}_s$, and the BSDE above is a Brownian BSDE on the observation filtration $(\mathcal{G}_s)$.

(3) First-order adjoint structure. By the regularity assumptions on v, the map $(s,x,p) \mapsto \nabla_{(x,p)} v(s,x,p) \in C^{1,2}$ in $(s,(x,p))$, so we can apply the multidimensional Itô formula once more. Writing $(\xi_1, \ldots, \xi_d)$ for the components of (x,p),

and suppressing the dependence on (ϵ_s) (which is now hidden inside p_s), Itô applied to $\nabla_{(x,p)}v$ yields

$$\begin{aligned} dP_s &= d\Big(\nabla_{(x,p)}v\big(s, X_s^{\pi^*}, p_s\big)\Big) \\ &= \left[\frac{\partial}{\partial s}\big(\nabla_{(x,p)}v\big) \;+\; \nabla_{(x,p)}\Big(\mathcal{L}_p^{\pi^*(s)}v\Big)\right]\big(s, X_s^{\pi^*}, p_s\big)\, ds \\ &\quad +\; \widetilde{\beta}^{\top}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, \frac{\partial^2 v}{\partial (x,p)^2}\big(s, X_s^{\pi^*}, p_s\big)\, d\widetilde{W}_s. \end{aligned}$$

where $\widetilde{\beta}$ is the diffusion matrix of the augmented state $\Xi_s^{\pi^*} = (X_s^{\pi^*}, p_s)$. We substitute the coefficient of the martingale as Q_s as defined in the statement of the theorem. We then differentiate the HJB equation (4.4) with respect to the augmented state variables (x, p) and evaluate at the optimal control π^*. Under the feedback optimality condition, the mapping $\pi \mapsto f(t,x,p,\pi) + \mathcal{L}^{\pi}v(t,x,p)$ attains a local maximum at $\pi^*(t,x,p)$ for each (t,x,p), so the derivative of the HJB equation with respect to (x,p) at the optimal trajectory $(s, X_s^{\pi^*}, p_s)$ gives

$$\nabla_{(x,p)}\left(\frac{\partial v}{\partial s} \;+\; \mathcal{L}_p^{\pi^*(s)}v \;+\; f\big(\cdot, \pi^*(s)\big)\right)\big(s, X_s^{\pi^*}, p_s\big) = 0.$$

Expanding this gradient and substituting P_s and Q_s we get

$$\begin{aligned} \frac{\partial}{\partial s}\big(\nabla_{(x,p)}v\big) \;+\; \nabla_{(x,p)}\Big(\mathcal{L}_p^{\pi^*(s)}v\Big) &= -\nabla_{(x,p)}f\big(s, X_s^{\pi^*}, p_s, \pi^*(s)\big) \\ &\quad -\; \nabla_{(x,p)}\Big(\widetilde{\alpha}^{\top}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, P_s\Big) \\ &\quad -\; \nabla_{(x,p)}\Big(\mathrm{Tr}\big(\tilde{\beta}\big(s, \Xi_s^{\pi^*}, \pi^*(s)\big)\, Q_s\big)\Big). \end{aligned}$$

Substituting the definition of Hamiltonian $\mathcal{H}$ back into the drift of dP_s we get

$$dP_s = -\nabla_{(x,p)}\mathcal{H}\big(s, \Xi_s^{\pi^*}, \pi^*(s), P_s, Q_s\big)\, ds + Q_s^{\top}\, d\widetilde{W}_s,$$

for $s \in [t,T]$, with terminal condition $P_T = \nabla_{(x,p)}g\big(X_T^{\pi^*}, p_T\big)$ obtained by differentiating the terminal condition $Y_T = g(X_T^{\pi^*}, p_T)$ with respect to (x,p) at time T. Finally, the results of steps (1)–(3), we conclude that the triple $\big(\Xi_s^{\pi^*}, Y_s, P_s\big)$ satisfies the coupled forward–backward system (4.7), with the given initial and terminal conditions. This concludes the proof. □

Theorem 4.1.1 establishes a clear pathway for converting a partially observable stochastic control (POSC) problem into a fully observable Markovian one. In the

POSC setting, the latent regime process ϵ_t is unobservable. By introducing its posterior filter p_t, we obtain an augmented state $\Xi_t^\pi = (X_t^\pi, p_t)$ that evolves on the probability simplex $\Delta(S)$. Under this augmented filtration, the value function becomes a function of (t, x, p), satisfies an HJB equation on the enlarged state space, and admits a forward–backward SDE representation driven by the innovation process W_t, mirroring the structure of the fully observed Markovian case. Conceptually, this demonstrates that certain non-Markovian problems can be transformed into Markovian ones by augmenting the state space with appropriate belief variables. In this sense, the non-Markovianity is "information-driven," arising from the mismatch between the true state (X_t, ϵ_t) and the observable filtration $\mathcal{G}_t$ and the augmentation with p_t exactly bridges this gap.

Having established this theorem and the context around it, we now examine how the partial-information problem compares to its full-information counterpart (when regimes are observable), both at the level of attainable primal value functions and their corresponding dual bounds.

Proposition 4.1.2 (Information monotonicity under partial observation)**.** Let $\mathbb{F} = (\mathcal{F}_t)_{t\in[0,T]}$ denote the *full-information* filtration under which the regime process $\epsilon_t \in S$ is observable, and let $\mathbb{G} = (\mathcal{G}_t)_{t\in[0,T]}$ denote the *observation* filtration generated by asset prices, with $\mathcal{G}_t \subseteq \mathcal{F}_t$ for all $t \in [0, T]$. Let $\mathcal{A}^{\mathbb{F}}$ and $\mathcal{A}^{\mathbb{G}}$ denote the corresponding sets of admissible controls, and assume the wealth dynamics are well defined for any admissible control.

Define the *full-information* value function conditional on the regime by

$$v^{\mathbb{F}}(t, x, \epsilon_i) := \sup_{\pi\in\mathcal{A}^{\mathbb{F}}} \mathbb{E}\big[U\big(X_T^{t,x,\pi}\big) \,\big|\, \epsilon_t = \epsilon_i\big], \qquad \epsilon_i \in S,$$

and the *partial-information* value function by

$$v^{\mathbb{G}}(t, x, p) := \sup_{\pi\in\mathcal{A}^{\mathbb{G}}} \mathbb{E}\big[U\big(X_T^{t,x,\pi}\big) \,\big|\, p_t = p\big], \qquad p \in \Delta(S).$$

Then, for all $(t, x, p) \in [0, T] \times \mathbb{R} \times \Delta(S)$,

$$v^{\mathbb{G}}(t, x, p) \;\le\; \sum_{i\in S} p^{(i)}\, v^{\mathbb{F}}(t, x, i). \tag{4.9}$$

Moreover, consider a weak dual formulation in which the primal value admits an upper bound of the form

$$v^{\mathbb{H}}(t, x, \cdot) \;\le\; v^{\mathbb{H}}_{\text{dual}}(t, x, \cdot) := \inf_{\eta\in\mathcal{D}^{\mathbb{H}}} J(t, x, \cdot; \eta), \qquad \mathbb{H} \in \{\mathbb{F}, \mathbb{G}\},$$

for some dual objective functional J and feasible sets $\mathcal{D}^{\mathbb{H}}$. Assume that dual feasibility is monotone in information, i.e. $\mathcal{D}^{\mathbb{G}} \subseteq \mathcal{D}^{\mathbb{F}}$ Then the corresponding dual values satisfy

$$v^{\mathbb{G}}_{\text{dual}}(t,x,p) \;\geq\; \sum_{i\in S} p^{(i)}\, v^{\mathbb{F}}_{\text{dual}}(t,x,i), \qquad (t,x,p) \in [0,T]\times\mathbb{R}\times\Delta(S). \quad (4.10)$$

Proof. Fix $(t,x,p) \in [0,T]\times\mathbb{R}\times\Delta(S)$. Since $\mathcal{G}_t \subseteq \mathcal{F}_t$, any $\mathbb{G}$-progressively measurable control is also $\mathbb{F}$-progressively measurable. It is easy to see that $\mathcal{A}^{\mathbb{G}} \subseteq \mathcal{A}^{\mathbb{F}}$.

Under full information, the controller observes the regime $\epsilon_t = \epsilon_i$ and may choose a control adapted to $\mathbb{F}$ conditional on this realization. Under partial information, the controller observes only the posterior belief $p_t = p$ and must select a single $\mathbb{G}$-adapted control that is feasible across all regimes consistent with p.

Therefore, for any $\mathbb{G}$-admissible control π,

$$\mathbb{E}\big[U\big(X_T^{t,x,\pi}\big)\,\big|\,p_t = p\big] = \sum_{i\in S} p^{(i)}\,\mathbb{E}\big[U\big(X_T^{t,x,\pi}\big)\,\big|\,\epsilon_t = i\big] \leq \sum_{i\in S} p^{(i)}\, v^{\mathbb{F}}(t,x,i).$$

Taking the supremum over $\pi \in \mathcal{A}^{\mathbb{G}}$ yields (4.9).

For the dual ordering, by assumption $\mathcal{D}^{\mathbb{G}} \subseteq \mathcal{D}^{\mathbb{F}}$. Taking an infimum over a smaller feasible set can only increase the value, and linearity in the initial distribution implies

$$v^{\mathbb{G}}_{\text{dual}}(t,x,p) = \inf_{\eta\in\mathcal{D}^{\mathbb{G}}} \mathcal{J}(t,x,p;\eta) \geq \inf_{\eta\in\mathcal{D}^{\mathbb{F}}} \sum_{i\in S} p^{(i)}\,\mathcal{J}(t,x,i;\eta) = \sum_{i\in S} p^{(i)}\, v^{\mathbb{F}}_{\text{dual}}(t,x,i),$$

which proves (4.10). □

Proposition 4.1.2 formalizes a fundamental information monotonicity principle in UMP under POSC. On the primal side, the inequality (4.9) says that restricting the controller to operate under the observation filtration $\mathbb{G}$ can only reduce the attainable expected utility relative to the full-information benchmark. Essentially, the relevant comparison is not to a single full-information value, but to the prior-weighted average of the regime-conditional full-information values. Economically, this tells us that a partially informed investor must commit to a single strategy that is feasible across all regimes consistent with the current belief p_t, whereas a fully informed investor can condition the control directly on the realized regime.

On the dual side, the ordering in (4.10) shows that information restrictions further weaken the corresponding dual bound. Since the set of admissible dual variables

shrinks as information is removed, the infimum defining the dual value increases, leading to a larger dual value under partial information. Together, these inequalities explain why partial observation induces both a lower primal value and a higher dual bound, and therefore leads to a wider duality gap as informational uncertainty increases.

This proposition highlights the role of information as a structural source of non-Markovianity, i.e, while the underlying dynamics remain Markovian under full observation after state-augmentation, the restriction to the observation filtration induces an effective path dependence through the evolution of beliefs. In the following sections, we move beyond this information-driven setting and consider UMP in which non-Markovianity arises intrinsically from the dynamics themselves, such as through long-memory or rough path-dependent coefficients that do not admit finite-dimensional state augmentations.

Remark 4.1.3. In theorem 4.1.1, we specifically focus on a class of UMP problems in which the drift is regime-dependent and unobservable, meanwhile the diffusion coefficient is constant, known, and independent of the latent regime. This assumption isolates the informational source of non-Markovianity namely, uncertainty about the unobserved regime without introducing additional filtering complexity through regime-dependent volatility. Under this restriction, Theorem 4.1.1 ensures that the innovation process is well defined and that the problem admits a Markovian reformulation via augmentation with the posterior filter.

From an application point of view, this modeling choice (to have constant diffusion but unknown drift) is natural. In financial markets, expected returns are often viewed as regime-dependent, while volatility can be treated as approximately constant and known over moderate horizons. Such cases have a dominant source of uncertainty stemming from regime misclassification rather than instantaneous volatility noise. Moreover, this framework is not just practically applicable specifically to finance. Similar structures arise in control and signal processing problems with latent operational modes, where system dynamics exhibit mode-dependent drift but are subject to stable, well-calibrated noise (diffusion). In these settings, uncertainty originates from hidden state transitions rather than through the diffusion term, and the control problem admits an analogous filtering-based reformulation.

We now turn to a fundamentally different source of non-Markovianity, where the dynamics themselves exhibit long memory or path dependence. In such UMP

problems, the coefficients at time t explicitly depend on the history $S_{[0,t]}$, and the problem cannot be reduced to a finite-dimensional Markovian formulation through state augmentation alone.

Path-dependent Coefficients: Solving UMP with Long-Memory Dynamics

We now turn our attention to a different form of non-Markovianity, one in which the dynamics themselves exhibit long memory. In this setting, the coefficients at time t depend not only on the current state but also on a portion of the past trajectory or, in some cases, on the entire history $S_{[0,t]}$. These problems capture many practically relevant features especially in finance, where volatility clustering, roughness, and persistent autocorrelation are well documented. However, the downside is that, unlike in the incomplete-information setting, there is in general no finite-dimensional state augmentation capable of restoring Markovian structure. This is because the non-Markovianity is dynamics-driven rather than information-driven, i.e, even when the regimes are fully observable, the coefficients themselves incorporate long memory through functionals of the past. As a result, the value function may naturally live in an infinite-dimensional path space, which significantly complicates both the solution and the analysis of such UMPs.

To formalize these problems, we again let the wealth process evolve under an observable price process S and a finite-state continuous-time Markov chain $(\epsilon_t)_{t\in[0,T]}$ with generator Q. For simplicity, we assume for now that the investor observes both the price and the regime. The purely path-dependent case is a special instance of this model, obtained by taking S to have only a single regime.

We now consider a setting in which, for each regime $\epsilon_i \in S$, the risky asset price is modeled as a functional stochastic differential equation (FSoDE). Again, for simplicity, we work with a single risky asset and a single Brownian motion, although the arguments extend naturally to the multi-dimensional case. We denote the price path up to time t by $S_{[0,t]} := \{S_s : 0 \le s \le t\}$, which we view as an element of the canonical space of continuous functions. Furthermore, we assume that the drift and diffusion coefficients for each regime ϵ_i are given by non-anticipative functionals

$$\mu_{\epsilon_i} : [0,T] \times C_t \to \mathbb{R}, \qquad \sigma_{\epsilon_i} : [0,T] \times C_t \to \mathbb{R},$$

where $C_t = C([0,t];\mathbb{R})$ denotes the space of continuous paths up to time t. In this framework, the asset dynamics in regime ϵ_i take the form

$$dS_t = S_t\, \mu_{\epsilon_i}\big(t, S_{[0,t]}\big)\, dt + S_t\, \sigma_{\epsilon_i}\big(t, S_{[0,t]}\big)\, dW_t, \qquad \text{when } \epsilon_t = \epsilon_i. \tag{4.11}$$

Subsequently, we can also write the wealth dynamics (using the same approach as in (2.5)) as

$$dX_t^\pi = \alpha\big(t, \epsilon_t, X_t^\pi, S_{[0,t]}, \pi_t\big)\, dt + \beta\big(t, \epsilon_t, X_t^\pi, S_{[0,t]}, \pi_t\big)^\top dW_t, \qquad X_0^\pi = x_0. \quad (4.12)$$

Remark 4.1.4. One analytically tractable and particularly useful representation that encompasses many empirically relevant models can be obtained by introducing, for each regime ϵ_i, a memory functional defined by

$$\Phi_{\epsilon_i}(t, S_{[0,t]}) := \int_0^t K_{\epsilon_i}(t-s)\, S_s\, ds, \quad (4.13)$$

where K_{ϵ_i} is a regime-dependent memory kernel. This kernel may be exponential (short memory), polynomial (long memory), or more general, including kernels designed to capture rough dynamics. Under this structure, the coefficients take the form

$$\mu_{\epsilon_i}(t, S_{[0,t]}) = \mu_{\epsilon_i}\big(t, S_t, \Phi_{\epsilon_i}(t, S_{[0,t]})\big), \qquad \sigma_{\epsilon_i}(t, S_{[0,t]}) = \sigma_{\epsilon_i}\big(t, S_t, \Phi_{\epsilon_i}(t, S_{[0,t]})\big),$$

and (4.11) becomes

$$dS_t = S_t\, \mu_{\epsilon_i}\big(t, S_t, \Phi_{\epsilon_i}(t, S_{[0,t]})\big)\, dt + S_t\, \sigma_{\epsilon_i}\big(t, S_t, \Phi_{\epsilon_i}(t, S_{[0,t]})\big)\, dW_t. \quad (4.14)$$

We emphasize that this construction defines one *class* of non-Markovian models, rather than an exhaustive characterization of all path-dependent dynamics. In fact, for certain choices of the kernel K_{ϵ_i}, the augmented state $(S_t, \Phi_{\epsilon_i}(t, S_{[0,t]}))$ admits a finite-dimensional Markovian representation. In such cases, a genuinely path-dependent HJB formulation is not required, and the problem can be treated using standard state-augmented dynamic programming techniques (like we did for hidden-regime case in the previous section).

Nonetheless, to *illustrate* the flexibility of the functional representation in (4.13), consider the case in which, for each regime ϵ_i, the coefficients depend only on the most recent five days of price history. This finite-memory specification is common in finance, as it captures moving-average momentum or short-term mean reversion. Such dependence can be modeled by choosing the regime-specific memory kernel

$$K_{\epsilon_i}(u) = \frac{1}{5}\, \mathbf{1}_{\{0 \le u \le 5\}},$$

which yields the memory functional

$$\Phi_{\epsilon_i}(t, S_{[0,t]}) = \frac{1}{5} \int_{t-5}^t S_s\, ds,$$

corresponding to the 5-day moving average of the price. This structure admits a natural economic interpretation. For example, in a bull regime (R0), the drift functional μ may place greater emphasis on recent upward trends as reflected in the moving average, whereas in a bear regime (R1), the drift may react more strongly to short-term declines or incorporate additional mean-reversion effects. In essence, the same path-dependent structure is interpreted through a regime-specific lens. Moreover, the model readily accommodates regime-specific kernels, for instance, a 5-day rolling average in the bull regime and a 7-day rolling average in the bear regime. We stress that this particular specification is included primarily as a benchmarking mechanism, i.e, while it falls within the broader path-dependent framework, it admits a Markovian reduction (can be solved by state-augmentation under Markovian settings) and therefore serves as a controlled example for validating the proposed path-dependent numerical methodology.

The primary motivation for introducing the path-dependent formulation lies in more general settings such as rough volatility models with Volterra-type kernels (e.g., fractional kernels of the form $K(u) = u^{H-\frac{1}{2}}$) and other infinite-memory or non-functional dependencies on $S_{[0,t]}$, which do not admit finite-dimensional state augmentations and therefore require a genuinely path-dependent treatment. These examples will be studied explicitly in subsequent sections.

Finally, we can now consider a terminal utility function $g : \mathbb{R} \times C_T \to \mathbb{R}$ and, for maximum generality, allow for a running utility (or cost) function (possibly regime-dependent)

$$f : [0,T] \times \mathbb{R} \times C_T \times K \to \mathbb{R},$$

Then the *path-dependent* value function can be written as

$$v(t,x,\epsilon_i,S_{[0,t]}) = \sup_{\pi\in\mathcal{A}} \mathbb{E}\Bigg[\int_t^T f\big(s, X_s^\pi, S_{[0,s]}, \pi_s\big)\, ds + g\big(X_T^\pi, S_{[0,T]}\big) \;\Big|\; X_t^\pi = x,\ \epsilon_t = \epsilon_i,\ S_{[0,t]} \Bigg] \tag{4.15}$$

Now, in order to characterize the optimality, we would need to an analytical setting that can handle path-dependence. For this purpose we employ the functional Itô calculus developed by Dupire (Dupire, 2009) and Cont–Fournié (Cont and Fournié, 2013). In that regard we define the following:

Definition 4.1.5. A scalar-valued functional

$$u : [0,T] \times C_T \to \mathbb{R}$$

is called *non-anticipative* if, for any $t \in [0,T]$ and any $\omega, \tilde{\omega} \in C_T$,

$$\omega(s) = \tilde{\omega}(s) \text{ for all } s \le t \quad \Longrightarrow \quad u(t,\omega) = u(t,\tilde{\omega}).$$

Thus $u(t,\omega)$ depends only on the path segment $\omega_{[0,t]}$.

Definition 4.1.6. For $t \in [0,T]$ and $\omega \in C_t$, define the perturbed path

$$\omega^h(s) := \begin{cases} \omega(s), & s < t, \\ \omega(t) + h, & s = t, \end{cases} \qquad h \in \mathbb{R}.$$

The *vertical derivative* of u at (t,ω) is

$$\frac{\partial u}{\partial \omega}(t,\omega) := \lim_{h \to 0} \frac{u(t,\omega^h) - u(t,\omega)}{h},$$

when the limit exists. The second vertical derivative is

$$\frac{\partial^2 u}{\partial \omega^2}(t,\omega) := \lim_{h \to 0} \frac{\frac{\partial u}{\partial \omega}(t,\omega^h) - \frac{\partial u}{\partial \omega}(t,\omega)}{h}.$$

Definition 4.1.7. For $\Delta > 0$ with $t + \Delta \le T$, define the flat extension

$$\omega^{t,\Delta}(s) := \begin{cases} \omega(s), & s \le t, \\ \omega(t), & t < s \le t + \Delta. \end{cases}$$

The *horizontal derivative* of u at (t,ω) is

$$\frac{\partial u}{\partial t}(t,\omega) := \lim_{\Delta \downarrow 0} \frac{u(t+\Delta, \omega^{t,\Delta}) - u(t,\omega)}{\Delta},$$

whenever the limit exists.

Definition 4.1.8. A non-anticipative functional u is of class $C^{1,2}$ in the functional sense if:

1. the horizontal derivative $\frac{\partial u}{\partial t}(t,\omega)$ exists and is continuous in (t,ω);
2. the first and second vertical derivatives $\frac{\partial u}{\partial \omega}(t,\omega)$ and $\frac{\partial^2 u}{\partial \omega^2}(t,\omega)$ exist and are continuous in (t,ω);
3. u and its derivatives have at most polynomial growth in ω.

We now recall the functional Itô formula, which generalizes the classical Itô formula to path-dependent functionals.

Theorem 4.1.9. Let S_t satisfy (4.11), and let u be a non-anticipative functional of class $C^{1,2}$ in the functional sense. Then

$$\begin{aligned} du(t, S_{[0,t]}) = \frac{\partial u}{\partial t}(t, S_{[0,t]})\, dt + \frac{\partial u}{\partial \omega}(t, S_{[0,t]})\, dS_t \\ + \frac{1}{2}\frac{\partial^2 u}{\partial \omega^2}(t, S_{[0,t]})\, \sigma^2_{\epsilon_t}(t, S_{[0,t]})\, dt. \end{aligned} \tag{4.16}$$

Proof. The proof can be found in the construction of Dupire's functional Itô calculus Dupire, 2009. The procedure approximates the path $S_{[0,t]}$ by piecewise-constant paths, applies the classical Itô formula to the finite-dimensional projections, and then shows that the error terms vanish in the limit. The horizontal derivative $\frac{\partial u}{\partial t}$ accounts for the effect of extending the path over time, while the vertical derivatives $\frac{\partial u}{\partial \omega}$ and $\frac{\partial^2 u}{\partial \omega^2}$ describe the sensitivity to instantaneous perturbations of the endpoint S_t. For details, see Dupire, 2009 and Cont and Fournié, 2013. □

Applying the multidimensional functional Itô formula to $Z_t := v_{\epsilon_t}\big(t, X^\pi_t, S_{[0,t]}\big)$ we can derive the HJB:

Theorem 4.1.10. Let S_t satisfy the functional SDE (4.11) and let us say that $v_i(t, x, S_{[0,t]}) := v\big(t, x, \epsilon_i, S_{[0,t]}\big)$ denote the value functions associated with the control problem. Furthermore, assume the following structural conditions (based on Peng, 1992, Pham, 2009, and Ekren, Touzi, and Zhang, 2014):

(A1) **Non-anticipativity:** $\mu_{\epsilon_i}(t, \omega)$, $\sigma_{\epsilon_i}(t, \omega)$, $\alpha(t, \epsilon_i, x, \omega, \pi)$, and $\beta(t, \epsilon_i, x, \omega, \pi)$ depend only on $\omega_{[0,t]}$.

(A2) **Path regularity:** $\mu_{\epsilon_i}(t, \cdot)$ and $\sigma_{\epsilon_i}(t, \cdot)$ (and similarly α, β) are continuous on C_t under the supremum norm.

(A3) **Lipschitz and linear growth:** μ_{ϵ_i}, σ_{ϵ_i}, α, β satisfy uniform Lipschitz and linear-growth bounds in (x, ω, π).

(A4) **Admissible controls:** K is compact, and admissible controls are progressively measurable and stable under concatenation.

(A5) **Well-posedness:** The state equation admits a unique strong solution and is stable under perturbations of paths and controls in the sense of functional Itô calculus.

Under assumptions (A1)–(A5) and additional regularity and integrability conditions ensuring the validity of the DPP and the applicability of the functional Itô formula, the DPP holds true (see Peng, 1992). Furthermore, provided that each v_i is sufficiently smooth (of class $C^{1,2}$ in the functional sense of Dupire) and satisfies appropriate growth conditions ensuring integrability of the stopped processes, the value functionals satisfy[1], for each $i \in S$,

$$\begin{aligned} 0 = \frac{\partial v_i}{\partial t}(t, x, S_{[0,t]}) + \sup_{\pi \in K} \Bigg\{ & \alpha(t, \epsilon_i, x, S_{[0,t]}, \pi)\, \frac{\partial v_i}{\partial x} + \frac{1}{2} \big\| \beta(t, \epsilon_i, x, S_{[0,t]}, \pi) \big\|^2 \frac{\partial^2 v_i}{\partial x^2} \\ & + \mu_{\epsilon_i}(t, S_{[0,t]})\, S_t\, \frac{\partial v_i}{\partial \omega} + \frac{1}{2} \sigma^2_{\epsilon_i}(t, S_{[0,t]})\, S_t^2\, \frac{\partial^2 v_i}{\partial \omega^2} \\ & + \sigma_{\epsilon_i}(t, S_{[0,t]})\, S_t\, \beta(t, \epsilon_i, x, S_{[0,t]}, \pi)\, \frac{\partial^2 v_i}{\partial x\, \partial \omega} + f_i(t, x, S_{[0,t]}, \pi) \Bigg\} \\ & + \sum_{j \neq i} q_{ij} \big(v_j(t, x, S_{[0,t]}) - v_i(t, x, S_{[0,t]}) \big), \end{aligned} \tag{4.17}$$

with terminal condition

$$v_i(T, x, S_{[0,T]}) = g\big(x, S_{[0,T]}\big). \tag{4.18}$$

Proof. The proof follows the same steps as the Markovian case, with the only difference that the sensitivity to the path enters through the vertical and horizontal derivatives of Dupire's functional calculus. Again, fix $(t, x, S_{[0,t]})$ and an admissible control π. Let $Z_s := v_{\epsilon_s}(s, X_s^\pi, S_{[0,s]})$. Applying the multidimensional functional Itô formula in the variables (x, ω) yields, up to the exit time τ_n,

$$\begin{aligned} dZ_s = \Big[& \frac{\partial v_i}{\partial s} + \alpha\, \frac{\partial v_i}{\partial x} + \mu_{\epsilon_i} S_s\, \frac{\partial v_i}{\partial \omega} + \frac{1}{2} \|\beta\|^2\, \frac{\partial^2 v_i}{\partial x^2} + \frac{1}{2} \sigma^2_{\epsilon_i} S_s^2\, \frac{\partial^2 v_i}{\partial \omega^2} \\ & + \sigma_{\epsilon_i} S_s\, \beta\, \frac{\partial^2 v_i}{\partial x\, \partial \omega} \Big] ds + dM_s, \end{aligned}$$

where M_s is a local martingale with zero expectation. Integrating from t to τ_n, taking expectations gives the usual subsolution inequality. Invoking the dynamic programming principle under assumptions (A1)–(A5), and using the regularity and growth conditions on v_i to justify localization and dominated convergence, this implies

$$\frac{\partial v_i}{\partial t} + \sup_{\pi \in K} \{ \cdots \} \leq 0.$$

[1] In the absence of sufficient regularity, the equation should be interpreted in the viscosity sense within the framework of path-dependent PDEs (Ekren, Touzi, and Zhang, 2014; Ekren, Touzi, and Zhang, 2016).

If an optimal feedback control exists and the supremum is attained, the inequality becomes an equality, resulting inthe HJB equation (4.17). □

The HJB system (4.17) is a coupled nonlinear system of partial differential equations on path space. The coupling through the generator Q is identical to the Markovian regime-switching case. However, the presence of functional derivatives makes the effective state infinite-dimensional. In particular, in general there is no finite-dimensional summary statistic of $S_{[0,t]}$ that preserves optimality (unless the model admits a finite-dimensional state augmentation).

We now derive a backward SDE representation for the value functional along an optimal trajectory in the path-dependent regime-switching setting. This provides the path-dependent analogue of the Markovian "value BSDE" representation and is fully consistent with Theorem 4.1.10.

Theorem 4.1.11. Assume the conditions of Theorem 4.1.10 hold, and let π^* be an optimal control. For $s \in [t, T]$ define

$$Y_s := v_{\epsilon_s}\big(s, X_s^{\pi^*}, S_{[0,s]}\big),$$

and

$$Z_s := \frac{\partial v_{\epsilon_s}}{\partial x}\big(s, X_s^{\pi^*}, S_{[0,s]}\big)\,\beta\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big) + \frac{\partial v_{\epsilon_s}}{\partial \omega}\big(s, X_s^{\pi^*}, S_{[0,s]}\big)\,\sigma_{\epsilon_s}\big(s, S_{[0,s]}\big).$$

Define the purely discontinuous local martingale

$$dM^{\epsilon}(s) := \sum_{i \neq j} \Big(v_j\big(s, X_s^{\pi^*}, S_{[0,s]}\big) - v_i\big(s, X_s^{\pi^*}, S_{[0,s]}\big)\Big)\, dM_{ij}(s),$$

where M_{ij} denotes the compensated jump martingale associated with transitions from regime i to regime j. Then (Y_s, Z_s) satisfies the regime-switching BSDE

$$dY_s = -f\big(s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)\, ds + Z_s\, dW_s + dM^{\epsilon}(s), \qquad Y_T = g\big(X_T^{\pi^*}, S_{[0,T]}\big). \tag{4.19}$$

Proof. Along the optimal trajectory define

$$Y_s = v_{\epsilon_s}\big(s, X_s^{\pi^*}, S_{[0,s]}\big).$$

Applying the multidimensional functional Itô formula to the map $(x, \omega) \mapsto v_i(s, x, \omega)$ on each regime i, and accounting for the jumps of the Markov chain $(\epsilon_s)_{s\geq 0}$, yields

$$\begin{aligned} dY_s = \Bigg[& \frac{\partial v_{\epsilon_s}}{\partial s} + \alpha\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big) \frac{\partial v_{\epsilon_s}}{\partial x} + \mu_{\epsilon_s}\big(s, S_{[0,s]}\big)\, S_s \frac{\partial v_{\epsilon_s}}{\partial \omega} \\ & + \frac{1}{2}\Big\|\beta\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)\Big\|^2 \frac{\partial^2 v_{\epsilon_s}}{\partial x^2} + \frac{1}{2}\sigma_{\epsilon_s}^2\big(s, S_{[0,s]}\big)\, S_s^2 \frac{\partial^2 v_{\epsilon_s}}{\partial \omega^2} \\ & + \sigma_{\epsilon_s}\big(s, S_{[0,s]}\big)\, S_s\, \beta\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big) \frac{\partial^2 v_{\epsilon_s}}{\partial x\, \partial \omega} \\ & + \sum_{j \neq \epsilon_s} q_{\epsilon_s j} \Big(v_j\big(s, X_s^{\pi^*}, S_{[0,s]}\big) - v_{\epsilon_s}\big(s, X_s^{\pi^*}, S_{[0,s]}\big) \Big) \Bigg] ds \\ & + Z_s\, dW_s + dM^{\epsilon}(s), \end{aligned}$$

where

$$dM^{\epsilon}(s) = \sum_{i \neq j} \Big(v_j\big(s, X_s^{\pi^*}, S_{[0,s]}\big) - v_i\big(s, X_s^{\pi^*}, S_{[0,s]}\big) \Big)\, dM_{ij}(s)$$

is the purely discontinuous local martingale term.

By optimality of π^*, the functional HJB equation (4.17) holds with equality alongikm along $(X_s^{\pi^*}, S_{[0,s]})$ for the active regime ϵ_s. Substituting this equality eliminates all derivative and generator terms and yields

$$dY_s = -f\big(s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)\, ds + Z_s\, dW_s + dM^{\epsilon}(s).$$

At T we have

$$Y_T = v_{\epsilon_T}\big(T, X_T^{\pi^*}, S_{[0,T]}\big) = g\big(X_T^{\pi^*}, S_{[0,T]}\big),$$

which matches the terminal condition of (4.19). □

The BSDE in Theorem 4.1.11 provides the value process (Y, Z) along the optimal state trajectory. In order to recover the full first-order optimality structure and to mirror the unified HJB/BSDE/adjoint characterization that we had for Markovian case in the previous chapter, we now differentiate the value function with respect to the state variable x. This will yield a first-order adjoint (backward) equation with an additional jump martingale term reflecting regime switches as can be seen in the theorem.

Theorem 4.1.12. Assume the conditions of Theorem 4.1.10 hold and let π^* be an optimal control. For $s \in [t,T]$, define

$$\begin{aligned} P_s &:= \frac{\partial v_{\epsilon_s}}{\partial x}\Big(s, X_s^{\pi^*}, S_{[0,s]}\Big), \\ \widehat{Q}_s &:= \frac{\partial^2 v_{\epsilon_s}}{\partial x^2}\Big(s, X_s^{\pi^*}, S_{[0,s]}\Big)\,\beta\Big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\Big) \\ &\quad + \frac{\partial^2 v_{\epsilon_s}}{\partial x\,\partial \omega}\Big(s, X_s^{\pi^*}, S_{[0,s]}\Big)\,\sigma_{\epsilon_s}\Big(s, S_{[0,s]}\Big). \end{aligned}$$

Assume the integrability conditions

$$P \in S^2(t,T;\mathbb{R}), \qquad \widehat{Q} \in H^2(t,T;\mathbb{R}^m).$$

Define the purely discontinuous local martingale

$$dM^{\epsilon,\nabla}(s) := \sum_{i\neq j}\left(\frac{\partial v_j}{\partial x}\big(s, X_s^{\pi^*}, S_{[0,s]}\big) - \frac{\partial v_i}{\partial x}\big(s, X_s^{\pi^*}, S_{[0,s]}\big)\right) dM_{ij}(s),$$

where M_{ij} denotes the compensated jump martingale associated with transitions from regime i to regime j.

Then the pair $(P_s, \widehat{Q}_s)$ satisfies the backward adjoint equation

$$\begin{aligned} dP_s &= -\nabla_x \mathcal{H}\big(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s), P_s, \widehat{Q}_s\big)\,ds + \widehat{Q}_s^{\top}\,dW_s + dM^{\epsilon,\nabla}(s), \\ P_T &= \frac{\partial g}{\partial x}\big(X_T^{\pi^*}, S_{[0,T]}\big) \end{aligned} \tag{4.20}$$

Proof. Fix $(t, x, S_{[0,t]})$ and assume that π^* be an optimal control. Throughout the proof, we use the definitions of P_s, $\widehat{Q}_s$, and $dM^{\epsilon,\nabla}(s)$ given in the statement of the theorem. Since each v_i is of class $C^{1,2}$ in the functional sense and twice continuously differentiable in the state variable x, the non-anticipative functional

$$\zeta_i(s,x,\omega) := \frac{\partial v_i}{\partial x}(s,x,\omega), \qquad i \in S,$$

admits the multidimensional functional Itô formula. Applying this formula to the process $s \longmapsto \zeta_{\epsilon_s}\big(s, X_s^{\pi^*}, S_{[0,s]}\big)$ and accounting for the jumps of the regime process $(\epsilon_s)_{s\geq 0}$ via the compensated martingales M_{ij}, yields, up to the localization time $\tau_n := \inf\{u \geq t : |X_u^{\pi^*}| \geq n\} \wedge T$ the semimartingale decomposition

$$dP_s = \mathcal{D}_s\,ds + \widehat{Q}_s^{\top}\,dW_s + dM^{\epsilon,\nabla}(s), \qquad s \in [t, \tau_n], \tag{4.21}$$

where $\mathcal{D}_s$ denotes the drift term generated by the functional Itô expansion, evaluated along $\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)$. Along the optimal trajectory, the functional HJB

equation (4.17) holds with equality for the active regime ϵ_s and control π_s^*. Differentiating this equality with respect to the state variable x, and using the product rule together with the fact that μ_{ϵ_s} and σ_{ϵ_s} do not depend on x, yields

$$\begin{aligned} 0 = \frac{\partial}{\partial s}P_s + \alpha\,\frac{\partial}{\partial x}P_s + \frac{\partial \alpha}{\partial x}\,P_s + \frac{1}{2}\|\beta\|^2\,\frac{\partial^2}{\partial x^2}P_s + \left(\frac{\partial \beta}{\partial x}\right)^{\top}\widehat{Q}_s \\ + \mu_{\epsilon_s}S_s\,\frac{\partial}{\partial \omega}P_s + \frac{1}{2}\sigma_{\epsilon_s}^2 S_s^2\,\frac{\partial^2}{\partial \omega^2}P_s + \sigma_{\epsilon_s}S_s\,\frac{\partial^2}{\partial x\,\partial \omega}P_s \\ + \frac{\partial f}{\partial x}\big(s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big) + \sum_{j\neq\epsilon_s} q_{\epsilon_s j}\left(\frac{\partial v_j}{\partial x} - \frac{\partial v_{\epsilon_s}}{\partial x}\right), \end{aligned} \tag{4.22}$$

where the last bracket is evaluated at $\big(s, X_s^{\pi^*}, S_{[0,s]}\big)$. Comparing (4.22) with the drift term $\mathcal{D}_s$ appearing in the Itô expansion (4.21), and rearranging terms, yields

$$\mathcal{D}_s = -\nabla_x \mathcal{H}\big(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s), P_s, \widehat{Q}_s\big)$$

Substituting this expression for $\mathcal{D}_s$ into (4.21) yields the adjoint backward SDE (4.20) on $[t, \tau_n]$. Finally, by the terminal condition (4.18),

$$P_T = \frac{\partial v_{\epsilon_T}}{\partial x}\big(T, X_T^{\pi^*}, S_{[0,T]}\big) = \frac{\partial g}{\partial x}\big(X_T^{\pi^*}, S_{[0,T]}\big).$$

The integrability assumptions $P \in S^2(t,T;\mathbb{R})$ and $\widehat{Q} \in H^2(t,T;\mathbb{R}^m)$ allow removal of the localization by letting $n \to \infty$ (so that $\tau_n \uparrow T$ almost surely), which completes the proof. □

The three theorems above provide apt characterizations of optimality in the path-dependent regime-switching control problem. The functional HJB equation in Theorem 4.1.10 characterizes the value functional on path space, Theorem 4.1.11 provides a backward stochastic representation of the value process along an optimal trajectory, and Theorem 4.1.12 presents the corresponding first-order adjoint processes through a fully coupled backward SDE. The following proposition consolidates these results into a single path-dependent regime-switching FBSDE system, providing the natural parallel of the unified Markovian HJB/BSDE/adjoint characterization.

Proposition 4.1.13. Assume the standing conditions of Theorem 4.1.10 and let π^* be an optimal control (so that the functional HJB (4.17) holds with equality along the optimal trajectory). Let $(X^{\pi^*}, \epsilon_s, S)$ denote the corresponding controlled state/regime/path processes, with terminal payoff g as in (4.18).

Define the *value process* and its *martingale integrand* by (4.19), and define the *first-order adjoint* processes $(P, \widehat{Q})$ and the jump-martingale $M^{\epsilon,\nabla}$ by Theorem 4.1.12. In addition, define the regime-switching jump martingale M^{ϵ} as in Theorem 4.1.11.

Then the quadruple $(X_s^{\pi^*}, Y_s, P_s)$ together with the martingale terms $(Z_s, \widehat{Q}_s, M^{\epsilon}, M^{\epsilon,\nabla})$ satisfies, on $[t, T]$, the coupled path-dependent regime-switching forward–backward system

$$\begin{aligned} dX_s^{\pi^*} &= \alpha\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)\, ds + \beta\big(s, \epsilon_s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)^{\top} dW_s, \\ dY_s &= -\, f\big(s, X_s^{\pi^*}, S_{[0,s]}, \pi_s^*\big)\, ds + Z_s\, dW_s + dM^{\epsilon}(s), \\ dP_s &= -\nabla_x \mathcal{H}\big(s, \epsilon_s, X_s^{\pi^*}, \pi^*(s), P_s, \widehat{Q}_s\big)\, ds + \widehat{Q}_s^{\top}\, dW_s + dM^{\epsilon,\nabla}(s) \end{aligned} \tag{4.23}$$

with boundary conditions

$$Y_T = g\big(X_T^{\pi^*}, S_{[0,T]}\big), \qquad P_T = \frac{\partial g}{\partial x}\big(X_T^{\pi^*}, S_{[0,T]}\big), \qquad X_t^{\pi^*} = x. \tag{4.24}$$

Moreover, along the optimal trajectory the processes admit the (path-dependent) decoupling relations

$$Y_s = v_{\epsilon_s}\big(s, X_s^{\pi^*}, S_{[0,s]}\big), \qquad P_s = \frac{\partial v_{\epsilon_s}}{\partial x}\big(s, X_s^{\pi^*}, S_{[0,s]}\big), \tag{4.25}$$

and Z_s and $\widehat{Q}_s$ are given by the explicit representations in Theorems 4.1.11 and 4.1.12, respectively.

Proposition 4.1.13 is the path-dependent regime-switching analogue of the unified Markovian characterization where the functional HJB system (4.17) encodes optimality on path space, while Theorems 4.1.11 and 4.1.12 provide the fully adjoint BSDE structure satisfied along an optimal trajectory, together with the decoupling relations (4.25). Furthermore, the proposition also makes transparent the principal computational obstruction present in these classes of problems. The forward dynamics and the decoupling fields depend on the entire history $S_{[0,t]}$, so the effective state is inherently infinite-dimensional. In contrast to the hidden-regime setting, where partial information can often be restored to a finite-dimensional state through filtering techniques, here there exists no finite collection of summary statistics capable of capturing the full path dependence without loss of optimality. Therefore, state-of-art classical grid-based PDE solvers are therefore inapplicable, and standard Deep BSDE architectures that were designed for finite-dimensional Markovian states (as in the previous chapter) no longer fit the structure of (4.23).

The characterization in Proposition 4.1.13 highlights that the primary computational issue in genuinely path-dependent regime-switching UMPs stems from the infinite-dimensional nature of the effective state. While the functional HJB system (4.17) and the associated FBSDE representation provide a complete theoretical pathway for optimality, their explicit numerical implementation is not strictly necessary in order to recover optimal controls. An alternative approach would be to treat the UMP and its associated dual representation directly as stochastic optimization problems over a class of admissible, non-anticipative controls. In this *policy-based formulation*, one would parametrize the control process π_t itself as a function of the observed path history and optimizes either the expected utility functional (primal) or the corresponding dual objective, without explicitly enforcing the functional HJB equation during training. The functional HJB and BSDE structures derived above would then in that case serve as *certificates of optimality* and consistency relations that can characterize the optimizer at convergence, rather than as using them as hard constraints imposed during the numerical optimization and learning.

However, equation (4.25) presents a natural procedure to numerical approximation where one can learn the path-to-value and path-to-control maps directly from simulated trajectories. Any viable algorithm must ingest variable-length histories and learn which time segments are relevant for predicting $v_{\epsilon_t}(t, X_t, S_{[0,t]})$ and π_t^*. Transformer architectures are particularly well suited to this setting, since self-attention provides an adaptive mechanism for selecting and combining information across the entire past, capturing both short and long-range temporal dependencies as well as complex temporal cross-talk.

In essence, when applied to simulated trajectories $(S_{t_k})_{k\leq n}$, a transformer-based solver can approximate the mappings

$$(S_{[0,t]}, X_t, \epsilon_t) \longmapsto Q_{\epsilon_t}(t, X_t, S_{[0,t]}), \qquad (S_{[0,t]}, X_t, \epsilon_t) \longmapsto \pi_t^*,$$

and similarly the BSDE components (Y_t, Z_t) in (4.19), without requiring state augmentation or Markovian embeddings. To the best of our knowledge, numerical solution methods for this class of path-dependent regime-switching control problems have not been developed in the existing stochastic control literature.

In addition to the *value-based formulation* by the functional HJB, transformers also open the path to consider a primal-dual numerical approach that directly approximates the original UMP or its associated dual representation. Essentially, if $\{\pi^\theta\}_{\theta\in\Theta}$ denotes a parametric family of admissible controls, where each control is

non-anticipative function of the observed trajectory $\pi_t^\theta = \Pi_\theta(\mathcal{T}_t)$ then in the case of the primal problem, the numerical objective would be to maximize the expected utility

$$J(\theta) = \mathbb{E}\left[\int_0^T f\left(t, X_t^{\pi^\theta}, S_{[0,t]}, \pi_t^\theta\right) dt + g\left(X_T^{\pi^\theta}, S_{[0,T]}\right)\right] \tag{4.26}$$

subject to the path-dependent controlled dynamics (4.12). At the optimum parameter θ^*, the resulting control π^{θ^*} would satisfy the functional HJB (equation (4.17)) in an implicit sense. In the dual formulation, the same non-anticipative parametrization is used but here the expected utility function would be replaced by the dual objective that was introduced in the previous chapter. We will now move onto the next section of this chapter where we will start developing the transformer-based numerical algorithms for path-dependent regime-switching UMPs.

4.2 Transformer-Based Algorithms for Path-dependent Regime-Switching UMP

Building on the motivation and transformer characterization from the section above, the central and most critical numerical object of interest in the path-dependent regime-switching problem is the set of backward processes which are expressed deterministic functions of the current state and the history (decoupling fields) associated with the FBSDE system illustrated in Proposition 4.1.13. Along any trajectory which is deemed to be admissible, these fields determine the value process, its martingale representation, and the optimal control through the relations

$$Y_t = v_{\epsilon_t}(t, X_t, S_{[0,t]}), \quad P_t = \frac{\partial v_{\epsilon_t}}{\partial x}(t, X_t, S_{[0,t]}), \quad \pi_t^* = \arg\max_{\pi \in K} \mathcal{H}(t, \epsilon_t, X_t, S_{[0,t]}, \pi, \cdot)$$

Now in the path-dependent case (where no state-augmentation settings are possible), these aforementioned quantities cannot be parameterized as functions of a fixed-dimensional state evaluated pointwise in time. Instead, they must be understood as non-anticipative functionals of the entire observed history up to the current time. From an algorithmic standpoint, this means that the numerical problem is no longer to approximate a family of functions indexed by time, but rather to learn a mapping from partial trajectories to instantaneous decisions and value-related quantities. At each discretization time t_i, the solver observes the growing sequence $\left\{(S_{t_k}, X_{t_k}, \epsilon_{t_k})\right\}_{k \le i}$ and must produce estimates of the optimal control $\pi_{t_i}^*$ and the associated BSDE components Y_{t_i} and Z_{t_i}. The relevance of past observations in this case is neither uniform nor known *a priori*. Depending on the underlying dynamics, optimal decisions may depend predominantly on recent history (as in finite-memory

moving-average models) or on long-range temporal structure (as in the case of rough volatility and other Volterra-type models).

This reframes the numerical solution of path-dependent UMP as a sequence learning task with strict causality constraints. An algorithmic solution must process variable-length histories, respect the temporal ordering of information, and adaptively combine information across time in a non-anticipative manner. Any viable approximation must be able to represent nonlocal temporal interactions without relying on handcrafted memory features or problem-specific state augmentations. In that regard, we spend the remainder of this section developing a numerical framework that addresses these particular requirements directly. Rather than trying to compress path information into a finite-dimensional proxy, we parameterize these decoupling fields and the associated control policies as functions of discretized trajectories.

At this point, it is useful to differentiate between the two complementary numerical approaches (using transformers) to solving the path-dependent regime-switching UMP (simplified version expressed in table 4.1). The first approach, the *value-based formulation*, is rooted directly in the functional HJB equation (4.17) and the associated FBSDE system in Proposition 4.1.13. In this approach, the numerical algorithm aims to approximate the value function and its associated derivative on path space, including its martingale representations and adjoint processes. The optimal control is then recovered implicitly through the Hamiltonian maximization condition encoded in the HJB system. This approach generalizes classical dynamic programming and Deep BSDE method, but extended to the genuinely path-dependent regime-switching setting via functional Itô calculus.

The second approach, the *policy-based primal–dual formulation*, targets the original UMP or its associated dual representation directly. Here, the control process is parameterized as a non-anticipative functional of the observed trajectory, and the numerical objective is to optimize either the expected utility functional (4.26) (primal) or the corresponding dual objective introduced in the previous chapter. In this formulation, neither the functional HJB equation nor the BSDE system is explicitly enforced during training. Instead, these analytical objects serve as "checks" that characterize the optimizer at convergence, either through first-order optimality conditions in the primal problem or through saddle-point conditions in the dual formulation.

Both these approaches are theoretically equivalent in the sense that they admit the

same optimal control under the standing assumptions of Theorem 4.1.10. However, they differ substantially in their numerical behavior, stability properties, and suitability for high-dimensional or infinite-dimensional path-dependent settings. In particular, the policy-based primal–dual formulation avoids the explicit approximation of path-dependent value functionals and their derivatives, making it especially well suited for transformer-based architectures operating directly on trajectory prefixes. These approaches allows similar algorithmic structure to be applied across a variety of path-dependent UMPs, including finite-memory benchmarks and genuinely long-memory dynamics, and provides a unified basis for both the path-dependent primal and dual Deep BSDE formulations.

Table 4.1: Comparison of numerical formulations for path-dependent UMPs

	Value-based (Algorithm 6)	**Policy-based (Algorithm 7)**
Primary target	Value functional v and decoupling fields	Control π (path-dependent)
Optimization objective	HJB / BSDE consistency	Primal utility or dual objective
HJB enforcement	Explicit (during training)	Implicit at optimum
Role of BSDE	Central	Auxiliary / stabilizing
Numerical paradigm	Dynamic programming–based	Direct stochastic optimization
Sensitivity to path dimension	High	Lower

Sequence Representation of Paths and Model Inputs

In order to ensure that the approximation of path-dependent decoupling fields are computationally tractable, we specify how continuous-time trajectories are discretized and fed into to the learning algorithm. Throughout, we work on the already introduced fixed time grid $0 = t_0 < t_1 < \cdots < t_N = T$ and consider discrete-time approximations of the forward processes generated by the controlled dynamics. At any given time index i which is representing time t_i, the transformer has access to the partial trajectory $\{(S_{t_k}, X_{t_k}, \epsilon_{t_k})\}_{k=0}^{i}$ which is treated as a finite sequence whose length increases with time. Each element of the input sequence corresponds to the system state at a single discretized point and serves as a basic informational unit for the numerical solver.

Furthermore, we represent each time step t_k by a token that encodes the contempora-

neous values of the asset price S_{t_k}, the wealth process X_{t_k}, and the regime indicator ϵ_{t_k}. Additional features, such as normalized time indices or increments, may be included when beneficial for numerical stability, but the essential structure is that the input at time t_i consists of the ordered collection of tokens up to that time. This representation allows us to accommodate a wide range of path-dependent structures without any modification. For example, in finite-memory models, such as those induced by moving-average kernels, only the most recent subset of tokens influences the coefficients and optimal controls. On the other hand, in models driven by rough or Volterra-type kernels, relevant information may be distributed across the entire history, with weights that decay slowly or irregularly over time. Essentially, the sequence representation does not impose any prior assumptions on which portions of the path are relevant and this structure is learned implicitly by the model.

A key requirement of this representation is causality. At any time t_i, the solver must base its predictions solely on information available up to that time (no forward looking biases). Accordingly, the ordering of tokens in the sequence must be preserved, and the numerical architecture must prevent the use of future observations when producing current outputs. This constraint is enforced at the architectural level and is vital to the design of the transformer-based learning algorithm developed in this section. Now, the central algorithmic challenge after we obtain a unified input format is to construct a non-anticipative map from partial trajectories to instantaneous value-related quantities (value itself and value derivatives) and control, without relying on state-augmentation or finite-dimensional approximations. In essence, at each time t_i, the solver observing a growing trajectory prefix $\mathcal{T}_i = \left\{(S_{t_k}, X_{t_k}, \epsilon_{t_k}, t_k)\right\}_{k=0}^{i}$ must produce estimates of the path-dependent decoupling fields evaluated at the current time, namely

$$Y_{t_i} = v_{\epsilon_{t_i}}(t_i, X_{t_i}, S_{[0,t_i]}), \qquad Z_{t_i}, \qquad \pi^*_{t_i}.$$

Now, in contrast to the Markovian setting, these aforementioned quantities cannot be represented as evaluations of time-indexed functions acting on a fixed-state. Instead, they must be interpreted as non-anticipative functions of the entire observed history. From a numerical perspective, this problem is now a sequence learning task with strict causality constraints. So, the final output expected is no longer a family of functions indexed by time, but rather a mapping from variable-length sequences to instantaneous outputs.

Architectural composition of the Transformer-network

Transformer architectures here operates directly on sequences and provide a flexible mechanism for aggregating information across time. At the same time, the model also respects temporal causality and adaptively determines which portion of the past trajectory are relevant in the decision making. For this, each observation $(S_{t_k}, X_{t_k}, \epsilon_{t_k}, t_k)$ is first embedded into a fixed-dimensional vector space, producing a sequence of latent representations $h_0, h_1, \ldots, h_i \in \mathbb{R}^{d_{\text{model}}}$. These embeddings place heterogeneous quantities into a common representation space and scale further enabling shared processing across time. In our architecture, temporal structure is encoded explicitly, either through positional embeddings or through the inclusion of time indices as input features (the latter is easier to implement).

At a broad level, the transformer consists of multiple stacked layers, each combining a self-attention operation with feed-forward transformations, residual connections, and normalization. The output of the final layer is a sequence of latent vectors, from which the representation corresponding to the current time t_i is extracted and mapped to the desired outputs via task-specific readout heads. At the heart of these stacked layers is the self-attention mechanism. Essentially, for each token representation h_k, linear projections produce a query, key, and value:

$$q_k = W_Q h_k, \qquad k_k = W_K h_k, \qquad v_k = W_V h_k.$$

where at every time t_i, the attention output is given by

$$\widetilde{h}_i = \sum_{k=0}^{i} \alpha_{ik}\, v_k, \qquad \alpha_{ik} = \frac{\exp\Big(\langle q_i, k_k\rangle/\sqrt{d}\Big)}{\sum_{\ell=0}^{i} \exp\Big(\langle q_i, k_\ell\rangle/\sqrt{d}\Big)},$$

where the restriction $k \leq i$ enforces causality and non-anticipativity. The attention weights $\{\alpha_{ik}\}$ form a data-dependent probability distribution over past time indices and admit a natural interpretation in the context of stochastic control. They quantify the relevance of each past state $(S_{t_k}, X_{t_k}, \epsilon_{t_k})$ for determining the current value or control, conditional on the present context. Unlike fixed-memory constructions or predefined kernel-based summaries, this relevance is learned endogenously and adapts dynamically across trajectories and regimes. The learning via self-attention mechanism is implemented using multiple attention heads, each of which have its own set of projection matrices. Having multiple heads allows us to use different projections matrices to specifically learn distinct temporal patterns and interactions. Certain heads are allotted to learn short term fluctuations, some for longer term

fluctuations and other on regime-switching effects. The output of all these heads are concatenated and linearly transformed, outputting a rich aggregate representation. In addition to the multiple heads, the architecture also has stacked transformer layers, which induces a hierarchical temporal structure. The lower layers capture local interactions (mostly short-range interactions), meanwhile the deeper layers integrate information across longer horizons. This hierarchical aggregation reflects the multi-scale nature of path-dependent UMP models, and provides a natural mechanism for learning complex temporal dependencies.

In the subsequent subsections, we specify how this architecture is integrated into the primal and dual Deep BSDE algorithms, discuss the resulting algorithmic modifications, and introduce results that provide insight into how the transformer internally represents and utilizes path-dependent information.

Transformer-Based Primal and Dual Deep BSDE Algorithms

The goal of this subsection is to adapt the primal and dual Deep BSDE frameworks developed in the Markovian setting to the path-dependent case, where the decoupling fields and optimal controls depend on the entire observed history. As established previously, this requires replacing time-indexed, state-based neural networks with sequence models that operate directly on discretized paths. In classical Deep BSDE solvers for Markovian control problems, the unknown quantities appearing in the discretized FBSDE systems are parameterized by families of functions of the form $(t_i, X_{t_i}, \epsilon_{t_i}) \longmapsto (\pi_{t_i}, Z_{t_i}, Q_{t_i})$ implemented via separate neural networks at each time step. This approach relies fundamentally on the existence of a finite-dimensional state that fully summarizes the past. In the path-dependent regime-switching setting, such a representation is no longer feasible. Instead, at each time t_i, the relevant variables must be parameterized as functionals of the trajectory prefix $\mathcal{T}_i = \left\{(S_{t_k}, X_{t_k}, \epsilon_{t_k})\right\}_{k \le i}$. Accordingly, we replace the collection of time-indexed networks $\{\mathcal{N}_{\theta_{i,\cdot}}\}_{i=0}^{N-1}$ by a single transformer model $\mathcal{T}_\theta$ that maps trajectory prefixes to instantaneous outputs. In such a formulation, the dependence on time is encoded implicitly through the sequence ordering (and, when used, positional encodings), rather than explicitly passing t_i as an input. Operationally, this modification preserves the structure of the primal and dual value-based algorithms presented earlier, while changing only the parameterization step by which (π_i, Z_i, Q_i) are produced.

This transformer $\mathcal{T}_\theta$ produces a sequence of latent representations, one for each

Algorithm 4 Task-specific readouts from the final-token latent

Require: Final latent $\mathbf{h}_i \in \mathbb{R}^{d_{\text{model}}}$ at time t_i.

$\pi_{t_i} \leftarrow \text{Head}_\pi(\mathbf{h}_i)$ ▷ control output (apply constraints if needed)
$Z_{t_i} \leftarrow \text{Head}_Z(\mathbf{h}_i)$ ▷ BSDE martingale integrand
$Q_{t_i} \leftarrow \text{Head}_Q(\mathbf{h}_i)$ ▷ second-order or auxiliary term, if used
return $(\pi_{t_i}, Z_{t_i}, Q_{t_i})$

time step in the observed trajectory prefix. At each time t_i, the produced latent representation corresponding to the most recent token (i.e., the final token in the prefix) is passed through task-specific readout heads to obtain the quantities required by the discretized BSDE updates. Concretely, we use a shared backbone together with distinct readouts $\text{Head}_\pi, \text{Head}_Z, \text{Head}_Q$, so that a single forward pass yields $(\pi_{t_i}, Z_{t_i}, Q_{t_i})$ at the current time.

In the *primal* value-based formulation, these outputs parameterize the optimal control π_{t_i} and the BSDE martingale integrand Z_{t_i} (and, when needed, auxiliary quantities such as second-order terms Q_{t_i} in the adjoint formulation). These variables enter the discretized FBSDE system exactly as in the Markovian algorithm, with the sole difference that evaluation now depends on the full prefix $\mathcal{T}_i$ rather than only $(t_i, X_{t_i}, \epsilon_{t_i})$. On the other hand, in the *dual* value-based formulation, the same transformer backbone is used, but with loss functions and readouts appropriate to the dual variables and minimization objective (the ones that have been discussed in the chapter before). This shared representation is especially natural in the present setting because both primal and dual problems require learning path-to-decision maps, and the architectural burden is concentrated in the ability to represent non-anticipative dependence on histories. The task-specific readouts from the final-token latent are explicitly summarized in Algorithm 4.

In addition to the above, we still need the learned controls to be admissible. This we implement by enforcing the self-attention mechanism to be causal. Essentially, the attention mechanism is masked to so that the representation at position i only attends to the tokens corresponding to the times $t_k \leq t_i$. This guarantees that the outputs $(\pi_{t_i}, Z_{t_i}, Q_{t_i})$ are non-anticipative functions of $\mathcal{T}_i$. Intuitively, causal masking is the algorithmic analogue of measurability, where it ensures that no future increments or regime changes are used when producing outputs at time t_i. The resulting model therefore acts as a valid functional approximation of the decoupling fields in Proposition 4.1.13. The Algorithm 5 formalizes the evaluation of the causal transformer on a trajectory prefix.

Algorithm 5 Causal transformer evaluation on a trajectory prefix

Require: Prefix $\mathcal{T}_i = \{(S_{t_k}, X_{t_k}, \epsilon_{t_k})\}_{k \le i}$; transformer layers $\{\mathrm{Attn}_\ell, \mathrm{FFN}_\ell\}_{\ell=1}^{L}$.

Embed tokens: $\mathbf{e}_k \leftarrow \mathrm{Embed}(S_{t_k}, X_{t_k}, \epsilon_{t_k}) + \mathrm{PosEnc}(k)$.
Apply causal mask $M_{k\ell} = \mathbf{1}_{\{\ell \le k\}}$ in self-attention.
for $\ell = 1$ **to** L **do**
    $\mathbf{E} \leftarrow \mathrm{Attn}_\ell(\mathbf{E}; M)$ ▷ no access to future tokens
    $\mathbf{E} \leftarrow \mathrm{FFN}_\ell(\mathbf{E})$
end for
Return final-token latent $\mathbf{h}_i \leftarrow \mathbf{E}_i$.

One thing to note is that the with respect to the discretized dynamics and loss construction, the primal and dual value-based Deep BSDE algorithms remain structurally identical to the Markovian algorithms described in the chapter before. The key change however is in the parameterization step where at each time t_i, the control and BSDE quantities are obtained by evaluating the transformer on the trajectory prefix $\mathcal{T}_i$, rather than by calling a time-indexed neural network. This can be expressed compactly by a unified training step that covers both primal and dual variants. The procedure contains: (i) a terminal L^2 regression step enforcing the terminal condition (and derivative condition, in the primal/adjoint setting), (ii) a per-time-step control update step that maximizes (primal) or minimizes (dual) the discrete stage objective induced by the HJB operator, and (iii) for the dual algorithm only, an additional minimization step for the discretized dual functional. Algorithm 6 states this training logic explicitly in transformer form.

As can be seen, the transformer-based formulation can therefore be viewed as an architectural substitution: $(\pi_i, Z_i, Q_i) \leftarrow \big(\mathcal{N}_{\theta_{i,\pi}}, \mathcal{N}_{\theta_{i,Z}}, \mathcal{N}_{\theta_{i,Q}}\big)(t_i, X_{t_i}, \epsilon_{t_i}) \quad \rightsquigarrow$ $(\pi_i, Z_i, Q_i) \leftarrow \big(\mathrm{Head}_\pi, \mathrm{Head}_Z, \mathrm{Head}_Q\big)\big(\mathcal{T}_\theta(\mathcal{T}_i)\big)$ with causal masking enforcing adaptedness. In order to keep things simple and to avoid redundancy, we state the transformer generalization explicitly in Algorithms 5–6, and retain Algorithms 1–2 as canonical references for the Markovian baseline.

Remark 4.2.1. Although the substitution as mentioned above looks a purely architectural modification at first glance, it reflects a deeper mathematical shift in the nature of the objects being numerically obtained. In the Markovian setting, the Deep BSDE algorithm relied on the classical Itô calculus. In contrast, path-dependent value-based formulation requires functional Itô calculus in the sense of Dupire, where the value functions depends on the entire past trajectory and admits both horizontal and vertical derivatives rather than ordinary partial derivatives. The backward representations in the Theorems 4.1.11 and 4.1.12 therefore hinges on the

Algorithm 6 One training step: Value-based Transformer Deep BSDE (primal/dual, regime switching)

Require: Batch size b_{size}, grid $\{t_i\}_{i=0}^{N}$, step Δt; trainable initial variables (v_0, p_0) (primal) or (v_0, p_0, y_0, η) (dual); transformer $\mathcal{T}_\theta$ with causal masking; readout heads $\text{Head}_\pi, \text{Head}_Z, \text{Head}_Q$.

Sample b_{size} paths of $\{\Delta W_i^j\}_{i=0}^{N-1}$ and regime jumps $\{\Delta M_i^{\epsilon,j}, \Delta M_i^{\epsilon,\nabla,j}\}$ according to Q.

▷ Substep 1: terminal L^2 fit (enforce terminal conditions)

for $j = 1$ **to** b_{size} **do**
  Initialize $(S_0^j, X_0^j, \epsilon_0^j)$ and set $v_0^j \leftarrow v_0$, $P_0^j \leftarrow p_0$.
end for
for $i = 0$ **to** $N-1$ **do**
  for $j = 1$ **to** b_{size} **do**
    Construct prefix trajectory $\mathcal{T}_i^j = \{(S_{t_k}^j, X_{t_k}^j, \epsilon_{t_k}^j)\}_{k \le i}$.
    $\mathbf{h}_i^j \leftarrow \mathcal{T}_\theta(\mathcal{T}_i^j)$ ▷ causal attention
    $\pi_i^j \leftarrow \text{Head}_\pi(\mathbf{h}_i^j)$; $Z_i^j \leftarrow \text{Head}_Z(\mathbf{h}_i^j)$; $Q_i^j \leftarrow \text{Head}_Q(\mathbf{h}_i^j)$
    Propagate $(S_{i+1}^j, X_{i+1}^j, v_{i+1}^j, P_{i+1}^j)$ via the discretized forward–backward updates.
  end for
end for
Compute terminal regression loss

$$\text{loss}_1 \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Big(\|v_{t_N}^j - g(X_{t_N}^j, S_{[0,t_N]}^j)\|^2 + \|P_{t_N}^j - \partial_x g(X_{t_N}^j, S_{[0,t_N]}^j)\|^2 \Big).$$

Update (v_0, p_0) (and relevant dual parameters if applicable) with one SGD step w.r.t. loss_1.

▷ Substep 2: control improvement (primal maximization / dual minimization)

for $i = 0$ **to** $N-1$ **do**
  Re-simulate prefixes $\{\mathcal{T}_i^j\}_{j=1}^{b_{\text{size}}}$ and evaluate π_i^j via $\mathcal{T}_\theta$ as above.
  Define stage loss $\text{loss}_{2,i}$:
  if primal then
    $\text{loss}_{2,i} \leftarrow -\frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Big\{ f(\cdot) + \widehat{\mathcal{L}}_i^\pi v(\cdot) \Big\}$.
  else ▷ **dual**
    $\text{loss}_{2,i} \leftarrow +\frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Big\{ f(\cdot) + \widehat{\mathcal{L}}_i^\pi v(\cdot) \Big\}$.
  end if
  Update transformer control head parameters (and optionally shared θ) w.r.t. $\text{loss}_{2,i}$.
end for

▷ Substep 3 (dual only): minimize discrete dual functional

if dual then
  Compute $\text{loss}_{\text{dual}} \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Psi(Y_\eta; \text{path}^j)$.
  Update (η, y_0) with one SGD step w.r.t. $\text{loss}_{\text{dual}}$.
end if

functional HJB system derived in theorem 4.1.10. The role of the transformer is not to alter the analytical structure, but also to provide a flexible non-anticipative parametrization of the resulting decoupling fields on path space. In this sense, the architectural substitution allows for numerical approximation of functional objects whose existence and interpretation are guaranteed by Dupire's calculus, while preserving the mathematical structure of the primal and dual BSDE algorithms.

On the other hand, in the case of policy-based primal-dual formulation, the approach is to directly optimize a stochastic control objective over a parametrized family of anticipative, path-dependent controls, without explicitly enforcing the functional HJB equation or the associated BSDE constraints during training. As can be seen in Algorithm 7, for each iteration the transformer evaluates on the growing trajectories to produce adapted controls. These are then used to propagate the controlled dynamics forward in time. Depending on the nature of the problem (primal or dual), the resulting trajectories are used to construct an unbiased Monte Carlo estimator of the expected utility function in the case of primal problem (or an estimator of the dual objective for the dual problem). The transformer parameters are then subsequently updated via stochastic gradient-based optimization. Although the algorithm 7 is stated in a unified manner (primal and dual together), the distinction between the primal and dual formulations enters only through the choice of loss functional and optimization direction. The architectural role of the transformer as a non-anticipative path-to-decision map remains unchanged in either case. Furthermore, the policy-based primal–dual algorithm does not explicitly enforce the functional HJB equation (4.17) or the BSDE system (4.19) during training, rather, it remains fully consistent with the theoretical characterization of optimality developed earlier in this chapter. Under the standing assumptions of Theorem 4.1.10, any optimal control π^* must satisfy the functional HJB equation along optimal trajectories, and the associated value and adjoint processes must solve the path-dependent BSDEs in Theorems 4.1.11 and 4.1.12. In that direction, Algorithm 7 can be interpreted as a *direct method* in stochastic control. Rather than enforcing optimality conditions pointwise on path space, it searches over a larger class of adapted controls and recovers an optimizer by directly optimizing the primal or dual objective. At convergence, the learned control implicitly satisfies the first-order optimality conditions encoded in the functional HJB equation, while the BSDE representations provide *a posteriori* figure of merit.

Remark 4.2.2. The distinction between the value-based and policy-based formu-

Algorithm 7 One training step: Policy-based primal–dual optimization on path space

Require: Batch size b_{size}; grid $\{t_i\}_{i=0}^{N}$ with step Δt; causal transformer Π_θ; objective type (**primal** or **dual**).

Sample b_{size} independent realizations of Brownian increments $\{\Delta W_i^j\}_{i=0}^{N-1}$ and regime-switching jumps $\{\Delta M_i^{\epsilon,j}\}$.

▷ Forward simulation under learned policy

for $j = 1$ **to** b_{size} **do**
  Initialize $(S_0^j, X_0^j, \epsilon_0^j)$.
end for
for $i = 0$ **to** $N - 1$ **do**
  for $j = 1$ **to** b_{size} **do**
    Construct trajectory prefix

$$\mathcal{T}_i^j = \{(S_{t_k}^j, X_{t_k}^j, \epsilon_{t_k}^j)\}_{k \le i}.$$

    Compute adapted control

$$\pi_i^j \leftarrow \Pi_\theta(\mathcal{T}_i^j).$$

    Propagate $(S_{t_{i+1}}^j, X_{t_{i+1}}^j)$ using (4.11) and (4.12).
  end for
end for

▷ Objective evaluation

if primal then
  Compute Monte Carlo estimator of expected utility:

$$\widehat{J}_{\text{primal}}(\theta) \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \left(\sum_{i=0}^{N-1} f(t_i, X_i^j, S_{[0,t_i]}^j, \pi_i^j)\, \Delta t + g(X_{t_N}^j, S_{[0,t_N]}^j) \right).$$

  Set loss loss $\leftarrow -\widehat{J}_{\text{primal}}(\theta)$.
else ▷ **dual**
  Compute Monte Carlo estimator of the dual objective:

$$\widehat{J}_{\text{dual}}(\theta) \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Psi\big(\mathcal{T}_N^j; \theta\big).$$

  Set loss loss $\leftarrow \widehat{J}_{\text{dual}}(\theta)$.
end if

▷ Policy update

Update transformer parameters θ with one stochastic gradient step w.r.t. loss.

lations mirrors the classical dichotomy between indirect (Euler–Lagrange or HJB-based) and direct (shooting or collocation) methods in deterministic optimal control. Both approaches characterize the same (or atleast similar) optimizer, but differ in how optimality conditions are used during training and computation. The policy-based primal–dual formulation adopted here is particularly well suited to genuinely path-dependent problems, where explicit enforcement of functional PDE constraints is numerically fragile (rough volatility case).

With the algorithms decided, it remains that several architectural choices play an important role in the stability and performance of the transformer-based solver. These choices would include the number of attention heads, the depth L, the embedding dimension d_{model}, and the choice of positional encoding. In all experiments, we use causal masking and shared weights across time, ensuring that the model generalizes across discretized horizons without requiring separate transformer-networks at each time step. Furthermore, there is no explicit architectural choice where assumptions of path-dependence is imposed. Essentially, the same model is used for both finite-memory models (such as moving-average kernels) or with long-memory model (such as rough volatility), allowing the learned attention pattern to adapt endogenously to the structure of the problem.

Remark 4.2.3. Now, despite the fact that we have a single transformer as compared to the time-index networks, this approach incurs additional computation cost due to the quadratic complexity of the self-attention in sequence length. Average computational expense is manageable in the case when the grids are not finely discretized. Furthermore, the cost can also further be reduced when we employ windowing strategies whenever appropriate (mostly us trying to simplify the problem).

4.3 Numerical Experiments: Solving Non-Markovian Dynamics

Experiment 1: Hidden-regime UMP using Higher-Dimensional Markovian State Augmentation

We now return to the single stock, two-regime example introduced in Experiment 1 (cf. Section 2.6) in the previous chapter, but we now remove the unrealistic assumption that the current regime is directly observed. In line with Theorem 4.1.1, the drift is treated as regime-dependent and unobservable, while the diffusion coefficient remains constant, known, and observable. As a recap, the market contains two

latent regimes $\epsilon_t \in \{\epsilon_0, \epsilon_1\}$ governed by the generator

$$Q = \begin{pmatrix} -\lambda_{01} & \lambda_{01} \\ \lambda_{10} & -\lambda_{10} \end{pmatrix}, \qquad (\lambda_{01}, \lambda_{10}) = (0.3, 0.5).$$

The regime-specific model parameters are

$$(\mu_{\epsilon_0}, r_{\epsilon_0}) = (0.11, 0.04), \qquad (\mu_{\epsilon_1}, r_{\epsilon_1}) = (0.06, 0.01), \qquad \sigma_{\epsilon_t} \equiv \sigma = 0.25.$$

When the regime is latent, the investor observes only asset prices and must form a posterior belief about the state of the market. As described in Theorem 4.1.1, this incomplete-information problem becomes a completely observed control problem once we augment the state with the posterior probability

$$p_t^{(0)} = p_t = \mathbb{P}(\epsilon_t = \epsilon_0 \mid \mathcal{G}_t), \qquad p_t^{(1)} = 1 - p_t,$$

where ϵ_0 denotes the Bull regime (R0) and ϵ_1 denotes the Bear regime (R1). The augmented state (X_t, p_t) becomes Markovian under the filtering formulation and therefore admits a standard dynamic programming representation (which we have already derived). Because the utility function is of CRRA type, we can obtain a fully analytical benchmark for this partially observed utility maximization problem. This benchmark will serve as the ground truth against which we validate our deep-learning solvers.

Moreover, Theorem 4.1.1 ensures the existence of an innovation Brownian motion W_t under which the augmented state evolves according to the following dynamics:

$$\frac{dX_t}{X_t} = \widehat{r}(p_t)\, dt + \pi_t\big(\widehat{\mu}(p_t) - \widehat{r}(p_t)\big)\, dt + \pi_t \sigma\, dW_t,$$

$$dp_t = b_p(p_t)\, dt + \Sigma_p(p_t)\, dW_t,$$

where the Brownian motion W_t is the innovation process and the filter coefficients b_p and Σ_p follow the classical Wonham filter (innovation form) for a finite-state hidden Markov chain with regime-dependent drift and known constant volatility. The filtered coefficients can be written as

$$\widehat{\mu}(p) = p\mu_{\epsilon_0} + (1-p)\mu_{\epsilon_1}, \qquad \widehat{r}(p) = pr_{\epsilon_0} + (1-p)r_{\epsilon_1}, \qquad \sigma = 0.25.$$

the value function

$$v(t, x, p) = \sup_{\pi \in \mathcal{A}} \mathbb{E}[U(X_T) \mid X_t = x, p_t = p], \qquad U(x) = \frac{x^{1-\gamma}}{1-\gamma}, \quad \gamma = \tfrac{1}{2},$$

satisfies the filtered HJB equation given in (4.4). The CRRA utility U is homogeneous in x, so we will again use the separable ansatz $v(t, x, p) = A(t, p)U(x)$.

Substituting the ansatz into $\mathcal{L}_p^\pi v$ (given by (4.5) and simplifying using CRRA identities gives

$$\mathcal{L}_p^\pi v = U(x)\Big[b_p A_p + \tfrac{1}{2}\Sigma_p^2 A_{pp} + (1-\gamma)A\widehat{r} + (1-\gamma)A\pi(\widehat{\mu}-\widehat{r}) - \tfrac{\gamma(1-\gamma)}{2}A\pi^2\sigma^2 + (1-\gamma)\pi\sigma\Sigma_p A_p\Big]$$

Since $U(x) > 0$, the ansatz reduces the HJB to a PDE for $A(t, p)$:

$$\begin{aligned} &A_t + b_p A_p + \frac{1}{2}\Sigma_p^2 A_{pp} + (1-\gamma)A\widehat{r} \\ &\quad + (1-\gamma)\sup_{\pi\in\mathbb{R}}\left\{A\,\pi(\widehat{\mu}-\widehat{r}) - \frac{\gamma}{2}A\pi^2\sigma^2 + \pi\sigma\Sigma_p A_p\right\} = 0, \qquad A(T,p) = 1 \end{aligned} \tag{4.27}$$

The term inside the sup is a concave quadratic in π. Define the quadratic

$$\Phi(\pi) = \pi\big(A(\widehat{\mu}-\widehat{r}) + \sigma\Sigma_p A_p\big) - \frac{\gamma}{2}A\,\pi^2\sigma^2.$$

Differentiating:

$$\frac{\partial\Phi}{\partial\pi} = A(\widehat{\mu}-\widehat{r}) + \sigma\Sigma_p A_p - \gamma A\pi\sigma^2.$$

Setting this to zero gives

$$\pi^*(t,p) = \frac{\widehat{\mu}(p)-\widehat{r}(p)}{\gamma\sigma^2} + \frac{\Sigma_p(p)}{\gamma\sigma}\,\frac{A_p(t,p)}{A(t,p)}. \tag{4.28}$$

The first part of this optimal control is the myopic (Merton-type) demand, while the second term represents an intertemporal hedging demand (arising from the uncertainty from the unobserved regimes). Furthermore, substituting π^* into Φ yields

$$\sup_\pi \Phi(\pi) = \frac{\big(A(\widehat{\mu}-\widehat{r}) + \sigma\Sigma_p A_p\big)^2}{2\gamma A\sigma^2}.$$

Substituting back into (4.27) yields the one-dimensional nonlinear PDE:

$$\begin{aligned} &A_t(t,p) + b_p(p)\,A_p(t,p) + \frac{1}{2}\Sigma_p^2(p)\,A_{pp}(t,p) + (1-\gamma)A(t,p)\widehat{r}(p) \\ &\qquad + (1-\gamma)\frac{\Big(A(t,p)(\widehat{\mu}(p)-\widehat{r}(p)) + \sigma\Sigma_p(p)\,A_p(t,p)\Big)^2}{2\gamma A(t,p)\,\sigma^2} = 0, \\ &A(T,p) = 1. \end{aligned} \tag{4.29}$$

Finally, the value function is going to be $v(t = 0, x = 1, p) = 2A(0, p_0)$. This nonlinear parabolic PDE constitutes the complete analytical benchmark for the hidden-regime UMP.

Table 4.2: Deep BSDE Hidden-Regime UMP Results: Primal, Dual, Duality Gap, and Approximate and True Optimal Controls

p_0	v^{primal}	v^{dual}	Duality Gap	$\pi^*(0, p_0)$	π^{primal}	π^{dual}
1.0	2.106554	2.110134	0.003850	2.2400	2.2558	2.2138
0.8	2.096072	2.100640	0.004568	2.1174	2.1399	2.0983
0.6	2.088809	2.093826	0.005017	1.9840	1.9994	1.9716
0.5	2.081268	2.086699	0.005431	1.9200	1.9285	1.9110
0.4	2.074679	2.079391	0.004712	1.8560	1.8725	1.8320
0.2	2.067869	2.071701	0.003832	1.7280	1.7398	1.7036
0.0	2.057715	2.060933	0.003218	1.6000	1.6294	1.5751

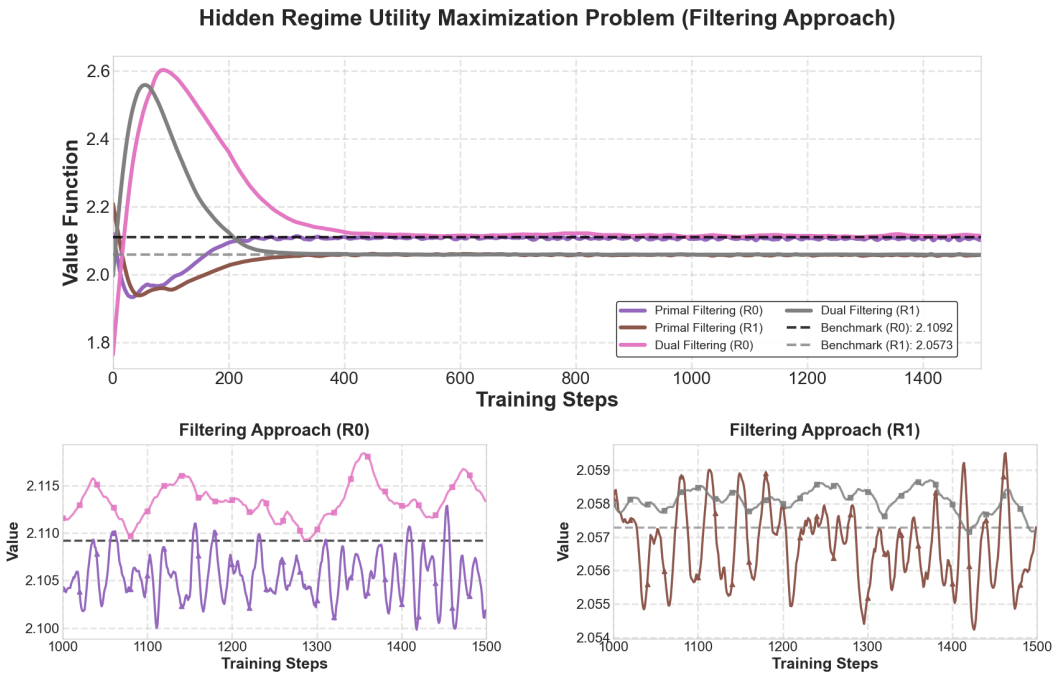

Figure 4.1: **Hidden-Regime Utility Maximization Problem.** Convergence of the primal and dual Deep FBSDE value estimates under extreme prior beliefs $p_0 = 1$ (Bull) and $p_0 = 0$ (Bear). The primal estimates remain below the dual estimates, consistent with weak duality, and both trajectories converge tightly toward the PDE benchmark obtained by solving equation (4.29).

Having established the analytical structure of the hidden-regime UMP and the subsequent nonlinear PDE for $A(t, p)$ (Equation (4.29)), we now solve this PDE numerically to obtain benchmark values $v(0, 1, p_0) = 2A(0, p_0)$ for each prior $p_0 \in [0, 1]$. The PDE here is solved backward in time using a second-order finite-difference discretizations on a uniform grid in the belief variable p. Central differences are used

for the first and second derivatives in p, combined with an implicit backward Euler step in time to ensure stability. Since the diffusion term $\Sigma_p^2(p)$ vanishes only at $p = 0$ and $p = 1$, homogeneous Neumann boundary conditions, $A_p(t, 0) = A_p(t, 1) = 0$ are imposed to maintain well-posedness. The resulting numerical solution $A(0, p)$ provides the ground-truth benchmark for each p_0. These values are used to evaluate the performance of our primal and dual Deep FBSDE algorithms solving the system of equations given by equation (4.7).

Figure 4.1 illustrates the convergence behavior in the extreme cases $p_0 = 1$ and $p_0 = 0$, i.e, 100% belief that regime is either the bull regime (R0) or the bear regime (R1). In both cases, the estimates converge tightly around the numerically obtained PDE benchmark, validating the filtering-based FBSDE formulation. The approximate controls π^{primal} and π^{dual} reported in Table 4.2 also align closely with the optimal analytical control $\pi^*(0, p_0)$ of (4.28), further confirming the accuracy of both solvers in capturing the hedging demand induced by uncertainty of the belief. Furthermore, table 4.2 summarizes the results across a range of prior beliefs $p_0 \in \{1.0, 0.8, 0.6, 0.5, 0.4, 0.2, 0.0\}$. For each p_0, the table reports the primal and dual Deep BSDE value functions, their duality gap, and the optimal portfolio weight $\pi^*(0, p_0)$ computed from (4.28) and $\pi^*(0, p_0)$ from both the deep FBSDE algorithms. Because the optimal control depends on the filtered coefficients and on the ratio A_p/A, the values of π^* decrease smoothly as p_0 becomes more Bear-weighted. This matches economic intuition, i.e, one would allocate more aggressively to the risky asset when they are more confident in the Bull regime and vice-versa.

Furthermore, two qualitative patterns emerge. Firstly, both primal and dual value functions decrease monotonically in p_0, reflecting the decline in expected returns as the prior places more weight on the Bear regime. Secondly, the duality gap is uniformly small typically on the order of 10^{-3} indicating that the numerical solutions closely align with the true values. Notably though however, is the fact that the gap is largest near $p_0 = 0.5$, where regime uncertainty is maximal and belief volatility contributes most strongly to the value-function curvature.

An important consistency check following Proposition 4.1.2 arises in the extreme prior cases $p_0 = 1$ and $p_0 = 0$, corresponding to full certainty that the market at time $t = 0$ in the Bull regime (R0) or the Bear regime (R1), respectively. In these cases, the posterior is degenerate and the hidden-regime problem collapses to the fully observed regime-setting studied in Experiment 1 in the last chapter. This reduction is reflected directly in the structure of the optimal control (4.28) as the

posterior belief is degenerate, $\Sigma_p(p)$ vanishes at the boundaries $p \in \{0, 1\}$, and the intertemporal hedging term disappears. Consequently, the optimal strategy reduces to the classical myopic Merton portfolio corresponding to the known regime. This behavior is clearly visible both in Table 4.2, where the optimal controls align with their full-information counterparts, and in Figure 4.1, where the primal and dual estimates coincide closely with the PDE benchmark.

Furthermore, a boundary effect is also observed in the duality gap. At $p_0 = 0$ and $p_0 = 1$, the duality gap is smallest, reflecting the reduced curvature and uncertainty in the value function when regime risk (as the regime is determinant) is absent. In contrast, the gap attains its maximum near $p_0 = 0.5$, where the uncertainty of the regime is highest, i.e, the informational constraint is most severe, and the volatility of beliefs contributes most strongly to the nonlinear terms in the reduced HJB equation (4.29). This increase in the duality gap is accompanied by the largest discrepancies between the primal and dual approximations of the optimal control π^*, indicating that intermediate belief states leads to deviation in the numerical solutions obtained by both solvers. Overall, these patterns further confirm that the Deep FBSDE framework correctly captures the transition between full-information and partial-information regimes, both at the level of value functions and optimal controls.

Experiment 2: Path-dependent Regime Switching UMP

Lets consider a path-dependent regime-switching UMP where non-Markovianity arises endogenously from rough volatility dynamics. For this lets say that the risk asset follows

$$\frac{dS_t}{S_t} = \mu_{\epsilon_t} dt + \sqrt{V_t}\, dW_t \tag{4.30}$$

where the ϵ_t is as usual the finite-state Markov chain with the generator Q, and the volatility process V_t exhibits rough dynamics of Volterra type. Specifically, we consider a fractional Volterra representation:

$$V_t = V_0 + \int_0^t K(t-s)\, b(\epsilon_s, V_s)\, ds + \int_0^t K(t-s)\, \nu(\epsilon_s, V_s)\, dB_s, \tag{4.31}$$

where B_t is a Brownian motion (possibly correlated with W_t but for the sake of this example we take that correlation to be 0) and

$$K(t) = \frac{t^{H-\frac{1}{2}}}{\Gamma(H+\frac{1}{2})}, \qquad H \in (0, \tfrac{1}{2}), \tag{4.32}$$

is a fractional kernel where H is the Hurst parameter. Due to the presence of the fractional kernel K, the volatility process V_t depends on the entire past trajectory. Consequently, the joint process (S_t, V_t, ϵ_t) is not Markovian in any finite dimensional state space. Any attempt to augment the state with finitely many auxiliary variables necessarily results in a misspecified approximation. p such a scheme, the wealth process becomes

$$\frac{dX_t}{X_t} = r_{\epsilon_t}\, dt + \pi_t\big(\mu_{\epsilon_t} - r_{\epsilon_t}\big)\, dt + \pi_t\sqrt{V_t}\, dW_t. \tag{4.33}$$

and under CRRA utility the final objective being the terminal expected utility of wealth becomes

$$v(t, x, \mathcal{F}_t) = \sup_{\pi\in\mathcal{A}} \mathbb{E}\left[\frac{X_T^{1-\gamma}}{1-\gamma}\,\middle|\, X_t = x,\ \mathcal{F}_t\right], \tag{4.34}$$

where it is fully path-dependent as $\mathcal{F}_t$ denotes the filtration is generated by the full past history of (S, V, ϵ). As established in theorem 4.1.10, the corresponding value function satisfies a functional HJB on the path space and the optimal control can be characterized through a coupled path-dependent FBSDE system. However, due to the nature of the state, the formulation doesn't admit a analytical or a grid-based numerical solution. In that regard, the two approaches developed (i.e, Algorithm 6 and Algorithm 7) will be used to determine the value function and the associated optimal control. For this experiment, unless stated otherwise, we fix the time horizon $T = 1$ and use $N = 30$ time steps. The rough volatility parameters are set to $H = 0.1, \nu = 1.9, V_0 = 0.04$ corresponding to a highly non-Markovian volatility regime. The regime-switching dynamics consist of two states with transition intensities $\lambda_{01} = 0.5$ and $\lambda_{10} = 0.3$. Market coefficients are given by

$$(\mu_0, r_0) = (0.11,\ 0.04), \qquad (\mu_1, r_1) = (0.06,\ 0.01),$$

and the initial conditions are $S_0 = 1$ and $X_0 = 1$. The correlation between the asset and volatility drivers is set to $\rho = -0.7$, capturing the empirically observed leverage effect. Unless stated otherwise, all expectations are approximated via Monte Carlo simulation using mini-batches of size 64.

Remark 4.3.1. The non-Markovian nature of the problem arises directly from the fractional kernel $K(t - s)$, which assigns non-negligible weight to the entire past trajectory of the volatility process. Unlike classical stochastic volatility models, where the state can be represented by a finite-dimensional Markov process, the Volterra structure implies that V_t depends on the full history $\{B_s : s \le t\}$ through

a slowly decaying memory kernel. In particular, for $H < \frac{1}{2}$, the kernel exhibits a singularity at the origin and a power-law decay, leading to long-range temporal dependence. As a result, the effective state space is infinite-dimensional, and no finite collection of auxiliary variables can fully summarize the system without approximation.

This nuances in this behavior are shown in Figure 4.2, which compares rough and Markovian dynamics along a representative trajectory of a sample. The rough volatility path exhibits pronounced spikes and clustering, while the Markovian counterpart remains comparatively smooth. Correspondingly, the induced wealth dynamics under rough volatility display visibly different evolution patterns, reflecting the influence of long-range temporal dependence.

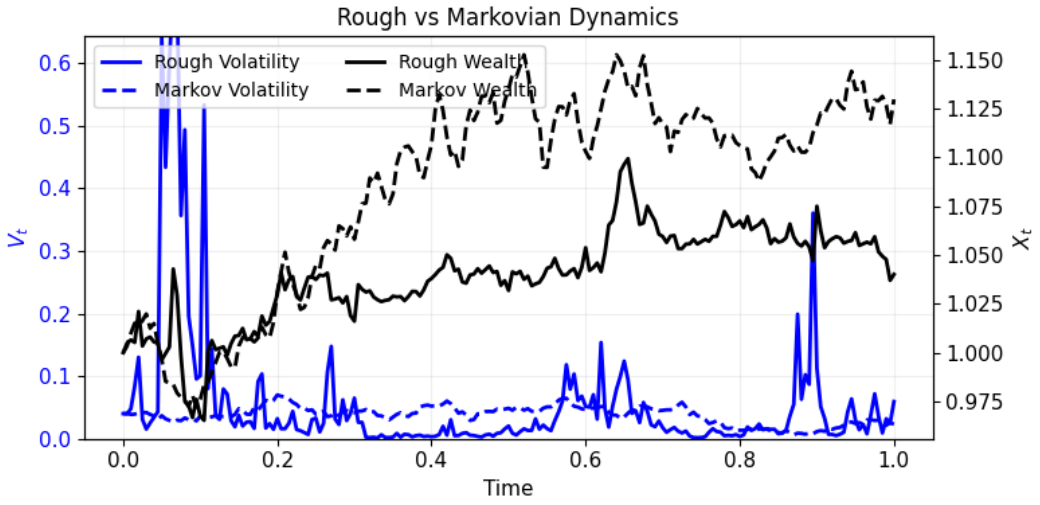

Figure 4.2: **Pathwise Comparison of Rough and Markovian Dynamics.** Evolution of the volatility process V_t (left axis, blue) and wealth process X_t (right axis, black) along a representative sample path (parameters used $H = 0.1, \nu = 1.9, V_0 = 0.04, \rho = -0.7$). Solid lines correspond to the rough volatility model, while dashed lines correspond to the Markovian model. The rough volatility trajectory exhibits burst-like behavior and clustering, reflecting its long-memory structure. These features propagate into the wealth dynamics, resulting in path-dependent deviations relative to the Markovian case.

Before moving on to the transformer-based solvers, we first construct a benchmark numerical policy for the same rough-volatility regime-switching environment. The purpose of this benchmark is not to solve the full functional HJB system derived earlier, but instead to provide a stable reference ground truth point against which both

Algorithm 8 Benchmark Policy Optimization via Monte Carlo Simulation

Require: Batch size b_{size}, time grid $\{t_i\}_{i=0}^{N}$, step Δt; policy network Π_θ; learning rate α; penalty parameter η.

Initialize policy parameters θ.

▷ Training loop

for training step $k = 1$ to K **do**

  Sample b_{size} trajectories of:

  - Brownian increments $\{\Delta W_i^j, \Delta B_i^j\}$
  - Regime process $\{\epsilon_i^j\}$ via generator Q

  Simulate rough volatility paths $\{V_i^j\}$ using the fractional kernel.

  for $j = 1$ to b_{size} **do**

    Initialize (X_0^j, S_0^j).

  end for

  for $i = 0$ to $N - 1$ **do**

    for $j = 1$ to b_{size} **do**

      Evaluate control:

$$\pi_i^j \leftarrow \Pi_\theta\big(X_i^j, \log S_i^j, \sqrt{V_i^j}, \epsilon_i^j, t_i\big)$$

      Update wealth and asset:

$$X_{i+1}^j \leftarrow X_i^j + X_i^j\big(r_{\epsilon_i^j} + \pi_i^j(\mu_{\epsilon_i^j} - r_{\epsilon_i^j})\big)\Delta t + X_i^j \pi_i^j \sqrt{V_i^j}\,\Delta W_i^j$$

    end for

  end for

  Compute objective:

$$\widehat{J}(\theta) = \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \left[g(X_T^j) - \eta \cdot \frac{1}{N} \sum_{i=0}^{N-1} (\pi_i^j)^2 \right]$$

  Update parameters:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \widehat{J}(\theta)$$

end for

▷ Evaluation

Estimate value and control statistics using large-scale Monte Carlo simulation.

the transformer-based value-based and policy-based formulations can be compared. For this reason, we train a feedforward NN policy $\pi_t^\theta = \Pi_\theta\big(X_t, \log S_t, \sqrt{V_t}, \epsilon_t, t\big)$, which maps the currently observed state variables into a portfolio weight. Thus, while the underlying market dynamics remain genuinely non-Markovian through the rough volatility process, the benchmark policy itself is parameterized only through a finite-dimensional collection of contemporaneous features rather than the entire trajectory prefix. Do note that the value of the benchmark provides a practical learned control under the same simulated dynamics, but it is not itself a path-space architecture. Furthermore, in order to avoid unrealistically aggressive portfolio control values and to improve numerical stability, we consider a regularized (with a penalty) version of the UMP. Instead of optimizing terminal utility alone in this experiment, we maximize the penalized objective

$$J_{\text{bench}}(\theta) = \mathbb{E}\left[g\big(X_T^{\pi^\theta}\big) - \eta \cdot \frac{1}{N} \sum_{i=0}^{N-1} \big(\pi_{t_i}^\theta\big)^2 \right], \tag{4.35}$$

where g denotes the CRRA specification with risk aversion parameter $\gamma = \frac{1}{2}$ and $\eta \geq 0$ is a regularization coefficient penalizing large control magnitudes. The quadratic penalty on the control π actually serves two purposes. Economically and for interpretation, it discourages excessive leverage and yields smoother, more realistic portfolio policies. Numerically on the other hand, it stabilizes training by preventing the optimizer from exploiting extreme exposures in highly volatile regions of the state space. The objective in (4.35) does not admit a closed-form solution, nor can it be handled by naive Monte Carlo over static controls. The key difficulty is that the control process is adapted and enters multiplicatively in both the drift and diffusion of the wealth dynamics, so the law of the controlled trajectory depends on the policy itself. Due to this reason, both the simulation and optimization are intrinsically coupled due to which one must simulate trajectories under the current policy, estimate the resulting objective, and then iteratively update the policy parameters. We employ a direct policy-optimization procedure based on stochastic gradient ascent over simulated paths. At each training step, we generate regime paths, rough-volatility trajectories, and the corresponding controlled wealth dynamics under the current policy. The empirical regularized utility is then used as the optimization objective for updating the network parameters. The full benchmark training procedure is summarized in Algorithm 8. Unlike classical Monte Carlo methods for uncontrolled problems, where expectations can be estimated independently of the decision rule, here the decision rule directly alters the state evolution

and therefore must be learned jointly with the trajectory distribution. Note that although this benchmark is useful as a reference model, it has an important structural limitation where the policy depends only on current state variables and therefore cannot explicitly encode the full path-dependence induced by rough volatility. This limitation is essentially the main reason what motivates the transformer-based formulations developed in this chapter, where the control and value-related quantities are parameterized directly as non-anticipative functions of trajectory prefixes.

We begin by evaluating the transformer-based policy (TP) and value (TV) formulations on a set of Markovian benchmark problems for which reliable reference solutions are possible through primal and dual FBSDE methods (refer the numerical experiments section of chapter 2). The results are reported in Table 4.3, while Figure 4.3 visualizes the corresponding deviations from a duality-centered reference value. The goal of this comparison is to assess whether the transformer-based approaches recover consistent value estimates in settings where the solution structure is well understood. The observed agreement of TP and TV across all experiments indicates that the learned representations capture the essential features of the underlying control problem.

While the agreement with FBSDE-based solutions provides an important validation, the primary motivation for the transformer-based formulations lies in their structural generality. As discussed before FBSDE methods rely fundamentally on the existence of a tractable Markovian representation of the state dynamics, or on an equivalent formulation in terms of a finite-dimensional backward equation. As a result, their applicability is inherently tied to problems where the underlying system can be reduced to a manageable state space. In contrast, the transformer-based approaches (both TP and TV) are model-agnostic in the sense that they do not require any explicit Markovian structure or possible closed-form characterization of the value function. Instead, both the policy and value representations are learned directly as non-anticipative functionals of the observed trajectory. This allows the same architecture to be applied uniformly across Markovian and non-Markovian settings, without any modification to the underlying formulation. Therefore, classical methods remain highly efficient when their structural assumptions are satisfied, meanwhile the transformer-based framework provides a more universal approach that extends naturally to class of UMP where such assumptions break down.

Looking more closely into Table 4.3 and Figure 4.3, we see a systematic difference between the transformer-based value (TV) and policy (TP) formulations in terms of

Table 4.3: Value function comparison across methods and experiments under different initial regimes.

Initial Regime $R0$ (Bull)						
Experiment	P-FBSDE	D-FBSDE	TV-P	TV-D	TP-P	TP-D
RS Merton	2.1470	2.1740	2.1315	2.1695	2.1290	2.1920
Regime Utility	2.2152	2.2968	2.2020	2.2790	2.1735	2.2885
Consumption	3.0699	3.1220	3.0790	3.1320	3.0345	3.1170
Risk-Sensitive	1.2937	1.3433	1.3040	1.3640	1.2785	1.3990
High-Dim (Constr. B)	2.3817	2.3891	2.3225	2.4285	2.2880	2.4425
Initial Regime $R1$ (Bear)						
Experiment	P-FBSDE	D-FBSDE	TP-P	TP-D	TV-P	TV-D
RS Merton	2.0606	2.0642	2.0312	2.0833	2.0309	2.0938
Regime Utility	2.0082	2.0408	2.0185	2.0515	1.9913	2.0965
Consumption	2.8075	2.8662	2.8001	2.8715	2.7835	2.9090
Risk-Sensitive	1.2200	1.2678	1.2235	1.2755	1.2075	1.3020
High-Dim (Constr. B)	2.4562	2.4685	2.4450	2.5345	2.3915	2.5175

Note: The table summarizes value function estimates across all five numerical experiments presented in this chapter. Experiment 1 corresponds to the regime-switching Merton benchmark, Experiment 2 to regime-dependent utility (homothetic case), Experiment 3 to the consumption–investment problem with running utility (proportional case), Experiment 4 to risk-sensitive terminal utility (entropic case), and Experiment 5 to the high-dimensional regime-switching model under leverage and short-sale constraints (Case B). Transformer-based policy (TP) and value (TV) methods are evaluated on the same Markovian setups.

dual gaps. Across all experiments and both regimes, the TV estimates tend to remain closer to the reference band defined by the primal and dual FBSDE solutions, while the TP estimates exhibit a visibly larger spread. This behavior can be attributed to the structural differences between the two approaches. The value-based formulation is trained to approximate the solution of the underlying backward equation, implicitly enforcing terminal conditions and consistency with the associated HJB structure (equivalent to FBSDE system itself). As a result, the learned representation is constrained to learn the intrinsic duality relationship between primal and dual formulations, leading to tighter learning dynamics around the reference value. On the other hand, the policy-based formulation directly optimizes the control objective without explicitly enforcing value function consistency or boundary conditions. While this approach provides greater flexibility to solve different types of UMPs, it also introduces additional degrees of freedom, which manifest as a larger duality

gap in the numerical estimates.

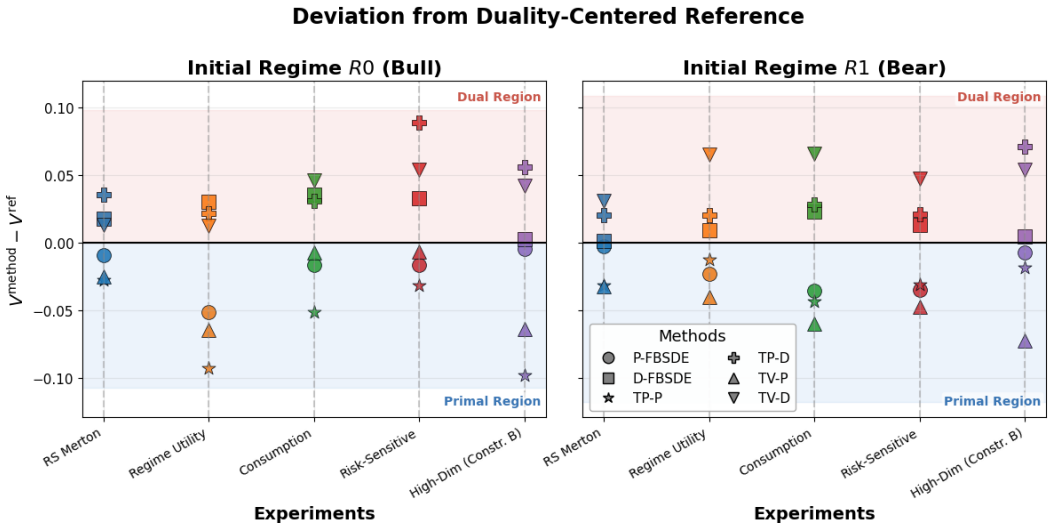

Figure 4.3: **Deviation from Duality-Centered Reference Across Initial Regimes.** Signed errors $V^{\text{method}} - V^{\text{ref}}$ are shown for both initial regimes, with the left panel corresponding to $R0$ (Bull) and the right panel to $R1$ (Bear). The reference value V^{ref} is defined as the midpoint between the primal and dual FBSDE estimates, so the zero line separates primal (underestimation) and dual (overestimation) regions. Marker shapes denote the method, while colors correspond to the different experiments. Across both regimes, classical FBSDE methods remain tightly concentrated around the reference, whereas transformer-based approaches exhibit larger dispersion, with the effect becoming more pronounced in more complex settings.

It is important to take a step back here and see that the transformer-based formulations do not provide a computational advantage over the FBSDE methods in the Markovian settings considered here. Deep BSDE methods are specifically tailored to exploit the underlying structure of such problems, leveraging finite-dimensional state representations and well-understood backward dynamics. This leads to efficient training, strong numerical stability, and tight duality gaps, as reflected in the results above. The transformer architectures introduce a significantly larger parameter space and a more flexible, but less structured, representation of the solution. While this flexibility enables the methods to recover semi-accurate value estimates, it also results in slower convergence and increased variability relative to other approaches. In essence, using transformers for Markovian problems will actually be an overkill (result also seen in Table 2.1), as the additional modeling capacity doesn't translate to improved learning and result. That being said, the fact that both TP and TV formulations remain consistent with the FBSDE reference

solutions demonstrates that the framework is robust, even when applied outside its most natural domain.

The advantage of the transformer-based framework becomes apparent when moving beyond the Markovian setting to problems with genuine path dependence. The key mechanism that enables transformers to address this challenge is self-attention. At each time step, the model processes the entire observed trajectory prefix and assigns data-dependent weights to past states, effectively learning which portions of the history are most relevant for the current decision. In this way, the control or value at time t is constructed as a nonlinear aggregation of past observations, allowing the model to adaptively encode long-range temporal dependencies without imposing any predefined memory structure. This perspective is particularly natural in the context of rough volatility dynamics, where the evolution of the volatility process is governed by a Volterra-type kernel that assigns nontrivial weight to the entire past. The attention mechanism can be interpreted as learning an effective memory kernel directly from simulated data, thereby approximating the functional dependence induced by the underlying stochastic system. As a result, the transformer-based formulations provide a flexible and unified approach for solving stochastic control problems on path space, making them especially well-suited for the non-Markovian regime-switching models considered in the remainder of this numerical example.

Having established the relative behavior of the four methods through benchmarking Markovian example, we now turn to the primary objective of this experiment which is to solve the fully path-dependent regime-switching UMP under rough volatility dynamics. The transformer-based approaches are no longer being validated against a known ground truth, but are used as the primary computational tools for approximating both the value function and the optimal control. To better understand how the proposed transformer-based formulations perform, we study a structured study across four key dimensions of the model: the Hurst parameter H, the volatility-of-volatility coefficient ν, the regularization strength η, and the effective history length L (the input length to the transformer). The objective of this study is not merely to report performance differences, but to identify which aspects of the problem most strongly affect the learned value function and control, and to clarify the relative strengths of the value-based and policy-based transformer formulations. Figures 4.4 and 4.5 summarize these results separately for the Bull and Bear regimes, allowing us to disentangle regime effects from architectural effects. Taken together, the results show that the transformer methods remain sensitive to the right economic

and stochastic features of the model, while also revealing a consistent hierarchy in stability, which is value-based methods produce tighter primal–dual agreement and lower variance, whereas policy-based methods are more flexible but more dispersed.

As can be seen in Panels (a–b) for the bull case, we vary the Hurst parameter H[2], which directly controls the roughness of the volatility process. Lower values of H correspond to rougher paths with stronger short-term irregularity and long-range dependence, increasing the effective dimensionality of the control problem. In this regime, all transformer-based methods consistently outperform the baseline (set using 8), with the largest obtained value functions observed at the lowest roughness level: at H=0.05, TV-P achieves 2.2872 versus 2.2594 for the baseline (+1.23%), while TP-P reaches 2.2745 (+0.67%). This provides clear evidence that causal self-attention captures path-dependent structure that cannot be represented through finite-dimensional Markovian features. Moreover, the distinction between value-based and policy-based formulations becomes more pronounced. The TV methods exhibit consistently tighter primal–dual sandwich across all values of H - at H=0.05, the TV gap is 0.0166 compared to 0.0453 for TP, narrowing to 0.0121 and 0.0332 respectively at H=0.49. This indicates that the HJB-consistent value function approximation acts as a stabilizer for the optimization. As H increases toward the nearly Markovian regime, both the difficulty of the problem and the duality gaps decrease, while the relative ordering of methods remains unchanged. Finally, the optimal control $\bar{\pi}$ decreases monotonically with H, from 4.348 at H=0.05 to 3.844 at H=0.49 under TV-P. This aligns with economic intuition which is smoother volatility reduces short-term fluctuations and thus diminishes exploitable structure, leading to more conservative optimal strategies. These results are consistent with existing literature on UMP under rough volatility (Han and Wong, 2021), which show that optimal control is sensitive to the roughness of the volatility process and must be solved numerically.

Remark 4.3.2. One thing to note is that under CRRA utility, the optimal allocation can be decomposed into a myopic component (Merton control) and an intertemporal hedging demand. In stochastic volatility models, the hedging component essentially reflects the sensitivity of future investment opportunities to changes in the volatility process which depends critically on both the risk aversion parameter γ and the leverage correlation ρ. In rough volatility models, the volatility process exhibits

[2]Unless otherwise stated, all numerical values reported in this subsection correspond to the Bull regime (R_0), with the Bear regime (R_1) exhibiting the same qualitative behavior at lower absolute value levels.

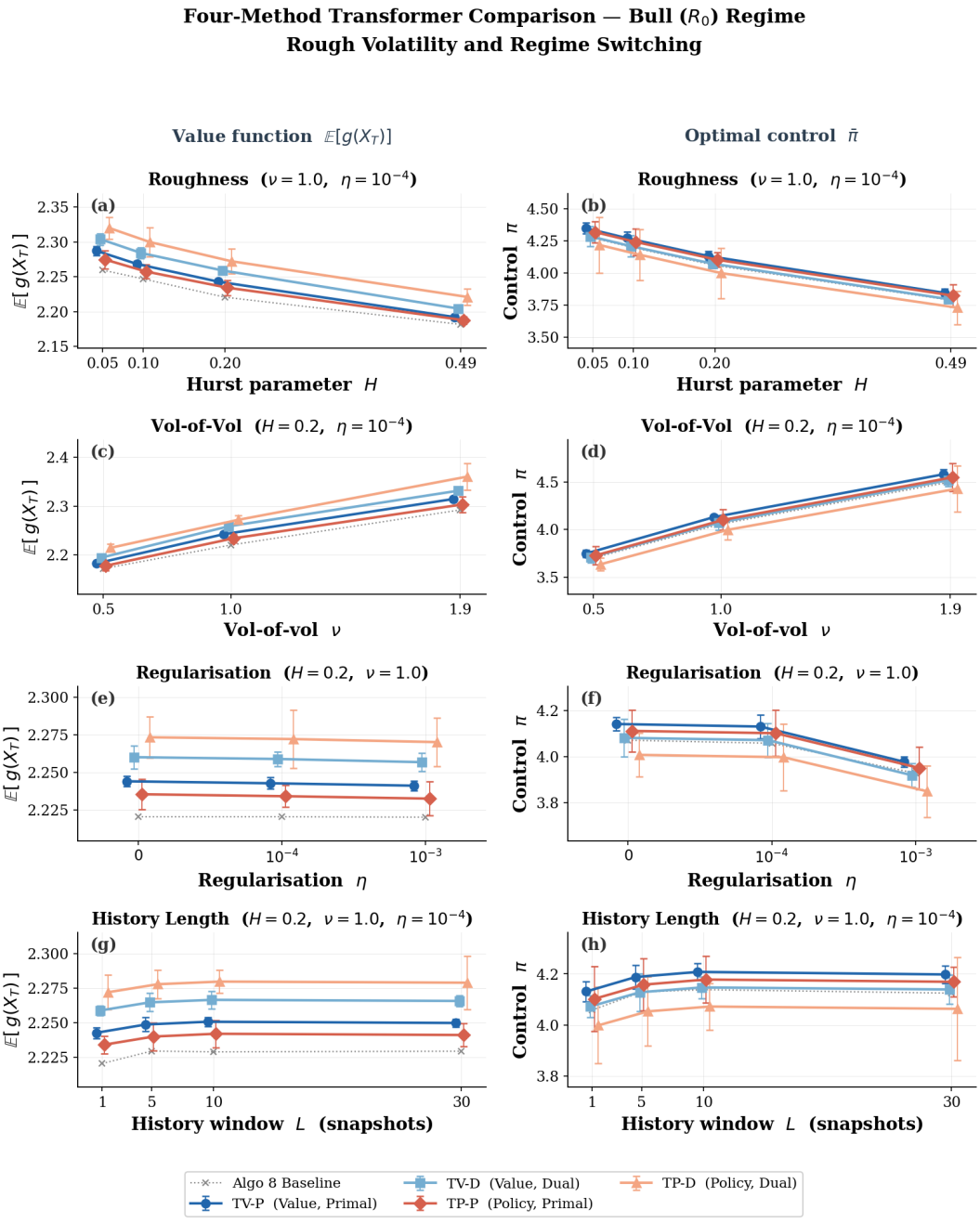

Figure 4.4: **Transformer-based solutions under rough volatility — Bull (R_0) regime.** Value function (left column) and optimal control (right column) across variations in roughness H, volatility-of-volatility ν, regularization η, and history length L. Marker shapes denote the method (TV vs TP, primal vs dual), while colors correspond to different experimental configurations. Value-based methods exhibit tighter clustering and improved stability, while policy-based methods show greater dispersion, particularly in more challenging regimes.

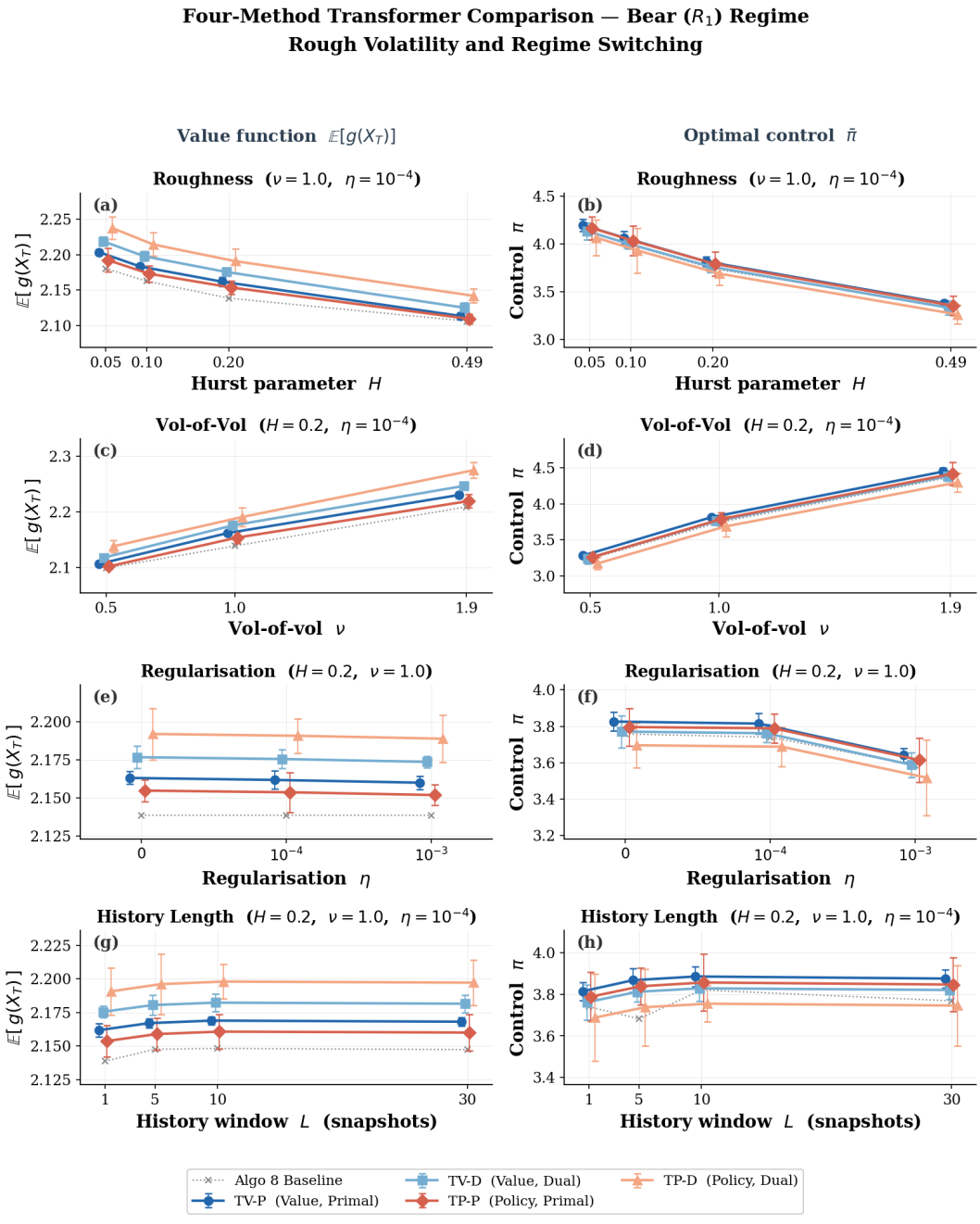

Figure 4.5: **Transformer-based solutions under rough volatility — Bear (R_1) regime.** Same configuration as Figure 4.4, evaluated under the Bear regime. Overall value levels are lower relative to the Bull regime, while maintaining consistent structural behavior across methods. The separation between value-based and policy-based formulations persists, with value-based methods exhibiting greater numerical stability.

strong temporal dependence, and the investment opportunity set becomes path-dependent. As shown in Han and Wong (2021), optimal strategies in this case must be computed numerically and depend on the full trajectory of the state variables. While the literature does not establish a universal relationship between roughness and optimal allocation, the numerical results we present indicates that roughness significantly affects the magnitude of the hedging demand. In our setting, with $\gamma < 1$ and $\rho < 0$, the hedging demand contributes positively to the optimal allocation. The observed increase in $\bar{\pi}$ as H decreases is therefore consistent with a strengthening of the path-dependent component of the investment opportunity set, which amplifies the impact of intertemporal hedging.

Remark 4.3.3. From an economic perspective, the behavior of the optimal allocation can be understood through the interaction between leverage effects and intertemporal hedging demand. When the correlation between leverage and returns satisfies $\rho < 0$, negative return shocks are associated with increases in volatility, implying that adverse market movements coincide with deterioration of investment opportunities. In such situations, a risk-averse investor anticipates that high-volatility states are unfavorable, as they are typically associated with lower expected returns or higher uncertainty. This gives rise to a nontrivial intertemporal hedging demand where the investor seeks to adjust current positions to account for future changes in the investment opportunity set. Under negative correlation, the risky asset itself partially spans these fluctuations, since its returns co-move negatively with volatility innovations. As a result, holding a larger position in the risky asset provides a partial hedge against adverse shifts in volatility, leading to an optimal allocation that can exceed the myopic (Merton) control. This mechanism is well documented in stochastic volatility models, such as the Heston framework, where the sign and magnitude of the hedging demand depend critically on ρ and the investor's risk preferences (Liu, 2007; Kim and Omberg, 1996). Empirically, the case $\rho < 0$ is particularly relevant for equity markets, where volatility tends to rise during market downturns.

Remark 4.3.4. As $H \to \frac{1}{2}$, the volatility process approaches a Markovian diffusion, and the model converges toward a classical stochastic volatility setting with a finite-dimensional state representation. In this limit, the investment problem can be characterized through a standard HJB equation, and the optimal allocation admits a decomposition into myopic and hedging components that depend only on the current state variables (Kim and Omberg, 1996; Liu, 2007). In the special case

of constant investment opportunities, this further reduces to the classical Merton solution, where the optimal allocation is given by a constant proportion and no hedging demand arises. As the model transitions from rough to Markovian dynamics, the influence of path dependence diminishes, and the contribution of intertemporal hedging becomes entirely determined by a finite-dimensional state. The numerical convergence of both the duality gap and the optimal allocation as H increases is consistent with this transition: as the problem becomes increasingly Markovian, the advantage of path-dependent representations decreases, and the solution approaches that of classical stochastic control formulations.[3] This limiting behavior provides a natural benchmark for the proposed methods, demonstrating that the transformer-based framework recovers classical solutions in the near-Markovian regime while retaining the flexibility to capture genuinely path-dependent effects when roughness is significant.

We next consider Panels (c–d), where we vary the volatility-of-volatility parameter ν, which controls the intensity of stochastic fluctuations in the variance process. In contrast to the roughness parameter, increasing ν amplifies uncertainty and creates more pronounced variability in the underlying dynamics. This is reflected directly in both the value function and the optimal control. For instance, under TV-P, the value increases from 2.1828 at ν=0.5 to 2.3148 at ν=1.9 (+6.0%), while the average allocation $\bar{\pi}$ rises from 3.748 to 4.583, indicating that the investor responds to increased volatility dispersion by taking more aggressive positions to exploit potential upside. From an optimization perspective, higher ν also exposes differences in stability across methods. The duality gap for value-based methods remains relatively controlled, increasing from 0.0119 at ν=0.5 to 0.0170 at ν=1.9 (1.43×), whereas the gap for policy-based methods widens more significantly, from 0.0367 to 0.0570 (1.55×). This again highlights the greater sensitivity of policy-based approaches to stochastic complexity, as they lack the structural regularization induced by the explicit HJB constraint. This effect is further reflected in estimation variance where the standard error of TP-D increases from 0.0118 to 0.0181, the largest absolute increase among all methods, indicating that the loosest dual formulation becomes increasingly unstable as volatility-of-volatility grows. Overall, these results reinforce

[3] At H=0.49, the TV duality gap narrows to 0.0121 in the Bull regime and 0.0115 in the Bear regime—the tightest across the entire Hurst grid—while the optimal allocation converges to $\bar{\pi} \approx 3.84$ and $\bar{\pi} \approx 3.38$, respectively. These are close to the Markovian Algo 8 reference values of 3.79 and 3.34, with discrepancies below 1.5%. By contrast, at H=0.05, the gap widens to 0.0166 and 0.0153, and the allocation deviates from the MLP reference by more than 3%, reflecting the increasing importance of path dependence as roughness intensifies.

that while all methods respond directionally correctly to increased stochasticity, value-based formulations maintain superior numerical stability in more challenging regimes.

We now turn to Panels (e–f), where we vary the regularization parameter η, which penalizes large control magnitudes and serves as a stabilizing mechanism in the learing and optimization process. In contrast to the previous parameters, regularization has a minimal impact on the value function across all methods. For example, under TV-P, the value changes only marginally from 2.2441 at η=0 to 2.2412 at η=10^{-3}, a difference of just 0.0029, indicating that the optimal objective remains largely unaffected by the penalty. However, the effect on the control is more pronounced. The average allocation $\bar{\pi}$ decreases from 4.142 to 3.978 (-3.9%) under TV-P, confirming that the regularization term effectively controls the portfolio aggressiveness without sacrificing the value. Importantly, the relative ordering of all methods remains unchanged across all values of η, suggesting that the performance hierarchy observed earlier is robust to regularization choices. This highlights that the penalty primarily acts as a control stabilizer rather than altering the fundamental structure of the learned solution.

Finally, we examine Panels (g–h), where we vary the effective history length L, which controls the amount of past information explicitly provided to the transformer at every time t. Across all transformer-based methods, the value function remains remarkably stable as L increases from 1 to 30. For instance, under TV-P, the value varies only between 2.2428 and 2.2508, indicating that performance is largely insensitive to the length of the explicitly supplied history. A small improvement is observed when increasing L from 1 to 5 (+0.0060), after which the gains quickly saturate. This behavior provides strong empirical evidence that the transformer architecture is not relying on explicit history concatenation to capture temporal dependencies. Instead, the model is able to internally extract and weight relevant information from the observed trajectory through its attention mechanism. In contrast, the baseline exhibits greater variation across different values of L, reflecting its dependence on manually constructed input windows. These results support the interpretation that self-attention acts as a learned memory mechanism, allowing the model to approximate path-dependent functionals without requiring explicit control over the history length.

Taken together, these results demonstrate that the proposed transformer-based formulations successfully solve the fully path-dependent regime-switching UMP under

rough volatility dynamics. Both value-based and policy-based approaches recover stable and economically consistent solutions across all ablation dimensions, while the attention mechanism effectively captures the underlying non-Markovian structure without requiring explicit state augmentation. Overall, these findings confirm that the framework provides a practical and robust solution methodology for stochastic control problems that lie beyond the reach of classical Markovian techniques.

Experiment 3: Path-dependent drift through a historical-average rule

We now consider a path-dependent control problem, where the path-dependence is not in the form of equation (4.13) but in which the non-Markovianity enters through the one of the coefficient. We replicate the experimental setup of Weidermann (Wiedermann, 2022) exactly in order to enable a direct comparison, applying our four transformer-based methods to a problem where only algorithmic benchmarks are available rather than closed-form solutions. Following Weidermann (Wiedermann, 2022), we study an unconstrained multi-asset problem, with no explicit regime-switching, and with $m = 5$ risky assets and terminal utility $U(x) = 2\sqrt{x}$, with initial wealth $x_0 = 1$ and time horizon $T = 0.5$. The volatility matrix is taken to be constant, with diagonal entries equal to 0.2 and all off-diagonal entries equal to 0.05, while the risk-free rate is fixed at $r = 0.1$. The path dependence is introduced by letting the drift of each asset depend on whether its current value lies above or below its own running historical mean. More precisely, for each asset $i \in \{1, \dots, 5\}$ and $t \in (0, T]$, we set

$$\mu_i(t) = \begin{cases} \mu_{\text{high}}, & \text{if } S_i(t) \geq \dfrac{1}{t} \displaystyle\int_0^t S_i(s)\, ds, \\ \mu_{\text{low}}, & \text{otherwise,} \end{cases} \tag{4.36}$$

with $\mu_{\text{low}} = 0.08$ and $\mu_{\text{high}} = 0.12$. This specification creates a simple trend(momentum)-following mechanism in the drift and makes the control problem genuinely path-dependent, since the coefficient at time t depends on the entire past trajectory of the asset rather than on the current state alone. In particular, the drift depends on a running time-average, which cannot be summarized by a finite-dimensional state without augmentation, making classical HJB-based approaches inapplicable in their standard form. As a result, prior work relies (Wiedermann, 2022) on stochastic maximum principle (SMP)-based algorithms to produce lower and upper bounds for the value function. We follow this approach and use the reported bounds in Weidermann's work as a reference, while applying the four transformer-based methods developed in this chapter. This allows us to assess not only the numerical efficiency

Table 4.4: Final value estimates and duality gaps for transformer-based methods in the path-dependent drift problem.

Method	Primal (Lower)	Dual (Upper)	Gap
TV-P	2.0512	2.0524	0.0012
TV-D	2.0513	2.0525	0.0012
TP-P	2.0509	2.0558	0.0049
TP-D	2.0510	2.0562	0.0052

and flexibility of the transformer formulations, but also their ability to capture path-dependent structure in a setting where only algorithmic benchmarks are available rather than exact solutions.

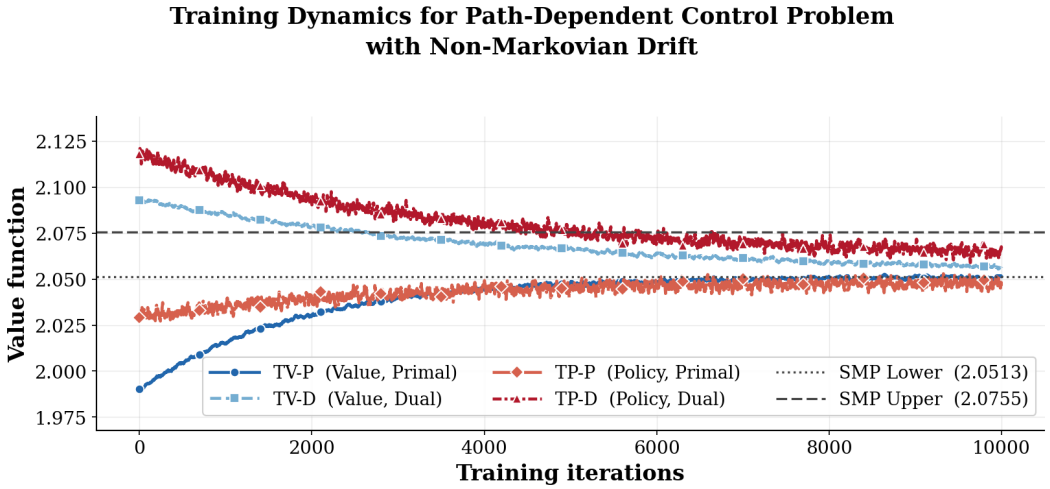

Figure 4.6: **Training dynamics for path-dependent control problem with non-Markovian drift.** Convergence trajectories of the four transformer-based methods (TV-P, TV-D, TP-P, TP-D) over 10,000 training iterations, alongside the SMP-derived lower and upper bounds (dashed reference lines). Value-based methods (TV-P, TV-D) exhibit smooth and monotone convergence with a small, stable duality gap, while policy-based methods (TP-P, TP-D) display pronounced fluctuations and a wider primal–dual separation throughout training, reflecting the higher variance inherent to direct policy optimisation. All four methods ultimately converge to value levels consistent with the SMP bounds.

For the numerical implementation, we follow the discretization and simulation framework[4] used in the original experiments in order to ensure we make 1-1 comparison. The time interval $[0, T]$ is discretized using N time steps, and expectations

[4]We also ensure that the learning schedules closely replicate the learing behavior as in Weidermann's work. Therefore, the learning in this experiments will be highly hyperparameter tuned.

are approximated via Monte Carlo sampling with a batch size of $b_{\text{size}} = 64$ and $N_{\text{MC}} = 10^5$ simulated trajectories. All transformer-based methods are trained for 10,000 iterations, with piecewise constant learning rate schedules analogous to those used in the SMP-based algorithms. The network architectures are chosen to be consistent across methods, differing only in their output heads to accommodate value-based and policy-based formulations. The path-dependent feature $\frac{1}{t}\int_0^t S_i(s)\,ds$ is not explicitly given as an input, but instead must be inferred implicitly through the attention mechanism, allowing us to test whether the transformer architecture can internalize such running functionals directly from trajectory data.

The numerical results confirm that all four transformer-based methods successfully capture the path-dependent structure of the problem and produce consistent value estimates (Table 4.4, Figure 4.6). In particular, both value-based methods (TV-P and TV-D) yield tightly aligned lower and upper bounds throughout training, resulting in a small and stable duality gap of 0.0012 in both cases. On the other hand, the policy-based methods (TP-P and TP-D) exhibit a larger dispersion between primal and dual estimates with duality gaps of 0.0049 and 0.0052 respectively, roughly four times wider than the value-based counterparts with visibly higher variance during training, although they ultimately converge to comparable value levels (see Table 4.4). Quantitatively, the final estimates produced by the transformer methods are in close agreement with the bounds reported by the SMP-based algorithms. We actually also see a mild downward bias (of the order of 10^{-3}) toward the SMP lower bound. The finite capacity of the transformer networks introduces a systematic underestimation, as the learned representation cannot perfectly approximate the optimal control for either the primal or dual formulation.

One thing to note, however, is that the SMP-based algorithms converge considerably faster (approximately 2,000 iterations (Wiedermann, 2022)) versus transformer-based models, which take considerably longer due to the quadratic scaling of time complexity originating from the self-attention mechanisms. Moreover, as visible in Figure 4.6, the convergence behaviour differs markedly across formulations as the value-based methods (TV-P and TV-D) display smoother and more monotone trajectories, while the policy-based methods (TP-P and TP-D) show more pronounced fluctuations, particularly in the early stages of training. Overall, this example is a simple demonstration that the transformer based networks we have developed are versatile and can help us solve a variety of UMPs. One can take this example as a benchmark and solve the other examples that (Wiedermann, 2022) had constructed

as well.

Experiment 4: Linear–Quadratic Stochastic Control (Engineering Benchmark)

To further demonstrate that the proposed framework is not restricted to financial applications or rough-volatility settings, we next consider a classical linear–quadratic (LQ) stochastic control problem as an engineering benchmark case. The purpose of this experiment is twofold. First, it shows that the four transformer-based formulations and the Deep BSDE methods developed in this thesis can be applied to a broader class of stochastic control problems. Second, unlike the path-dependent rough-volatility experiment, the LQ problem admits a closed-form analytical solution, allowing us to directly benchmark the learned value function and control against the exact optimum. This makes it an ideal test case for verifying that the proposed methods recover the correct structure when the underlying problem is fully understood.

We consider the one-dimensional controlled diffusion given by

$$dX_t = (aX_t + bu_t)\,dt + \sigma\,dW_t, \qquad X_0 = x_0, \tag{4.37}$$

where $X_t \in \mathbb{R}$ denotes the system state, $u_t \in \mathbb{R}$ is the control process, and W_t is a standard Brownian motion. The coefficients a and b govern the intrinsic system dynamics and control effectiveness, respectively, while $\sigma > 0$ represents the magnitude of stochastic disturbances. This formulation is standard in stochastic control and captures a wide range of engineering systems where the state evolves linearly but is subject to noise, and the controller can influence the dynamics through a linear input. The main objective is to minimize a quadratic cost functional of the form

$$J(u) = \mathbb{E}\left[\int_0^T \left(qX_t^2 + ru_t^2\right)\,dt + q_T X_T^2\right], \tag{4.38}$$

where $q, r, q_T > 0$ are weighting coefficients that penalize deviations of the state from zero and the magnitude of the control. The running cost qX_t^2 encourages stabilization of the system, while the control penalty ru_t^2 prevents excessively aggressive control actions. The terminal cost $q_T X_T^2$ enforces stability at the final time. This quadratic structure leads to a well-known trade-off between state regulation and control effort, making the LQ problem a canonical benchmark in both deterministic and stochastic control theory. The value function associated with this problem is

defined by

$$V(t,x) = \inf_{u\in\mathcal{A}} \mathbb{E}\left[\int_t^T \left(qX_s^2 + ru_s^2\right) ds + q_T X_T^2 \,\middle|\, X_t = x\right]. \tag{4.39}$$

Since both the state dynamics and the cost functional are quadratic, it is natural to think of a value function of quadratic form

$$V(t,x) = P(t)x^2 + C(t), \tag{4.40}$$

where $P(t)$ and $C(t)$ are deterministic functions to be which can be found analytically. The presence of the additive term $C(t)$ is necessary because the diffusion term $\sigma\, dW_t$ contributes a state-independent correction to the value through the second derivative term in the HJB equation. Looking at the associated HJB equation is

$$0 = V_t(t,x) + \inf_{u\in\mathbb{R}} \left\{(ax+bu)V_x(t,x) + \frac{1}{2}\sigma^2 V_{xx}(t,x) + qx^2 + ru^2\right\}, \qquad V(T,x) = q_T x^2. \tag{4.41}$$

Using the ansatz (4.40), we obtain

$$V_t(t,x) = \dot{P}(t)x^2 + \dot{C}(t), \qquad V_x(t,x) = 2P(t)x, \qquad V_{xx}(t,x) = 2P(t).$$

Substituting these expressions into (4.41) yields

$$\begin{aligned} 0 &= \dot{P}(t)x^2 + \dot{C}(t) + \inf_{u\in\mathbb{R}} \left\{(ax+bu)\,2P(t)x + \sigma^2 P(t) + qx^2 + ru^2\right\} \\ &= \dot{P}(t)x^2 + \dot{C}(t) + \sigma^2 P(t) + \inf_{u\in\mathbb{R}} \left\{(2aP(t)+q)x^2 + 2bP(t)x\,u + ru^2\right\}. \end{aligned} \tag{4.42}$$

The control-dependent part is a convex quadratic in u, and the first-order optimality condition gives

$$2ru + 2bP(t)x = 0, \qquad \Longrightarrow \qquad u_t^* = -\frac{b}{r}P(t)X_t. \tag{4.43}$$

Substituting this optimizer back into the HJB equation gives

$$\inf_{u\in\mathbb{R}}\{2bP(t)x\,u + ru^2\} = -\frac{b^2}{r}P(t)^2x^2,$$

so that

$$0 = \left[\dot{P}(t) + 2aP(t) - \frac{b^2}{r}P(t)^2 + q\right]x^2 + \left[\dot{C}(t) + \sigma^2 P(t)\right]. \tag{4.44}$$

Since this identity must hold for all x, we obtain the coupled ODE system

$$\dot{P}(t) + 2aP(t) - \frac{b^2}{r}P(t)^2 + q = 0, \qquad P(T) = q_T, \tag{4.45}$$

$$\dot{C}(t) + \sigma^2 P(t) = 0, \qquad C(T) = 0. \tag{4.46}$$

Equivalently, the Riccati equation may be written in backward form as

$$-\dot{P}(t) = 2aP(t) - \frac{b^2}{r}P(t)^2 + q, \qquad P(T) = q_T. \tag{4.47}$$

Thus, the value function is fully characterized by the Riccati solution $P(t)$ together with the diffusion correction $C(t)$, while the optimal control is the linear state feedback given by equation (4.43). In particular, when $\sigma = 0$, the additive term vanishes and the deterministic LQ problem is recovered. This explicit solution provides an exact benchmark for both the learned value function and the learned control policy.

Moving onto the numerical implementation, we fix the time horizon to $T = 1$ and discretize the interval $[0, T]$ using $N = 30$ uniform time steps. Unless otherwise stated, we use the parameter values

$$a = 0.5, \qquad b = 1.0, \qquad \sigma = 0.4, \qquad q = 1.0, \qquad r = 0.5, \qquad q_T = 1.0,$$

with initial state $X_0 = 1.3$. These coefficients define a stable but nontrivial control problem where the positive drift coefficient a makes the uncontrolled state mildly unstable, the control coefficient b allows effective intervention, and the diffusion coefficient σ introduces persistent stochastic disturbance. The quadratic weights q and q_T penalize deviations of the state from the origin both along the trajectory and at maturity, while r regulates the cost of control effort. Under this parameterization, the optimal policy is sufficiently active to reveal meaningful differences across methods, while remaining analytically tractable through the Riccati solution. All expectations are approximated using Monte Carlo simulation with batch size 64, and the transformer-based and FBSDE-based methods are trained under the same discretization and simulation environment (learning rates an schedules) to ensure a fair comparison.

In contrast to the previous experiment, where no closed-form reference was available, the analytical solution provided by the Riccati equation allows us to directly quantify both value function accuracy and control recovery. For each method, we estimate the value function at the initial state X_0 and compare it against the exact value $V(0, X_0) = P(0)X_0^2 + C(0)$, where (P, C) are obtained by numerically solving the coupled ODE system in (4.45)–(4.46). In addition, we evaluate the learned control policies by comparing them against the optimal linear feedback $u_t^* = -\frac{b}{r}P(t)X_t$, allowing us to assess not only value approximation but also structural consistency of the learned control laws.

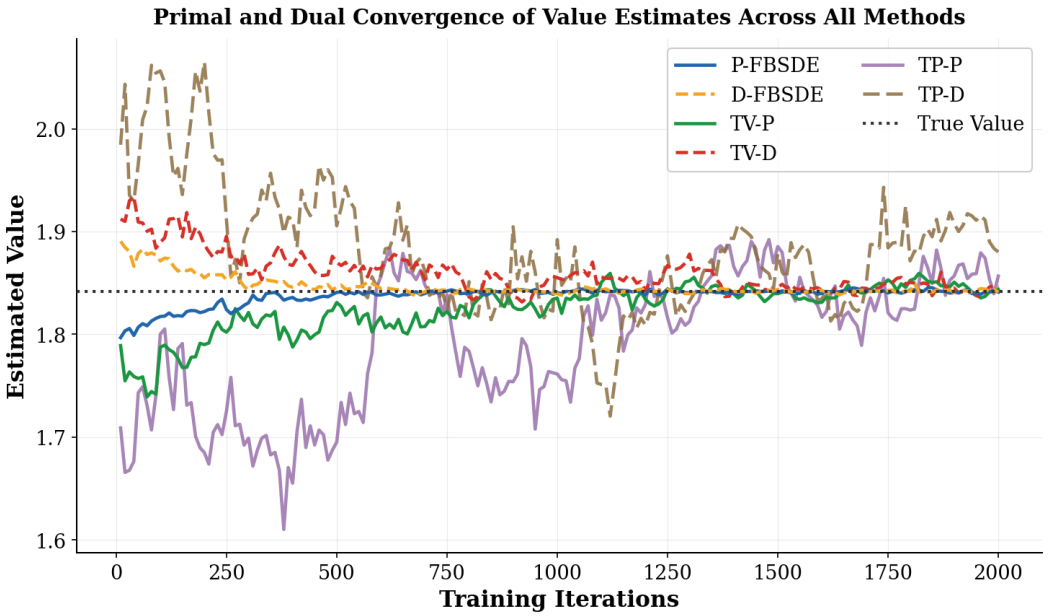

Figure 4.7: **Primal and dual convergence of value estimates across all methods.** Primal estimates approach the true value $V^* = 1.85$ from below, while dual estimates approach from above, forming a shrinking duality gap throughout training. FBSDE methods converge fastest with minimal variance. Value-based transformers (TV) exhibit smooth and stable trajectories, while policy-based transformers (TP) display slower convergence and higher variability.

The results, summarised in Table 4.5 and Figure 4.7, confirm that all methods successfully recover the analytical solution to a high degree of accuracy. As expected in this fully Markovian setting, the Deep FBSDE methods provide the most accurate estimates, both P-FBSDE and D-FBSDE produce primal values of 1.847 and 1.848 respectively against the true $V^* = 1.85$, with duality gaps of 0.006 (Table 4.5). This is consistent with its design, as these methods explicitly exploit the finite-dimensional Markovian structure through the associated backward equation. Among the transformer-based approaches, the value-based formulations (TV-P, TV-D) produce primal estimates of 1.844 and 1.845, with moderate duality gaps of 0.015 and 0.016. The policy-based formulations (TP-P, TP-D) also converge to the correct value level, with primal estimates of 1.836 and 1.838, but exhibit wider duality gaps of 0.039 and 0.042, reflecting the additional degrees of freedom introduced by directly parameterising the control without enforcing value-function consistency. As can be seen in Figure 4.7, this manifests as slower convergence and higher variability throughout training, while the FBSDE and TV methods display smooth and monotone trajectories that settle quickly near the analytical benchmark.

From a control perspective, all methods recover the correct linear feedback structure $u_t \approx -K(t)X_t$, with $K(t) = \frac{b}{r}P(t) = 2.00$. The recovered slopes $\hat{K}$, derived directly from the primal value estimates in Table 4.5, range from 2.00 for the FBSDE methods to 1.98 for TP-P (errors of at most 0.02) confirming that all architectures learn the underlying linear control law rather than merely fitting trajectory outputs. The slight underestimation in terms of $\hat{K}$ in the policy-based methods is consistent with their larger duality gaps which reflects the known approximation bias of direct policy optimization in the absence of value-function constraints.

Table 4.5: Value function estimates and recovered control slope $\hat{K}(0.5)$ for the LQ control problem ($X_0 = 1.3$, $V^* = 1.85$, $K^* = 2.00$).

Method	**Primal**	**Dual**	**Gap**	ΔV	$\hat{K}(0.5)$	ΔK
True Value		$V^* = 1.85$,		$K^* = 2.00$		
P-FBSDE	1.847	1.853	0.006	0.003	2.00	0.00
D-FBSDE	1.848	1.854	0.006	0.002	2.00	0.00
TV-P	1.844	1.859	0.015	0.006	1.99	0.01
TV-D	1.845	1.861	0.016	0.005	1.99	0.01
TP-P	1.836	1.875	0.039	0.014	1.98	0.02
TP-D	1.838	1.880	0.042	0.012	1.99	0.01

Note: $\Delta V = |V^* - V_{\text{primal}}|$. The slope $\hat{K}(0.5) = 2\hat{P}(0)$ is derived from the primal estimate via $\hat{P}(0) = (V_{\text{primal}} - \sigma^2 T)/X_0^2$, and $\Delta K = |K^* - \hat{K}|$. All methods recover the optimal linear feedback structure, with deviations growing monotonically from FBSDE to TV to TP, consistent with the hierarchy of duality gaps.

In conclusion, these results confirm that the proposed transformer-based/ deep BSDE formulations provide a flexible and general-purpose approach to stochastic control beyond UMPs in financial settings. In structured Markovian settings, they recover both the value function and the optimal control with high fidelity, reproducing the same qualitative hierarchy, FBSDE $\succ$ TV $\succ$ TP. This consistency across fundamentally different problem classes highlights the strength of the framework as a unified methodology. The LQ benchmark additionally serves as a structural diagnostic since the optimal policy is known to be exactly linear in the state, any deviation in the learned control directly quantifies approximation error. The tight alignment of the value-based transformer with this linear structure, and the slight but detectable deviation of the policy-based formulation, reinforces the interpretation that enforcing value-function consistency acts as an implicit regulariser.

4.4 Conclusion

In this chapter we developed a unified theoretical and computational framework for stochastic control under two distinct sources of non-Markovianity: latent regime uncertainty and path-dependent long-memory dynamics. These settings introduce fundamentally different departures from the classical Markovian formulation, requiring separate analytical treatments prior to admitting a common algorithmic treatment.

In the hidden-regime setting, state augmentation via the posterior belief process restores Markovianity under the observation filtration, yielding a filtered FBSDE system driven by the innovation process as established in Theorem 4.1.1. This formulation exposes an intrinsic information asymmetry captured by Proposition 4.1.2, which shows that partial observation simultaneously depresses the primal value and inflates the dual bound, thereby widening the duality gap in proportion to informational uncertainty. On the other hand in the case of the path-dependent setting, the functional HJB system of Theorem 4.1.10, together with the BSDE and adjoint characterizations of Theorems 4.1.11–4.1.12, provides a complete optimality theory on path space via Dupire's functional Itô calculus. Proposition 4.1.13 unifies these results into a coupled path-dependent regime-switching FBSDE system, establishing the natural analogue of the Markovian characterization while making explicit the central computational difficulty that the effective state is infinite-dimensional and does not admit a finite-dimensional reduction without loss of optimality.

To address this, we introduced transformer-based deep BSDE algorithms that replace time-indexed state-based networks with a single causal transformer operating on trajectory prefixes. Causal masking enforces non-anticipativity at the architectural level, providing an algorithmic analogue of measurability in the functional Itô framework. The value-based and policy-based formulations are equivalent at convergence but exhibit distinct structural properties where the former enforces HJB consistency during training and promotes tighter primal–dual alignment, while the latter directly optimizes the control objective and is particularly well-suited to settings where value function approximation on path space is numerically unstable. The numerical experiments across four problem classes validate these structural and computational properties. In Markovian benchmarks, transformer-based methods recover solutions consistent with Deep FBSDE references, confirming correctness while highlighting the computational advantage of classical methods when their structural assumptions hold. In the rough-volatility setting, the transformer ar-

chitecture successfully captures long-range temporal dependence and accurately approximates both value and control across a wide range of roughness parameters without explicit state augmentation. The path-dependent drift and LQ benchmarks further demonstrate that the framework generalizes beyond financial applications, recovering known structural properties such as linear feedback laws.

In this chapter, we establish transformer-based deep BSDE methods as a scalable approach for stochastic control on path space. By overcoming the curse of dimensionality inherent in functional state representations, the framework extends the reach of stochastic control beyond the limitations of classical DPP and finite-dimensional BSDE solvers, providing a unified methodology for high-dimensional, path-dependent, and partially observed control problems.

References

Cont, Rama and David-Antoine Fournié (2013). "Functional Itô calculus and functional Kolmogorov equations". In: *The Annals of Probability* 41.1, pp. 109–133.

Dupire, Bruno (2009). "Functional Itô calculus". In: *Bloomberg Portfolio Research Paper*.

Ekren, Ibrahim, Nizar Touzi, and Jianfeng Zhang (2014). "Viscosity Solutions of Fully Nonlinear Parabolic Path Dependent PDEs: Part I". In: *Annals of Probability* 42.1, pp. 204–236.

– (2016). "Viscosity solutions of fully nonlinear parabolic path dependent PDEs: Part II". In: *The Annals of Probability* 44.4, pp. 2507–2553. DOI: 10.1214/15-AOP1027. URL: https://doi.org/10.1214/15-AOP1027.

Gatheral, Jim, Thibault Jaisson, and Mathieu Rosenbaum (2018). "Volatility Is Rough". In: *Quantitative Finance* 18.6, pp. 933–949.

Han, Bingyan and Hoi Ying Wong (2021). "Merton's portfolio problem under Volterra Heston model". In: *Finance Research Letters* 39, p. 101580. ISSN: 1544-6123. DOI: https://doi.org/10.1016/j.frl.2020.101580. URL: https://www.sciencedirect.com/science/article/pii/S1544612319312917.

Kim, Tong Suk and Edward Omberg (1996). "Dynamic Nonmyopic Portfolio Behavior". In: *The Review of Financial Studies* 9.1, pp. 141–161. ISSN: 08939454, 14657368. URL: http://www.jstor.org/stable/2962368 (visited on 03/25/2026).

Krishnamurthy, Vikram, Elisabeth Leoff, and Jörn Sass (2018). "Filterbased stochastic volatility in continuous-time hidden Markov models". In: *Econometrics and Statistics* 6. STATISTICS OF EXTREMES AND APPLICATIONS, pp. 1–21. ISSN: 2452-3062. DOI: https://doi.org/10.1016/j.ecosta.2016.10.

007. URL: https://www.sciencedirect.com/science/article/pii/S2452306216300144.

Lim, Andrew E. B. and Xun Yu Zhou (2001). "Linear-Quadratic Control of Backward Stochastic Differential Equations". In: *SIAM Journal on Control and Optimization* 40.2, pp. 450–474.

Liptser, R. S. and A. N. Shiryaev (1977). *Statistics of Random Processes I: General Theory*. 1st ed. Applications of Mathematics. Berlin: Springer. DOI: 10.1007/978-3-662-12644-3.

Liu, Jun (2007). "Portfolio Selection in Stochastic Environments". In: *The Review of Financial Studies* 20.1, pp. 1–39. ISSN: 08939454, 14657368. URL: http://www.jstor.org/stable/4123484 (visited on 03/25/2026).

Peng, Shige (1992). "Stochastic Hamilton-Jacobi-Bellman Equations". In: *SIAM Journal on Control and Optimization* 30.2, pp. 284–304.

Pham, Huyên (2009). "The classical PDE approach to dynamic programming". In: *Continuous-time Stochastic Control and Optimization with Financial Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 37–60. ISBN: 978-3-540-89500-8. DOI: 10.1007/978-3-540-89500-8_3. URL: https://doi.org/10.1007/978-3-540-89500-8_3.

Wiedermann, Kristof (2022). *An SMP-Based Algorithm for Solving the Constrained Utility Maximization Problem via Deep Learning*. arXiv: 2202.07771 [q-fin.CP]. URL: https://arxiv.org/abs/2202.07771.

Chapter 5

ROBUST UTILITY MAXIMIZATION VIA DUAL MIN–MAX

5.1 Motivation and Setup

So far the focus of this thesis is to work with UMPs which operates under a single, known probability measure that governs the dynamics of both the underlying assets as well as the regime-switching process. This very fundamental assumption allows us to work and develop all the theory and numerical methods developed so far, while implying that the investor has the precise knowledge of the drift and volatility structure of the market. In practice however, the financial conditions are subject to Knightian uncertainty (Watkins, 1922), where the true probability distribution governing asset dynamics are unknown and can't be reliably estimated from historical data alone. In the most robust sense, one would need to seek a control that performs well across a family of plausible models rather than under a fixed probability measure, in order to guard against the possibility that the true market reality differs from any single reference specification. In this chapter, we are going to therefore deal with a form of UMP named robust UMP in a regime-switching environment, where we acknowledge the fact that the true probability measure that governs stochastic dynamics are unknown but may belong to a family of plausible measures $\mathcal{P}$. In essence, a robust UMP which is maximizing the terminal utility of wealth takes the form of a min-max problem given by

$$\sup_{\pi\in\Pi}\inf_{\mathbb{P}\in\mathcal{P}}\mathbb{E}_{\mathbb{P}}\left[U(X_T^{\pi})\right] \tag{5.1}$$

where we are maximizing the terminal utility U against the worst case probability measure selected by an adversarial nature. We construct $\mathcal{P}$ by varying the drift μ and diffusion σ of the asset dynamics within admissible process classes $\mathcal{M}$ and $\mathcal{S}$, so that $\mathcal{P} := \{\mathbb{P}_{\mu,\sigma} \mid (\mu,\sigma) \in \mathcal{M}\times\mathcal{S}\}$. The admissible drift class $\mathcal{M}$ and admissible diffusion class $\mathcal{S}$ are defined as

$$\mathcal{M} := \left\{\mu \in \mathbb{F}\text{-predictable} : \mu_{i,\epsilon} \in [\mu^0_{i,\epsilon} - \delta_\mu,\, \mu^0_{i,\epsilon} + \delta_\mu],\ \forall\, i = 1,\dots,d,\ \epsilon \in S\right\}, \tag{5.2}$$

$$\mathcal{S} := \left\{\sigma \in \mathbb{F}\text{-predictable} : \sigma_{i,\epsilon} \in [\sigma^0_{i,\epsilon} - \delta_\sigma,\, \sigma^0_{i,\epsilon} + \delta_\sigma],\ \sigma\sigma^\top \succ 0,\ \forall\, i = 1,\dots,d,\ \epsilon \in S\right\}, \tag{5.3}$$

where $\delta_\mu, \delta_\sigma > 0$ are fixed ambiguity radii and (μ^0, σ^0) are the coefficients associated with the reference measure $\mathbb{P}_0$. Both $\mathcal{M}$ and $\mathcal{S}$ are convex and compact in the

topology of uniform convergence on $[0,T] \times S$. It is essentially the compactness of $\mathcal{M} \times \mathcal{S}$ which ensures that the robust UMP in equation (5.1) is well-posed without requiring any additional penalty regularization, following the classical approach of (Alexander and Ching-Tang, 2005; Gundel, 2005; Tevzadze, Toronjadze, and Uzunashvili, 2013). Both the parameters δ_μ and δ_σ govern the degree of model ambiguity. As $\delta_\mu, \delta_\sigma \to 0$ the adversary is forced back to the reference model (μ^0, σ^0) and the robust value function V^{rob} collapses to the non-robust value function V which we solved in chapters before. Furthermore, as the value of $\delta_\mu, \delta_\sigma$ increases the investor faces greater uncertainty and the robust value declines. Rather than approaching the problem from the primal min–max perspective as in (Krach, Teichmann, and Wutte, 2025), in this chapter we are going to employ the Legendre–Fenchel transform to obtain a dual robust function $\Psi^{\text{rob}}(Y, \mu, \sigma)$ that depends jointly on a dual process Y and the adversarial parameters (μ, σ). The dual robust UMP then takes the form of a min–max problem in the dual space, $\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\text{robust}}(Y, \mu, \sigma)$, in which the dual investor minimizes the worst-case dual cost over all admissible dual processes, while the adversary simultaneously maximizes over all plausible market specifications. The theory and algorithm to solve the robust UMP via duality is developed in this chapter via four stages. First, in Section 5.2, we formalize the robust UMP in the regime-switching setting, derive the associated robust HJB system, and then establish a verification theorem for the saddle-point solution. Second, in Section 5.3, we derive the dual robust function via the Legendre–Fenchel transform, establish weak duality, and prove formally that the min–max equality

$$\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\text{robust}}(Y, \mu, \sigma) = \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \inf_{Y \in \mathbb{J}} \Psi^{\text{robust}}(Y, \mu, \sigma) \tag{5.4}$$

holds under appropriate saddle-point conditions. Third, in Section 5.4, we derive the FBSDE representation of the saddle-point solution, establishing the connection to the primal–dual FBSDE algorithms in Chapter 2. Finally, in Section 5.5, we develop a deep learning based dual min–max algorithm in which two NN, one parameterizing the dual process Y_η and one parameterizing the adversarial market (μ^ϕ, σ^ϕ), are trained against each other on the dual min–max objective $\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\text{robust}}(Y, \mu, \sigma)$. Throughout, the learning process the duality gap serves as a computable quantity to indicate near-optimality for both the investor and the adversary simultaneously, extending the adaptive gap framework from Chapter 3 to the robust setting.

5.2 Robust UMP under Regime Switching

Lets start by formalizing the robust UMP in a regime-switching setting. For this we retain the market model in equation (2.1) with some d risky assets and a finite-state continuous-time Markov chain $\epsilon_t \in S$ with generator $Q = (q_{ij})_{i,j\in S}$. Under a reference measure $\mathbb{P}_0$, the asset dynamics take the familiar form

$$dS_i(t) = S_i(t)\left(\mu^0_{i,\epsilon(t)}(t)\,dt + \sigma^0_{i,\epsilon(t)}(t)\,dW^\top(t)\right), \qquad i = 1,\dots,d, \tag{5.5}$$

where μ^0 and σ^0 denote the reference drift and diffusion coefficients under $\mathbb{P}_0$. In order to introduce model ambiguity/uncertainty, we will allow the adversary to perturb these coefficients by replacing μ^0 and σ^0 with progressively measurable processes $\mu \in \mathcal{M}$ and $\sigma \in \mathcal{S}$, where $\mathcal{M}$ and $\mathcal{S}$ denote the admissible classes of $\mathbb{F}$-predictable, $\mathbb{R}^d$- and $\mathbb{R}^{d\times d}$-valued processes respectively. Under the perturbed measure $\mathbb{P}_{\mu,\sigma}$, the asset dynamics become

$$dS_i(t) = S_i(t)\left(\mu_{i,\epsilon(t)}(t)\,dt + \sigma_{i,\epsilon(t)}(t)\,dW^\top(t)\right), \qquad i = 1,\dots,d, \tag{5.6}$$

and the corresponding wealth process $X^{\pi,\mu,\sigma}_t$ under portfolio policy $\pi \in \mathcal{A}$ satisfies

$$dX^{\pi,\mu,\sigma}_t = \alpha\big(t,\epsilon_t,X^{\pi,\mu,\sigma}_t,\mu_t,\sigma_t,\pi_t\big)\,dt + \beta^\top\big(t,\epsilon_t,X^{\pi,\mu,\sigma}_t,\sigma_t,\pi_t\big)\,dW(t), \tag{5.7}$$

where the drift and diffusion coefficients α and β are obtained from the self-financing condition as in equation (2.5), but now evaluated under the perturbed coefficients (μ_t,σ_t) rather than some reference values (μ^0_t,σ^0_t). The robust value function is then

$$V^{\text{rob}}(t,\epsilon,x) = \sup_{\pi\in\mathcal{A}}\inf_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} J^\pi(t,\epsilon,x;\mu,\sigma) \tag{5.8}$$

for $(t,\epsilon,x) \in [0,T]\times S\times\mathbb{R}$, where

$$J^\pi(t,\epsilon,x;\mu,\sigma) := \mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\int_t^T f(s,\epsilon_s,X^{\pi,\mu,\sigma}_s,\pi_s)\,ds + g(\epsilon_T,X^{\pi,\mu,\sigma}_T)\right],$$

and f and g are the running and terminal utility functions from Definition 2.1.3. To characterize the robust value function $V^{\text{rob}}(t,\epsilon,x)$, we derive the associated HJB system via dynamic programming. The key structural difference relative to the non-robust HJB is that the supremum over controls π and the infimum over adversarial parameters (μ,σ) now appear simultaneously inside the generator, yielding a coupled min–max HJB system. We proceed under the same regularity assumptions as in Theorem 2.2.4, namely that $V^{\text{rob}} \in C^{1,2}([0,T)\times\mathbb{R})$ for each regime $\epsilon_i \in S$, and that all relevant integrability conditions hold.

Theorem 5.2.1. Under suitable regularity and integrability conditions, the robust value function $V^{\text{rob}}(t, \epsilon_i, x)$ satisfies the following coupled min–max HJB system where for each $\epsilon_i \in S$ and $(t, x) \in [0, T) \times \mathbb{R}$,

$$\begin{aligned} -\frac{\partial V^{\text{rob}}}{\partial t}(t, \epsilon_i, x) = \sup_{\pi \in K} \inf_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} \left\{ f(t, \epsilon_i, x, \pi) + \mathcal{L}^{\pi,\mu,\sigma} V^{\text{rob}}(t, \epsilon_i, x) \right\} \\ + \sum_{j \in S} q_{ij} \left(V^{\text{rob}}(t, \epsilon_j, x) - V^{\text{rob}}(t, \epsilon_i, x) \right), \end{aligned} \tag{5.9}$$

with terminal condition $V^{\text{rob}}(T, \epsilon_i, x) = g(\epsilon_i, x)$ for all $x \in \mathbb{R}$ and $\epsilon_i \in S$. Here, $\mathcal{L}^{\pi,\mu,\sigma}$ denotes the differential generator of the wealth process under control π and perturbed coefficients (μ, σ), defined by

$$\mathcal{L}^{\pi,\mu,\sigma} V^{\text{rob}}(t, \epsilon_i, x) := \alpha(t, \epsilon_i, x, \mu, \sigma, \pi) \frac{\partial V^{\text{rob}}}{\partial x}(t, \epsilon_i, x) + \frac{1}{2} |\beta(t, \epsilon_i, x, \sigma, \pi)|^2 \frac{\partial^2 V^{\text{rob}}}{\partial x^2}(t, \epsilon_i, x). \tag{5.10}$$

Proof. The structure of the proof is the same as we did in Theorem 2.2.4, but extended to account for the simultaneous optimization over π and (μ, σ). We start by fixing $(t, \epsilon_i, x) \in [0, T) \times S \times \mathbb{R}$ and let $h > 0$ be small. Define the stopping time $\tau_h := \tau_{t,\epsilon_i,x,1} \wedge (t + h)$. Applying the robust dynamic programming principle (Guo et al., 2022) to the penalized objective (5.8) yields

$$\begin{aligned} V^{\text{rob}}(t, \epsilon_i, x) = \sup_{\pi \in \mathcal{A}} \inf_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} \mathbb{E}\Bigg[\int_t^{\tau_h} \left\{ f(s, \epsilon_s, X_s^{\pi,\mu,\sigma}, \pi_s) \right\} ds \\ + V^{\text{rob}}(\tau_h, \epsilon_{\tau_h}, X_{\tau_h}^{\pi,\mu,\sigma}) \Bigg]. \end{aligned} \tag{5.11}$$

Since $V^{\text{rob}} \in C^{1,2}([0, T) \times \mathbb{R})$ for each regime $\epsilon_i \in S$, we apply the Itô–Dynkin formula to $V^{\text{rob}}(s, \epsilon_s, X_s^{\pi,\mu,\sigma})$ on $[t, \tau_h]$, exactly as in the proof of Theorem 2.2.4, accounting for the regime-switching jump martingale $dM^{\epsilon}(s)$ and the perturbed generator $\mathcal{L}^{\pi,\mu,\sigma}$. Taking conditional expectations and invoking the zero-mean martingale properties of the Itô integral and M^{ϵ} as established in Remarks 2.2.5 and 2.2.6, we obtain

$$\begin{aligned} \mathbb{E}\left[V^{\text{rob}}(\tau_h, \epsilon_{\tau_h}, X_{\tau_h}^{\pi,\mu,\sigma}) \right] = V^{\text{rob}}(t, \epsilon_i, x) \\ + \mathbb{E}\left[\int_t^{\tau_h} \left(\frac{\partial V^{\text{rob}}}{\partial s} + \mathcal{L}^{\pi,\mu,\sigma} V^{\text{rob}} + \sum_{j \in S} q_{ij} (V_j^{\text{rob}} - V^{\text{rob}}) \right) ds \right]. \end{aligned} \tag{5.12}$$

Substituting (5.12) into (5.11), subtracting $V^{\text{rob}}(t, \epsilon_i, x)$ from both sides, dividing by h, and letting $h \to 0$ via the dominated convergence theorem under the localization and boundedness conditions inherited from Theorem 2.2.4, we obtain

$$\begin{aligned} 0 = \sup_{\pi \in K} \inf_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} & \left\{ f(t, \epsilon_i, x, \pi) + \mathcal{L}^{\pi,\mu,\sigma} V^{\text{rob}}(t, \epsilon_i, x) \right\} \\ & + \frac{\partial V^{\text{rob}}}{\partial t}(t, \epsilon_i, x) + \sum_{j \in S} q_{ij} \left(V^{\text{rob}}(t, \epsilon_j, x) - V^{\text{rob}}(t, \epsilon_i, x) \right). \end{aligned} \tag{5.13}$$

which is precisely the min–max HJB in equation (5.9). The terminal condition $V^{\text{rob}}(T, \epsilon_i, x) = g(\epsilon_i, x)$ follows from the definition of the value function at time T, exactly as in the non-robust case. This completes the proof. □

The min–max HJB (5.9) differs from the non-robust HJB in two ways. First, the supremum over controls π and the infimum over adversarial parameters (μ, σ) appear jointly inside the generator, rather than as a single supremum. This coupling means that the optimal control π^* and the worst-case model (μ^*, σ^*) must be determined simultaneously at each (t, ϵ_i, x), rather than sequentially. Second, the infimum over (μ, σ) is now constrained to the compact admissible set $\mathcal{M} \times \mathcal{S}$, rather than being taken over a single fixed model. The adversary therefore has to select the worst-case coefficients from within a bounded neighborhood of the reference model (μ^0, σ^0), and the optimal adversarial perturbation (μ^*, σ^*) is characterized as the constrained minimizer of the HJB generator over $\mathcal{M} \times \mathcal{S}$, which we formalize in the verification theorem below.

Theorem 5.2.2. Let $w(t, \epsilon_i, x)$ be a candidate value function $\in C^{1,2}([0, T) \times \mathbb{R}) \cap C([0, T] \times \mathbb{R})$ such that for every $\epsilon_i \in S$:

1. It satisfies the **robust HJB inequality** for all $(t, x) \in [0, T) \times \mathbb{R}$:

$$\begin{aligned} -\frac{\partial w}{\partial t}(t, \epsilon_i, x) \geq \sup_{\pi \in K} \inf_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} & \left\{ f(t, \epsilon_i, x, \pi) + \mathcal{L}^{\pi,\mu,\sigma} w(t, \epsilon_i, x) \right\} \\ & + \sum_{j \in S} q_{ij} \left(w(t, \epsilon_j, x) - w(t, \epsilon_i, x) \right). \end{aligned}$$

2. It satisfies the **terminal condition** for all $x \in \mathbb{R}$:

$$w(T, \epsilon_i, x) \geq g(\epsilon_i, x).$$

3. It satisfies a **polynomial growth condition**: there exist constants $C > 0$ and $p \geq 1$ such that

$$|w(t, \epsilon, x)| \leq C(1 + |x|^p), \quad \forall\, (t, x, \epsilon) \in [0, T] \times \mathbb{R} \times S.$$

Furthermore, assume there exists a saddle point $(\pi^*, \mu^*, \sigma^*) \in \mathcal{A} \times \mathcal{M} \times \mathcal{S}$ such that almost surely for almost every $s \in [t, T]$,

$$(\pi^*(s), \mu^*(s), \sigma^*(s)) \in \arg\sup_{\pi \in K} \arg\inf_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Big\{ f(s, \epsilon_s, X_s^{\pi^*,\mu^*,\sigma^*}, \pi) + \mathcal{L}^{\pi,\mu,\sigma} w(s, \epsilon_s, X_s^{\pi^*,\mu^*,\sigma^*}) \Big\},$$

and that the state process X^{π^*,μ^*,σ^*} satisfies $\mathbb{E}[\sup_{s \in [t,T]} |X_s^{\pi^*,\mu^*,\sigma^*}|^p] < \infty$. Then:

1. For any admissible $(\pi, \mu, \sigma) \in \mathcal{A} \times \mathcal{M} \times \mathcal{S}$, the candidate function w dominates the robust performance functional:

$$w(t, \epsilon_i, x) \geq \inf_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \left\{ \mathbb{E}_{\mathbb{P}_{\mu,\sigma}} \left[\int_t^T f(s, \epsilon_s, X_s^{\pi,\mu,\sigma}, \pi_s)\, ds + g(\epsilon_T, X_T^{\pi,\mu,\sigma}) \right] \right\}.$$

2. The saddle point (π^*, μ^*, σ^*) is optimal, and w coincides with the true robust value function:

$$\begin{aligned} V^{\text{rob}}(t, \epsilon_i, x) &= w(t, \epsilon_i, x) \\ &= \inf_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \left\{ \mathbb{E}_{\mathbb{P}_{\mu^*,\sigma^*}} \left[\int_t^T f(s, \epsilon_s, X_s^{\pi^*,\mu^*,\sigma^*}, \pi_s^*)\, ds + g(\epsilon_T, X_T^{\pi^*,\mu^*,\sigma^*}) \right] \right\} \end{aligned}$$

Proof. Again the proof mirrors the structure of the verification theorem for the non-robust case (Theorem 2.2.7). Fix $(t, \epsilon_i, x) \in [0, T] \times S \times \mathbb{R}$ and let $(\pi, \mu, \sigma) \in \mathcal{A} \times \mathcal{M} \times \mathcal{S}$ be any admissible triple. Define the sequence of stopping times $(\tau_n)_{n \geq 1}$ as in equation (2.32), localizing the stochastic integrals on $[t, \tau_n]$. Applying the Itô–Dynkin formula to $w(s, \epsilon_s, X_s^{\pi,\mu,\sigma})$ on $[t, \tau_n]$ under the perturbed measure $\mathbb{P}_{\mu,\sigma}$, and taking conditional expectations while invoking the zero-mean martingale properties of Remarks 2.2.5 and 2.2.6, yields

$$\mathbb{E}\big[w(\tau_n, \epsilon_{\tau_n}, X_{\tau_n}^{\pi,\mu,\sigma})\big] = w(t, \epsilon_i, x) + \mathbb{E}\left[\int_t^{\tau_n} \left(\frac{\partial w}{\partial s} + \mathcal{L}^{\pi,\mu,\sigma} w + \sum_{j \in S} q_{ij}(w_j - w) \right) ds \right] \tag{5.14}$$

By the robust HJB inequality satisfied by w, we have for all admissible (π, μ, σ),

$$\frac{\partial w}{\partial s} + \mathcal{L}^{\pi,\mu,\sigma} w + \sum_{j \in S} q_{ij}(w_j - w) \leq -f(s, \epsilon_s, X_s^{\pi,\mu,\sigma}, \pi_s) \tag{5.15}$$

Substituting this into (5.14) and rearranging gives

$$\begin{aligned} \mathbb{E}\big[w(\tau_n, \epsilon_{\tau_n}, X_{\tau_n}^{\pi,\mu,\sigma})\big] \leq\ & w(t, \epsilon_i, x) \\ & - \mathbb{E}\left[\int_t^{\tau_n} \left\{ f(s, \epsilon_s, X_s^{\pi,\mu,\sigma}, \pi_s) \right\} ds\right] \end{aligned} \tag{5.16}$$

By letting $n \to \infty$ via the polynomial growth condition on w and the dominated convergence theorem and applying the terminal condition $w(T, \epsilon, x) \geq g(\epsilon, x)$, we obtain

$$w(t, \epsilon_i, x) \geq \mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\int_t^T f(s, \epsilon_s, X_s^{\pi,\mu,\sigma}, \pi_s)\, ds + g(\epsilon_T, X_T^{\pi,\mu,\sigma})\right]$$

Taking the infimum over $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$ and then the supremum over $\pi \in \mathcal{A}$ establishes Part (1) of the theorem. Now for the second claim, under the saddle-point triple (π^*, μ^*, σ^*), the robust HJB inequality holds with equality along the optimal trajectory $(X_s^{\pi^*,\mu^*,\sigma^*}, \epsilon_s)$. Repeating the same argument but replacing the inequality with an equality at each step, and using the terminal condition $w(T, \epsilon, x) = g(\epsilon, x)$ which holds with equality at the saddle point, yields

$$\begin{aligned} w(t, \epsilon_i, x) &= \mathbb{E}_{\mathbb{P}_{\mu^*,\sigma^*}}\left[\int_t^T f(s, \epsilon_s, X_s^{\pi^*,\mu^*,\sigma^*}, \pi_s^*)\, ds + g(\epsilon_T, X_T^{\pi^*,\mu^*,\sigma^*})\right] \\ &= V^{\text{rob}}(t, \epsilon_i, x). \end{aligned}$$

which establishes that w coincides with the true robust value function and that (π^*, μ^*, σ^*) is a saddle-point optimal triple. □

Theorem 5.2.2 establishes essentially sufficiency where any smooth candidate function satisfying the robust HJB inequality and the terminal condition dominates the robust performance function, and equality is attained at the saddle point. But in order to obtain the candidate w, we are going to derive FBSDE representation and then use NN to numerically solve it. For that we are next going to first derive the dual robust UMP.

5.3 Dual Robust UMP

In this section we now derive the dual representation of the robust UMP (5.8) by applying the Legendre–Fenchel duality transform to the robust objective. The key

distinction relative to the non-robust dualization is that the robust objective again involves an additional infimum over adversarial parameters $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, which interacts non-trivially with the Legendre–Fenchel transform. To handle this properly, we will proceed in two stages. In the first stage, we dualize the primal optimization over admissible wealth processes $X \in \mathbb{I}_1$ treating the adversarial parameters (μ, σ) as fixed. This yields a dual robust functional that depends parametrically on (μ, σ). In the second stage, we restore the adversarial optimization by taking the supremum of the dual functional over $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, yielding the dual min–max problem. This two-stage construction preserves the convex-analytic structure of the non-robust dual while incorporating the adversarial layer in a natural and tractable manner.

Lets start with the introduction of the enlarged space $\mathbb{I}_1$ of feasible wealth processes and the admissible dual process class $\mathbb{J}$ as defined before in section 2.4 but now to depend parametrically on the adversarial coefficients (μ, σ). In particular, all admissible processes $X \in \mathbb{I}_1$ and dual processes $Y \in \mathbb{J}$ are assumed to be progressively measurable and integrable under $\mathbb{P}_{\mu,\sigma}$. Under these assumptions, the expectation operator $\mathbb{E}_{\mathbb{P}_{\mu,\sigma}}[\cdot]$ is well-defined and the Legendre–Fenchel transform can be applied pathwise under $\mathbb{P}_{\mu,\sigma}$ for each fixed (μ, σ), justifying the first stage of the construction. Therefore, for some fixed $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, the robust primal objective can be written as a minimization problem over $\mathbb{I}_1$ via the penalty function

$$\Phi^{\mu,\sigma}(X) := \Phi_1^{\mu,\sigma}(X) - \mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\int_t^T f(s, \epsilon_s, X_s, \pi_s)\, ds + g(\epsilon_T, X_T)\right] \tag{5.17}$$

where $\Phi_1^{\mu,\sigma}$ is the admissibility penalty (equation (2.56)) evaluated under the perturbed dynamics using equation (5.7). The robust primal value function then satisfies

$$-V^{\text{rob}}(t, \epsilon, x) = \sup_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} \inf_{X\in\mathbb{I}_1} \Phi^{\mu,\sigma}(X).$$

Now, for a fixed (μ, σ), we apply the Legendre–Fenchel transform to $\Phi^{\mu,\sigma}$, thereby introducing a dual process $Y \in \mathbb{J}$ and computing the convex conjugate with respect to the wealth process X. By the same sequence of steps as in the derivation of (equation (2.59)), the Legendre–Fenchel transform of $\Phi^{\mu,\sigma}$ with respect to X yields the *robust dual functional*

$$\begin{aligned}\Psi^{\text{rob}}(Y, \mu, \sigma) := xY_t &+ \mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\sup_{x\geq 0}\left\{g(\epsilon_T, x) - xY_T\right\}\right] \\ &+ \mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\int_t^T \sup_{\substack{x\geq 0\\ \pi\in K}}\left\{-f_s + x\dot{Y}_s + \alpha_s Y_s + \beta_s^\top \Lambda_s^Y\right\} ds\right],\end{aligned} \tag{5.18}$$

where $f_s := f(s, \epsilon_s, x, \pi)$, $\alpha_s := \alpha(s, \epsilon_s, x, \mu_s, \sigma_s, \pi)$, and $\beta_s := \beta(s, \epsilon_s, x, \sigma_s, \pi)$ and where $Y \in \mathbb{J}$ is a progressively measurable, $\mathbb{R}_+$-valued semimartingale with dynamics

$$dY_s = \dot{Y}_s\, ds + (\Lambda_s^Y)^\top dW_s + \sum_{\substack{\epsilon_i, \epsilon_j \in S \\ i \neq j}} \Gamma_{ij}^Y(s)\, dM_{ij}(s), \tag{5.19}$$

as in the non-robust dual formulation of Section 2.4, but now evaluated under the perturbed measure $\mathbb{P}_{\mu,\sigma}$. Furthermore, for each (s, ω) and fixed (Y, μ, σ), the mapping

$$(x, \pi) \;\mapsto\; -f(s, \epsilon_s, x, \pi) + x\dot{Y}_s + \alpha(s, \epsilon_s, x, \mu_s, \sigma_s, \pi)\, Y_s + \beta(s, \epsilon_s, x, \sigma_s, \pi)^\top \Lambda_s^Y$$

is assumed to be jointly concave in (x, π) on $\mathbb{R}_+ \times K$, with K compact and standard growth and integrability conditions ensuring the supremum is finite and attained for each (s, ω). The pointwise supremum therefore yields a measurable integrand, so that $\Psi^{\text{rob}}(Y, \mu, \sigma)$ is properly defined. Moreover, as $\mathcal{M} \times \mathcal{S}$ is compact by definition and $\Psi^{\text{rob}}(Y, \mu, \sigma)$ is continuous in (μ, σ) for each fixed $Y \in \mathbb{J}$, the adversarial optimization problem $\sup_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} \Psi^{\text{rob}}(Y, \mu, \sigma)$ is well-posed by the extreme value theorem. In particular, the supremum is also finite and attained for each $Y \in \mathbb{J}$.

The robust dual min–max problem is then obtained by restoring the adversarial optimization. Taking the supremum of $\Psi^{\text{rob}}(Y, \mu, \sigma)$ over $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$ and the infimum over $Y \in \mathbb{J}$ yields

$$\inf_{Y\in\mathbb{J}} \sup_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} \Psi^{\text{rob}}(Y, \mu, \sigma). \tag{5.20}$$

This formulation admits a natural game-theoretic interpretation. The dual process $Y \in \mathbb{J}$ acts as a stochastic discount factor or supermartingale deflator, selected by the dual investor to minimize the worst-case dual cost. The adversarial parameters $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$ act as nature, selecting the worst-case model from within the compact admissible set to maximize the same objective. The compactness of $\mathcal{M}\times\mathcal{S}$ ensures that the adversary cannot select arbitrarily extreme dynamics, playing the role here that the quadratic penalty played in (Krach, Teichmann, and Wutte, 2025). This zero-sum stochastic differential game structure forms the basis for the NN algorithm developed in this chapter. We now establish the fundamental weak duality relation between the primal robust value function and the dual min–max problem.

Theorem 5.3.1. For any initial condition $(t, x, \epsilon) \in [0, T] \times \mathbb{R} \times S$, the following weak duality inequality holds:

$$-V^{\mathrm{rob}}(t, \epsilon, x) \geq \inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\mathrm{rob}}(Y, \mu, \sigma). \tag{5.21}$$

Equivalently,

$$V^{\mathrm{rob}}(t, \epsilon, x) \leq -\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\mathrm{rob}}(Y, \mu, \sigma).$$

Proof. Fix any admissible primal triple $(\pi, \mu, \sigma) \in \mathcal{A} \times \mathcal{M} \times \mathcal{S}$ and any dual process $Y \in \mathbb{J}$. Applying Itô's formula to the product $X_s^{\pi,\mu,\sigma} Y_s$ under the perturbed measure $\mathbb{P}_{\mu,\sigma}$, and proceeding exactly as in the proof of the non-robust weak duality theorem in chapter 2, we obtain the identity

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[X_T^{\pi,\mu,\sigma} Y_T\right] - x Y_t \\ &\quad = \mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\int_t^T \left(X_s^{\pi,\mu,\sigma} \dot{Y}_s + Y_s \alpha_s^{\pi,\mu,\sigma} + (\beta_s^{\sigma,\pi})^\top \Lambda_s^Y\right) ds\right], \end{aligned} \tag{5.22}$$

where the martingale terms vanish in expectation under the standard integrability conditions. Applying the Legendre–Fenchel inequality to the terminal utility g and the running utility f as in equations (2.63) and (2.64), and combining with (5.22), yields the primal–dual inequality

$$\mathbb{E}_{\mathbb{P}_{\mu,\sigma}}\left[\int_t^T f(s, \epsilon_s, X_s^{\pi,\mu,\sigma}, \pi_s)\, ds + g(\epsilon_T, X_T^{\pi,\mu,\sigma})\right] \leq -\Psi^{\mathrm{rob}}(Y, \mu, \sigma) + \Phi^{\mu,\sigma}(X). \tag{5.23}$$

By construction of the admissibility penalty $\Phi_1^{\mu,\sigma}$, we have that $\Phi_1^{\mu,\sigma}(X) = 0$ for any admissible wealth process $X \in \mathbb{I}_1$, while $\Phi_1^{\mu,\sigma}(X) = +\infty$ otherwise. Consequently, for admissible X, the function $\Phi^{\mu,\sigma}(X)$ reduces to the negative of the primal objective $J^\pi(t, \epsilon, x; \mu, \sigma)$. By taking the infimum over (μ, σ) on the left and the supremum over (μ, σ) inside Ψ^{rob} on the right, then taking the supremum over $\pi \in \mathcal{A}$ and the infimum over $Y \in \mathbb{J}$, we obtain

$$V^{\mathrm{rob}}(t, \epsilon, x) \leq -\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\mathrm{rob}}(Y, \mu, \sigma),$$

which establishes the equation (5.21) and completing the proof. □

Weak duality establishes that the dual min–max provides an upper bound on the robust value function. We now address the stronger question of when equality holds, that is, when the min–max and maximin coincide and the duality gap vanishes. We

establish this result under structural conditions that are standard in convex-concave min–max theory (Rockafellar, 1970; Sion, 1958) and are satisfied in the regime-switching setting under compact admissible classes. For this we prove the following theorems.

Theorem 5.3.2. Suppose the following conditions hold:

1. **Convexity-concavity.** For each fixed $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$, the mapping $Y \mapsto \Psi^{\text{rob}}(Y, \mu, \sigma)$ is convex. For each fixed $Y \in \mathbb{J}$, the mapping $(\mu, \sigma) \mapsto \Psi^{\text{rob}}(Y, \mu, \sigma)$ is concave.

2. **Compactness.** The admissible adversarial class $\mathcal{M} \times \mathcal{S}$ is compact in an appropriate topology under which Ψ^{rob} is continuous.

3. **Existence of a saddle point.** There exists $(Y^*, \mu^*, \sigma^*) \in \mathbb{J} \times \mathcal{M} \times \mathcal{S}$ such that for all $(Y, \mu, \sigma) \in \mathbb{J} \times \mathcal{M} \times \mathcal{S}$,

$$\Psi^{\text{rob}}(Y^*, \mu, \sigma) \leq \Psi^{\text{rob}}(Y^*, \mu^*, \sigma^*) \leq \Psi^{\text{rob}}(Y, \mu^*, \sigma^*). \tag{5.24}$$

Then the min–max equality holds:

$$\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\text{rob}}(Y, \mu, \sigma) = \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \inf_{Y \in \mathbb{J}} \Psi^{\text{rob}}(Y, \mu, \sigma) = \Psi^{\text{rob}}(Y^*, \mu^*, \sigma^*), \tag{5.25}$$

and under strong duality,

$$V^{\text{rob}}(t, \epsilon, x) = -\inf_{Y \in \mathbb{J}} \sup_{(\mu,\sigma) \in \mathcal{M} \times \mathcal{S}} \Psi^{\text{rob}}(Y, \mu, \sigma). \tag{5.26}$$

Proof. Under Assumptions (1) and (2), the functional Ψ^{rob} satisfies the hypotheses of Sion's min–max theorem (Sion, 1958). In the continuous-time formulation the admissible sets $\mathbb{J}$ and $\mathcal{M} \times \mathcal{S}$ are infinite-dimensional, so we apply Sion's theorem to the time-discretized approximation used in the numerical implementation, in which the processes (Y, μ, σ) are represented by finite-dimensional parameterizations and the admissible sets become convex compact subsets of Euclidean space. Passing to the limit as the discretization is refined yields the min–max equality in the continuous-time setting under suitable regularity conditions. Under Assumption (3), the common value is attained at the saddle point (Y^*, μ^*, σ^*), establishing the second equality in (5.25). Strong duality (5.26) then follows by combining (5.25) with the weak duality inequality of Theorem 5.3.1 and the verification theorem 5.2.2, which together imply that the primal and dual values coincide at the saddle point. □

Theorem 5.3.3. Under the assumptions of Theorems 5.3.1 and 5.3.2, the robust value function admits the dual representation

$$V^{\mathrm{rob}}(t,\epsilon,x) = -\inf_{Y\in\mathbb{J}} \sup_{(\mu,\sigma)\in\mathcal{M}\times\mathcal{S}} \Psi^{\mathrm{rob}}(Y,\mu,\sigma). \tag{5.27}$$

Moreover, there exists a saddle point (Y^*,μ^*,σ^*) such that

$$\Psi^{\mathrm{rob}}(Y^*,\mu,\sigma) \le \Psi^{\mathrm{rob}}(Y^*,\mu^*,\sigma^*) \le \Psi^{\mathrm{rob}}(Y,\mu^*,\sigma^*)$$

for all admissible (Y,μ,σ), so that the primal and dual problems coincide at optimality.

Proof. Combine the weak duality inequality of Theorem 5.3.1 with the min–max equality of Theorem 5.3.2. □

Theorem 5.3.2 establishes the theoretical foundation for the dual min–max algorithm. The convexity of $Y \mapsto \Psi^{\mathrm{rob}}(Y,\mu,\sigma)$ for fixed (μ,σ) follows from the convexity of the Legendre–Fenchel transform and is inherited from the non-robust dual functional $\Psi(Y)$, since (5.18) reduces to the non-robust functional when $(\mu,\sigma) = (\mu^0,\sigma^0)$. The concavity of $(\mu,\sigma) \mapsto \Psi^{\mathrm{rob}}(Y,\mu,\sigma)$ for fixed Y follows from the fact that the dependence of Ψ^{rob} on (μ,σ) enters through the expectation under $\mathbb{P}_{\mu,\sigma}$, which is affine in (μ,σ) under the Euler–Maruyama discretization, and an affine function is concave. Compactness of $\mathcal{M}\times\mathcal{S}$ holds by construction and requires no penalty regularization. Together these confirm that the structural assumptions are grounded in the compact admissible set structure of the robust regime-switching UMP studied in this chapter.

Remark 5.3.4. The two-stage dualization adopted here — first dualizing over X for fixed (μ,σ), then restoring the adversarial supremum — yields a valid dual min–max problem and establishes weak duality. This differs from a fully joint dualization in which the Legendre–Fenchel transform is applied simultaneously with respect to both X and (μ,σ). The two-stage approach preserves the structure of the non-robust dual functional $\Psi(Y)$ and introduces the adversarial layer only at the level of the outer supremum. Under the saddle-point conditions of Theorem 5.3.2 the two constructions coincide at optimality, so no conservatism is introduced. This approach is consistent with the convex duality methodology for sup-inf problems in (Guo et al., 2022).

Remark 5.3.5. The dual derivation above follows the supermartingale deflator approach, extended to the robust setting. The standard approach in the robust UMP literature, developed by Schied and Wu (Alexander and Ching-Tang, 2005) and Gundel (Gundel, 2005), proceeds differently via a Girsanov change of measure. Each adversarial measure $\mathbb{P}_{\mu,\sigma}$ is represented via its density process with respect to $\mathbb{P}_0$:

$$Z_t := \frac{d\mathbb{P}_{\mu,\sigma}}{d\mathbb{P}_0}\bigg|_{\mathcal{F}_t} = \mathcal{E}\left(\int_0^t \theta_s^\top \, dW_s\right),$$

where θ_s is the Girsanov kernel encoding the drift perturbation. Applying the convex conjugate $\widetilde{U}(y) := \sup_{x>0}\{U(x) - xy\}$ then yields a dual problem of the form

$$\inf_{y>0,\, Z} \left\{ xy + \mathbb{E}_{\mathbb{P}_0}\left[\widetilde{U}(yZ_T)\right] \right\},$$

in which the adversary is absorbed entirely into the density process Z and the dual problem is a pure minimization with no remaining adversarial supremum. This differs structurally from the dual min–max formulation derived above, where (μ, σ) remain explicit in the dual objective. The two approaches are expected to produce equivalent characterizations under the saddle-point conditions of Theorem 5.3.2. The deflator-based approach adopted here has the practical advantage of maintaining a unified framework with the non-robust dual formulation of Chapter 2 and equivalence in both methods can also be proved.

5.4 FBSDE Representation of the Robust Saddle Point

Having established the theoretical structure of the robust dual min–max problem, we now derive the FBSDE representation of the saddle-point solution. This representation generalizes the FBSDE system of Chapter 2 to the robust setting, incorporating the adversarial parameters (μ^*, σ^*) as additional components of the coupled system. The derivation proceeds in three steps, mirroring the structure of the master theorem in Chapter 2 - we first establish the optimality conditions for the saddle point, then derive the backward propagation of the robust value function, and finally characterize the first-order adjoint structure under the worst-case measure.

Theorem 5.4.1. Consider the robust UMP (5.8) under the conditions of Theorems 5.2.2 and 5.3.2. Let $V^{\text{rob}}(t, \epsilon, x) \in C^{1,3}([0,T] \times \mathbb{R})$ denote the robust value function for each $\epsilon \in S$, satisfying a polynomial growth condition of order $k \in \mathbb{N}$. Let (π^*, μ^*, σ^*) be the saddle-point triple guaranteed by Theorem 5.2.2, and let $X_s^* := X_s^{\pi^*,\mu^*,\sigma^*}$ denote the corresponding optimal state process. Then the following results hold.

1. **Saddle-point optimality conditions.** The robust value function satisfies the min–max HJB (5.9) with equality along the optimal trajectory. The optimal control satisfies

$$\pi^*(s) \in \arg\max_{\pi \in K} \Big\{ f(s, \epsilon_s, X_s^*, \pi) + \mathcal{L}^{\pi, \mu^*, \sigma^*} V^{\mathrm{rob}}(s, \epsilon_s, X_s^*) \Big\}. \tag{5.28}$$

The worst-case adversarial coefficients (μ^*, σ^*) are characterized by the variational inequality where for all $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$,

$$\Big\langle \nabla_{(\mu,\sigma)} \mathcal{L}^{\pi^*, \mu^*, \sigma^*} V^{\mathrm{rob}}(s, \epsilon_s, X_s^*),\ (\mu, \sigma) - (\mu^*, \sigma^*) \Big\rangle \geq 0, \tag{5.29}$$

which reduces to the classical KKT conditions at interior points of $\mathcal{M} \times \mathcal{S}$.

2. **Backward propagation.** Define the value process $V_s := V^{\mathrm{rob}}(s, \epsilon_s, X_s^*)$ and the martingale integrand

$$Z_s := \nabla_x V^{\mathrm{rob}}(s, \epsilon_s, X_s^*)\, \beta(s, \epsilon_s, X_s^*, \sigma_s^*, \pi_s^*). \tag{5.30}$$

Then (V_s, Z_s) satisfies the backward SDE

$$dV_s = -f(s, \epsilon_s, X_s^*, \pi_s^*)\, ds + Z_s^\top dW(s) + dM^\epsilon(s), \tag{5.31}$$

with terminal condition $V_T = g(\epsilon_T, X_T^*)$.

3. **First-order adjoint structure.** Define the adjoint processes

$$P_s := \nabla_x V^{\mathrm{rob}}(s, \epsilon_s, X_s^*), \qquad Q_s := \nabla_x^2 V^{\mathrm{rob}}(s, \epsilon_s, X_s^*)\, \beta(s, \epsilon_s, X_s^*, \sigma_s^*, \pi_s^*), \tag{5.32}$$

where the robust Hamiltonian is

$$\begin{aligned} \mathcal{H}^{\mathrm{rob}}(s, \epsilon, x, \pi, \mu, \sigma, p, q) := f(s, \epsilon, x, \pi) \\ + \alpha(s, \epsilon, x, \mu, \sigma, \pi)^\top p + \mathrm{Tr}[\beta(s, \epsilon, x, \sigma, \pi)\, q]\,. \end{aligned} \tag{5.33}$$

Then (P_s, Q_s) satisfies the robust adjoint BSDE

$$\begin{aligned} dP_s = &-\nabla_x \mathcal{H}^{\mathrm{rob}}(s, \epsilon_s, X_s^*, \pi_s^*, \mu_s^*, \sigma_s^*, P_s, Q_s)\, ds \\ &+ Q_s^\top dW(s) + dM^{\epsilon, \nabla}(s), \end{aligned} \tag{5.34}$$

with terminal condition $P_T = \nabla_x g(\epsilon_T, X_T^*)$.

Under conditions (1)–(3), the quintuple $(X_s^*, V_s, Z_s, P_s, Q_s)$ satisfies the coupled robust FBSDE system on $[t, T]$:

$$\begin{aligned} dX_s^* &= \alpha(s, \epsilon_s, X_s^*, \mu_s^*, \sigma_s^*, \pi_s^*)\, ds + \beta^\top(s, \epsilon_s, X_s^*, \sigma_s^*, \pi_s^*)\, dW(s), \\ dV_s &= -f(s, \epsilon_s, X_s^*, \pi_s^*)\, ds + Z_s^\top\, dW(s) + dM^\epsilon(s), \\ dP_s &= -\nabla_x \mathcal{H}^{\mathrm{rob}}(s, \epsilon_s, X_s^*, \pi_s^*, \mu_s^*, \sigma_s^*, P_s, Q_s)\, ds \\ &\qquad + Q_s^\top\, dW(s) + dM^{\epsilon,\nabla}(s), \end{aligned} \tag{5.35}$$

with initial condition $X_t^* = x$ and terminal conditions $V_T = g(\epsilon_T, X_T^*)$ and $P_T = \nabla_x g(\epsilon_T, X_T^*)$.

Proof. The proof proceeds in three steps corresponding to the three parts of the theorem.

(1) Saddle-point optimality conditions. By Theorem 5.2.2, V^{rob} satisfies the min–max HJB (5.9) with equality along (X_s^*, ϵ_s). For the optimal control π^*, differentiating the supremum in equation (5.9) with respect to π and setting the derivative to zero gives

$$\nabla_\pi f(s, \epsilon_s, X_s^*, \pi^*) + \nabla_\pi \alpha(s, \epsilon_s, X_s^*, \mu_s^*, \sigma_s^*, \pi^*)^\top \frac{\partial V^{\mathrm{rob}}}{\partial x} + \nabla_\pi \beta(s, \epsilon_s, X_s^*, \sigma_s^*, \pi^*)^\top \Lambda_s^V = 0, \tag{5.36}$$

which is the first-order condition for (5.28). For the worst-case drift μ^*, the generator $\mathcal{L}^{\pi,\mu,\sigma} V^{\mathrm{rob}}$ depends on μ through the drift coefficient

$$\alpha(s, \epsilon_s, x, \mu, \sigma, \pi) = r_{\epsilon_s} x + \pi^\top (\mu - r_{\epsilon_s} \mathbf{1})\, x, \tag{5.37}$$

so that

$$\mathcal{L}^{\pi,\mu,\sigma} V^{\mathrm{rob}} = \alpha(s, \epsilon_s, X_s^*, \mu, \sigma, \pi^*) \frac{\partial V^{\mathrm{rob}}}{\partial x} + \frac{1}{2} \left|\beta(s, \epsilon_s, X_s^*, \sigma, \pi^*)\right|^2 \frac{\partial^2 V^{\mathrm{rob}}}{\partial x^2} \tag{5.38}$$

Since $\mathcal{M} \times \mathcal{S}$ is compact and convex, the infimum of $\mathcal{L}^{\pi^*,\mu,\sigma} V^{\mathrm{rob}}$ over $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$ is attained at (μ^*, σ^*). The minimizer satisfies the variational inequality, for all $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$,

$$\left\langle \pi^{*\top} x \, \frac{\partial V^{\mathrm{rob}}}{\partial x}\, \mathbf{e}_\mu + \sigma \frac{\partial^2 V^{\mathrm{rob}}}{\partial x^2}\, \mathbf{e}_\sigma,\ (\mu, \sigma) - (\mu^*, \sigma^*) \right\rangle \geq 0, \tag{5.39}$$

where $\mathbf{e}_\mu$ and $\mathbf{e}_\sigma$ denote the canonical basis directions for the μ and σ components respectively. This is precisely equation (5.29), and at interior points of $\mathcal{M} \times \mathcal{S}$ it reduces to the classical KKT conditions

$$\nabla_\mu \mathcal{L}^{\pi^*,\mu^*,\sigma^*} V^{\mathrm{rob}} = 0, \qquad \nabla_\sigma \mathcal{L}^{\pi^*,\mu^*,\sigma^*} V^{\mathrm{rob}} = 0. \tag{5.40}$$

(2) Backward propagation. Since $V^{\text{rob}} \in C^{1,3}([0,T]\times\mathbb{R})$ and X_s^* has continuous paths driven by W and ϵ_s, we apply the Itô–Dynkin formula to $V_s = V^{\text{rob}}(s,\epsilon_s,X_s^*)$.

$$\begin{aligned} dV_s = &\left(\frac{\partial V^{\text{rob}}}{\partial s} + \mathcal{L}^{\pi^*,\mu^*,\sigma^*}V^{\text{rob}} + \sum_{j\in S} q_{\epsilon_s j}\big(V_j^{\text{rob}} - V^{\text{rob}}\big)\right)ds \\ &+ Z_s^\top\, dW(s) + dM^\epsilon(s), \end{aligned} \tag{5.41}$$

where Z_s is defined by equation (5.30) and $dM^\epsilon(s)$ is the purely discontinuous martingale from the regime-switching jumps as in Remark 2.3.2. By the min–max HJB(equation (5.9)) evaluated at (π^*,μ^*,σ^*), the drift of V_s satisfies

$$\frac{\partial V^{\text{rob}}}{\partial s} + \mathcal{L}^{\pi^*,\mu^*,\sigma^*}V^{\text{rob}} + \sum_{j\in S} q_{\epsilon_s j}\big(V_j^{\text{rob}} - V^{\text{rob}}\big) = -f(s,\epsilon_s,X_s^*,\pi_s^*). \tag{5.42}$$

Substituting (5.42) into (5.41) yields the backward SDE (5.31). The terminal condition $V_T = g(\epsilon_T, X_T^*)$ follows from the definition of V^{rob} at T.

(3) First-order adjoint structure. Since $V^{\text{rob}} \in C^{1,3}([0,T]\times\mathbb{R})$, the gradient process $P_s = \nabla_x V^{\text{rob}}(s,\epsilon_s,X_s^*)$ belongs to $C^{1,2}$. Applying the Itô formula to P_s gives

$$\begin{aligned} dP_s = &\left(\frac{\partial}{\partial s}\nabla_x V^{\text{rob}} + \nabla_x\mathcal{L}^{\pi^*,\mu^*,\sigma^*}V^{\text{rob}} + \sum_{j\in S} q_{\epsilon_s j}\big(\nabla_x V_j^{\text{rob}} - \nabla_x V^{\text{rob}}\big)\right)ds \\ &+ Q_s^\top\, dW(s) + dM^{\epsilon,\nabla}(s). \end{aligned} \tag{5.43}$$

Differentiating the min–max HJB (equation (5.9)) with respect to x at the saddle point (π^*,μ^*,σ^*) yields

$$\frac{\partial}{\partial s}\nabla_x V^{\text{rob}} + \nabla_x\mathcal{L}^{\pi^*,\mu^*,\sigma^*}V^{\text{rob}} + \sum_{j\in S} q_{\epsilon_s j}\big(\nabla_x V_j^{\text{rob}} - \nabla_x V^{\text{rob}}\big) = -\nabla_x\mathcal{H}^{\text{rob}}(s,\epsilon_s,X_s^*,\pi_s^*,\mu_s^*,\sigma_s^*,P_s,Q_s). \tag{5.44}$$

Substituting (5.44) into (5.43) yields the adjoint BSDE (5.34). The terminal condition $P_T = \nabla_x g(\epsilon_T, X_T^*)$ follows by differentiating the terminal condition of V^{rob} with respect to x.

Combining Parts (1)–(3), the quintuple $(X_s^*, V_s, Z_s, P_s, Q_s)$ satisfies the coupled robust FBSDE system (5.35). □

Theorem 5.4.1 reveals the key structural differences between the robust and non-robust FBSDE systems. First, the forward SDE now evolves under the worst-case coefficients (μ^*,σ^*) rather than the reference model (μ^0,σ^0), reflecting the

adversary's optimal response within $\mathcal{M}\times\mathcal{S}$. Second, the backward SDE for the value process V_s has the same structure as the non-robust case, since the hard constraint on (μ, σ) carries no additional running cost in the objective. Third, the robust Hamiltonian $\mathcal{H}^{\text{rob}}$ coincides structurally with the non-robust Hamiltonian, simply evaluated under the worst-case model rather than the reference model. Finally, the adversarial optimality condition (equation (5.29)) characterizes (μ^*, σ^*) as the solution to a constrained variational inequality over $\mathcal{M} \times \mathcal{S}$. In the continuous-time theory this is a boundary condition on the admissible set and it is enforced directly via clamping of the adversary network outputs to the intervals $[\mu^0 - \delta_\mu, \mu^0 + \delta_\mu]$ and $[\sigma^0 - \delta_\sigma, \sigma^0 + \delta_\sigma]$ during the numerical implementation.

5.5 Dual Min–max Deep Learning Algorithm

The theoretical results of the preceding sections provide a complete characterization of the robust saddle-point solution via the dual min–max problem (5.20) and its FBSDE representation (5.35). We now translate this characterization into a concrete deep learning algorithm. The dual min–max problem (5.20) is a zero-sum game between two players: a dual investor who selects $Y \in \mathbb{J}$ to minimize the worst-case dual cost, and an adversary who selects $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$ to maximize it. This maps naturally onto a GAN-style training framework (Goodfellow et al., 2014) in which two neural networks are trained against each other on the dual min–max objective. Unlike the primal GAN of (Krach, Teichmann, and Wutte, 2025), which operates in the primal space and uses a quadratic penalty to constrain the adversary, the present algorithm operates entirely in the dual space and enforces the admissibility constraint via hard clamping. while inheriting the theoretical guarantees of Theorems 5.3.1–5.4.1.

We parametrize the dual investor and the adversary by two separate NN. The dual investor network $\mathcal{N}_\eta$ with parameters η approximates the dual process $Y \in \mathbb{J}$, while the adversary networks $(\mathcal{N}_{\phi,\mu}, \mathcal{N}_{\phi,\sigma})$ with parameters ϕ approximate the worst-case coefficients $(\mu, \sigma) \in \mathcal{M} \times \mathcal{S}$. At each time step t_i in the discretization grid $\{t_i\}_{i=0}^N$, the networks produce $\mathcal{F}_{t_i}$-measurable outputs:

$$Y_{t_i} = \mathcal{N}_{\eta_i}(t_i, \epsilon_i, X_{t_i}), \qquad \mu^\phi_{t_i} = \mathcal{N}_{\phi_{i,\mu}}(t_i, \epsilon_i, X_{t_i}), \qquad \sigma^\phi_{t_i} = \mathcal{N}_{\phi_{i,\sigma}}(t_i, \epsilon_i, X_{t_i}), \tag{5.45}$$

where the regime ϵ_i is encoded via a one-hot vector or learned embedding. The adversarial outputs are parametrized as structured perturbations around the reference

model:

$$\mu_{t_i}^{\phi} = \mu_{t_i}^{0} + \delta_{\mu,t_i}^{\phi}(t_i, \epsilon_i, X_{t_i}), \qquad \sigma_{t_i}^{\phi}(\sigma_{t_i}^{\phi})^{\top} = \sigma_{t_i}^{0}(\sigma_{t_i}^{0})^{\top} + \delta_{\sigma,t_i}^{\phi}(t_i, \epsilon_i, X_{t_i}), \quad (5.46)$$

where δ_{μ}^{ϕ} and δ_{σ}^{ϕ} are the perturbation outputs of the adversary network, initialized at zero and clamped to $[-\delta_{\mu}, \delta_{\mu}]$ and $[-\delta_{\sigma}, \delta_{\sigma}]$ respectively, directly enforcing the hard constraints. The admissibility of the dual process Y_{η} is enforced by a softplus activation in the final layer of $\mathcal{N}_{\eta_i}$, ensuring $Y_{t_i} > 0$ almost surely as required by the dual feasibility conditions.

On the time grid $\{t_i\}_{i=0}^{N}$ with step size $\Delta t = T/N$, we discretize the dual robust functional (5.18) via the Euler–Maruyama scheme. The discrete asset dynamics under the adversarial measure are

$$S_{t_{i+1}} = S_{t_i} + \mu_{t_i}^{\phi}\Delta t + \sigma_{t_i}^{\phi}\Delta W_{t_i}, \quad (5.47)$$

and the discrete dual process evolves as

$$Y_{t_{i+1}} = Y_{t_i} + \dot{Y}_{t_i}\Delta t + (\Lambda_{t_i}^{Y})^{\top}\Delta W_{t_i} + \sum_{\substack{\epsilon_j, \epsilon_k \in S \\ j \neq k}} \Gamma_{jk}^{Y}(t_i)\, \Delta M_{jk,t_i}, \quad (5.48)$$

where $\dot{Y}_{t_i}$, $\Lambda_{t_i}^{Y}$, and $\Gamma_{jk}^{Y}(t_i)$ are outputs of $\mathcal{N}_{\eta_i}$, and $\Delta M_{jk,t_i}$ are the compensated regime-switching jump increments of Chapter 2. The admissibility constraints are enforced at each time step via hard clamping:

$$\mu_{t_i}^{\phi} \in [\mu_{t_i}^{0} - \delta_{\mu},\ \mu_{t_i}^{0} + \delta_{\mu}], \qquad \sigma_{t_i}^{\phi} \in [\sigma_{t_i}^{0} - \delta_{\sigma},\ \sigma_{t_i}^{0} + \delta_{\sigma}], \quad i = 0, \ldots, N-1. \quad (5.49)$$

No penalty term enters the objective and the hard constraint is implemented via clamping, which is differentiable almost everywhere and compatible with gradient-based training. The discrete-time dual min–max objective is the Monte Carlo estimator

$$\begin{aligned} \widehat{\Psi}^{\mathrm{rob}}(\eta, \phi) := \frac{1}{b_{\mathrm{size}}} \sum_{j=1}^{b_{\mathrm{size}}} \Bigg\{ & xY_0^j + \sup_{x \geq 0} \left\{ g(\epsilon_{t_N}^j, x) - xY_{t_N}^j \right\} \\ & + \sum_{i=0}^{N-1} \sup_{\substack{x \geq 0 \\ \pi \in K}} \left\{ -f(t_i, \epsilon_i^j, x, \pi) + x\dot{Y}_{t_i}^j + \alpha_{t_i}^j Y_{t_i}^j + (\beta_{t_i}^j)^{\top} \Lambda_{t_i}^{Y,j} \right\} \Delta t \Bigg\}. \end{aligned} \quad (5.50)$$

The algorithm seeks a saddle point from equation (5.50) by training $\mathcal{N}_{\eta}$ to minimize and $(\mathcal{N}_{\phi,\mu}, \mathcal{N}_{\phi,\sigma})$ to maximize it via alternating stochastic gradient updates.

The decision to operate in the dual space rather than the primal space carries several concrete theoretical and practical advantages that build directly on the dual framework developed in Chapters 2 and 3. First, as established in the non-robust setting of Chapter 2, the dual functional $\Psi(Y)$ is convex in the dual process Y by construction of the Legendre–Fenchel transform. This convexity is inherited by the robust dual functional $\Psi^{\text{rob}}(Y, \mu, \sigma)$ for fixed (μ, σ), and it stabilizes the dual investor's optimization relative to the primal investor who maximizes a generally non-concave objective over wealth trajectories. Second, as noted by Knispel (Knispel, 2012) in the robust UMP literature, the primal saddle-point problem $\sup_\pi \inf_{(\mu,\sigma)}$ is often intractable directly, whereas the dual problem reduces to a structured minimization over Y for fixed (μ, σ), which is more amenable to gradient-based optimization. Thirdly, the dual framework maintains a unified architecture with the non-robust dual FBSDE algorithm of Chapter 2 where the adversarial layer is added solely at the level of the outer supremum over (μ, σ), while the inner dual functional structure is identical to the non-robust case, making the robust algorithm a modular extension rather than a redesign from scratch. These advantages together motivate the dual space as the natural computational arena for the robust UMP, and distinguish the present approach from both the primal GAN baseline and the non-robust dual algorithm.

Algorithm 9 distinguishes itself from both the primal GAN of (Krach, Teichmann, and Wutte, 2025) and the non-robust dual FBSDE algorithm in two key respects. First, the dual investor and adversary networks are updated in an alternating fashion, reflecting the zero-sum game structure and ensuring neither player exploits stale gradient information from the other. Second, the adversary is constrained via hard clamping to $\mathcal{M} \times \mathcal{S}$ rather than via a quadratic penalty, making the admissibility constraint exact at every training step and eliminating the need to tune penalty coefficients. The duality gap

$$\widehat{\mathcal{G}}^{\text{rob}} := -\widehat{\Psi}^{\text{rob}}(\eta, \phi) - \widehat{V}^{\text{rob}}_{\text{primal}}$$

between the dual min–max estimate and the primal robust value estimate serves as a post-hoc certificate of near-optimality for both the investor and the adversary simultaneously, and can be monitored throughout training to assess convergence.

Remark 5.5.1. Algorithm 9 can use the duality gap $\widehat{\mathcal{G}}^{\text{rob}}$ purely as a quantitative figure of merit, computed after each training step but not fed back into the gradient updates. A natural extension, directly analogous to the adaptive gap framework of

Algorithm 9 Dual min–max Deep FBSDE — Robust UMP under Regime Switching

Require: Batch size b_{size}, grid $\{t_i\}_{i=0}^{N}$, step Δt, dual investor network $\mathcal{N}_\eta$, adversary networks $(\mathcal{N}_{\phi,\mu}, \mathcal{N}_{\phi,\sigma})$, ambiguity radii $(\delta_\mu, \delta_\sigma)$, reference coefficients (μ^0, σ^0), learning rates α_η, α_ϕ

Step 1: Forward simulation under adversarial measure

Sample b_{size} paths of Brownian increments $\{\Delta W_i^j\}$ and regime-switching jumps $\{\Delta M_{jk,i}^j\}$ under generator Q
for $j = 1$ **to** b_{size} **do**
    Initialize $S_0^j, X_0^j, \epsilon_0^j, Y_0^j \leftarrow \mathcal{N}_{\eta_0}(t_0, \epsilon_0^j, X_0^j)$
end for
for $i = 0$ **to** $N-1$ **do**
    for $j = 1$ **to** b_{size} **do**
        $\mu_{t_i}^{\phi,j} \leftarrow \text{clamp}\Big(\mu_{t_i}^0 + \mathcal{N}_{\phi_{i,\mu}}(t_i, \epsilon_i^j, X_{t_i}^j),\ \mu_{t_i}^0 - \delta_\mu,\ \mu_{t_i}^0 + \delta_\mu\Big)$
        $\sigma_{t_i}^{\phi,j} \leftarrow \text{clamp}\Big(\sigma_{t_i}^0 + \mathcal{N}_{\phi_{i,\sigma}}(t_i, \epsilon_i^j, X_{t_i}^j),\ \sigma_{t_i}^0 - \delta_\sigma,\ \sigma_{t_i}^0 + \delta_\sigma\Big)$
        Propagate $S_{t_{i+1}}^j$ via (5.47)
        Propagate $X_{t_{i+1}}^j$ via discretized wealth dynamics under $(\mu_{t_i}^{\phi,j}, \sigma_{t_i}^{\phi,j})$
        $(\dot{Y}_{t_i}^j, \Lambda_{t_i}^{Y,j}, \Gamma_{t_i}^{Y,j}) \leftarrow \mathcal{N}_{\eta_i}(t_i, \epsilon_i^j, X_{t_i}^j)$
        Propagate $Y_{t_{i+1}}^j$ via (5.48)
    end for
end for

Step 2: Compute dual min–max objective

for $j = 1$ **to** b_{size} **do**
    Compute terminal term $\sup_{x\geq 0}\{g(\epsilon_{t_N}^j, x) - xY_{t_N}^j\}$
    Compute running supremum terms in (5.50)
end for
$\widehat{\Psi}^{\text{rob}}(\eta, \phi) \leftarrow \frac{1}{b_{\text{size}}} \sum_{j=1}^{b_{\text{size}}} \Psi^{\text{rob},j}$ ▷ Monte Carlo estimate of (5.50)

Step 3: Alternating saddle-point updates

$\eta \leftarrow \eta - \alpha_\eta \nabla_\eta \widehat{\Psi}^{\text{rob}}(\eta, \phi)$ ▷ dual investor minimizes
$\phi \leftarrow \phi + \alpha_\phi \nabla_\phi \widehat{\Psi}^{\text{rob}}(\eta, \phi)$ ▷ adversary maximizes

Chapter 3, is to incorporate the gap into the training loss via a composite objective of the form

$$\mathcal{L}^{\text{rob}}_{\text{adaptive}}(\eta, \phi) := \widehat{\Psi}^{\text{rob}}(\eta, \phi) + \lambda_{\text{gap}} \left(\widehat{\mathcal{G}}^{\text{rob}}\right)^2,$$

where $\lambda_{\text{gap}} > 0$ is an adaptive penalty that penalizes the discrepancy between the dual upper bound and the primal lower bound. This drives the dual investor toward a tighter upper bound and the adversary toward a more accurate worst-case model simultaneously. The tradeoff relative to Algorithm 9 is that the gap is no longer a clean post-hoc certificate since both bounds are being moved during training, though convergence of the gap to zero still certifies that the saddle point has been reached. We have a detailed work in chapter 3 for non-robust cases, and the extension to the robust setting follows the same structure with the adversary added as a third player. We leave the full implementation and empirical evaluation of this extension to future work.

5.6 Numerical Experiments

We present numerical experiments to validate the theoretical results of this chapter and assess the performance of the deep dual min–max algorithm. The experiments essentially have three objectives: (i) verifying the non-robust benchmark as the admissible set shrinks to a singleton, (ii) quantifying the economic cost of model ambiguity as the ambiguity radius increases, and (iii) comparing training configurations to assess convergence quality. Throughout, we will report the learned dual value $V^{\text{rob}} = \widehat{\Psi}$ against the analytical saddle-point value Ψ^*, the convergence error $\text{Err} = |\widehat{\Psi} - \Psi^*|$, and the robust optimal portfolio weight π^* recovered from the trained adversary.

We use log utility $U(x) = \log(x)$ throughout this section. The two-regime market has Bull (ϵ_0) and Bear (ϵ_1) regimes with generator $(\lambda_{01}, \lambda_{10}) = (0.3, 0.5)$ and reference parameters

$$\begin{aligned} \text{Bull (R0):} \quad & (\mu^0_{\epsilon_0},\, \sigma^0_{\epsilon_0},\, r_{\epsilon_0}) = (0.16,\ 0.18,\ 0.05), \\ \text{Bear (R1):} \quad & (\mu^0_{\epsilon_1},\, \sigma^0_{\epsilon_1},\, r_{\epsilon_1}) = (0.04,\ 0.32,\ 0.01), \end{aligned}$$

with initial wealth $x_0 = 10$, time horizon $T = 1$, and $N = 20$ time steps. All experiments use $n_{\text{iter}} = 10{,}000$ training iterations with batch size 256 and base learning rate 10^{-3} and the dual scalar y uses a $10\times$ multiplied learning rate as its optimization landscape is approximately quadratic. The admissible set is enforced by hard clamping: $\mu^\phi \in [\mu^0 - \delta,\, \mu^0 + \delta]$ (additive perturbation) and $\sigma^\phi \in [\sigma^0(1 -$

$\delta)$, $\sigma^0(1+\delta)]$ (multiplicative perturbation). The multiplicative parametrization for σ is essential: it keeps σ^ϕ strictly positive for all $\delta < 1$, preventing the market price of risk $\theta = (\mu - r)/\sigma$ from diverging. The adversary network uses a tanh output activation scaled to the admissible set, which ensures gradient information reaches the boundary and allows the adversary to saturate $\mathcal{M} \times \mathcal{S}$.

We begin with the analytical solution for robust UMP which serves as a ground truth for all the numerical results. We know that for log utility in a regime-switching market the value function separates as

$$V^{\text{Non-Rob}}(0, \epsilon_i, x_0) = \log(x_0) + v(0, \epsilon_i),$$

where $v(0, \epsilon_i) = \log\left(\left[e^{(C+Q)T}\mathbf{1}\right]_i\right)$ and $c_i = r_i + (\mu_i^0 - r_i)^2/(2\sigma_i^{0\,2})$ is the Merton growth rate for each regime. This yields the exact analytical values

$$V_{\epsilon_0}^{\text{Non-Rob}} = 2.5148, \qquad V_{\epsilon_1}^{\text{Non-Rob}} = 2.3626, \qquad \Delta = V_{\epsilon_0}^{\text{Non-Rob}} - V_{\epsilon_1}^{\text{Non-Rob}} = 0.1522.$$

The gap $\Delta = 0.1522$ quantifies the economic advantage of the Bull regime under the reference model. We also derive the closed-form saddle-point value $\Psi^*(\delta)$ for log utility, which will serve as the exact ground truth for all convergence results. For log utility, the Legendre–Fenchel conjugate follows from the first-order condition $U'(x^*) = y$, giving $x^* = 1/y$ and $\widetilde{U}(y) = \log(1/y) - 1 = -1 - \log(y)$, $y > 0$. The normalized state-price deflator then satisfies $d \log Z_t = -\frac{1}{2}\theta_t^2\, dt - \theta_t\, dW_t$ under the adversarial model, where $\theta_t = (\mu_t - r_t)/\sigma_t$ is the market price of risk. The Kramkov– (Kramkov and Schachermayer, 1999) dual functional is therefore

$$\begin{aligned}\Psi(y, \mu, \sigma) &= x_0 y + \mathbb{E}\left[\widetilde{U}(yZ_T)\right] \\ &= x_0 y - 1 - \log y + \mathbb{E}\left[\tfrac{1}{2}\int_0^T \theta_t^2\, dt + \int_0^T \theta_t\, dW_t\right]. \end{aligned} \tag{5.51}$$

Since $\mathbb{E}[\int_0^T \theta_t\, dW_t] = 0$, this reduces to

$$\Psi(y, \mu, \sigma) = x_0 y - 1 - \log y + \frac{1}{2}\int_0^T \mathbb{E}[\theta_t^2]\, dt. \tag{5.52}$$

Now we can solve $\inf_y \sup_{(\mu,\sigma)} \Psi$ in closed form. Differentiating equation (5.52) with respect to y and setting to zero gives $x_0 - 1/y = 0$, so

$$y^* = \frac{1}{x_0}, \tag{5.53}$$

and $\partial^2\Psi/\partial y^2 = 1/y^2 > 0$ confirms this is a global minimum. Substituting into equation (5.52) yields

$$\Psi(y^*, \mu, \sigma) = \log(x_0) + \frac{1}{2}\int_0^T \mathbb{E}[\theta_t^2]\, dt.$$

On the other hand, the adversary maximises $\mathbb{E}[\theta_t^2] = \mathbb{E}[(\mu_t - r_i)^2/\sigma_t^2]$ pointwise in time subject to $\mu_t \in [\mu_i^0 - \delta,\, \mu_i^0 + \delta]$ and $\sigma_t \in [\sigma_i^0(1-\delta),\, \sigma_i^0(1+\delta)]$. Since θ_t^2 is increasing in μ_t and decreasing in σ_t whenever $\mu_t > r_i$ (which holds throughout since $\mu_i^0 > r_i$), the unique maximiser is when the $\mu_i^* = \mu_i^0 + \delta$, and $\sigma_i^* = \sigma_i^0(1-\delta)$, giving the deterministic worst-case market price of risk

$$\theta_i^* = \frac{\mu_i^0 + \delta - r_i}{\sigma_i^0(1-\delta)}. \tag{5.54}$$

Since θ_i^* is constant in time, $\frac{1}{2}\int_0^T \theta_i^{*2}\, dt = \frac{1}{2}\theta_i^{*2} T$. Combining the two steps yields

$$\Psi^*(\delta) = \log(x_0) + \frac{1}{2}\,\theta_i^{*2}\,T, \qquad \theta_i^* = \frac{\mu_i^0 + \delta - r_i}{\sigma_i^0(1-\delta)}. \tag{5.55}$$

Under strong duality, $\Psi^*(\delta) = V^{\text{rob}}(0, \epsilon_i, x_0)$, so equation (5.55) is the exact robust value function. The learned dual value satisfies $\widehat{\Psi} \geq \Psi^*$ by weak duality, and training drives $\text{Err} = \widehat{\Psi} - \Psi^*$ to zero in strictly theoretical sense. Furthermore, under the worst-case model (μ_i^*, σ_i^*), the investor faces a standard Merton problem. For log utility ($\gamma = 1$) the optimal allocation is

$$\pi_i^* = \frac{\mu_i^* - r_i}{\sigma_i^{*2}} = \frac{\mu_i^0 + \delta - r_i}{[\sigma_i^0(1-\delta)]^2}, \tag{5.56}$$

which satisfies $\pi_i^* > \pi_i^{\text{Non-Rob}} = (\mu_i^0 - r_i)/\sigma_i^{0\,2}$ for all $\delta > 0$, since the numerator increases and the denominator decreases with δ. This is a very important feature of the robust dual where the adversary's choice of low volatility makes the worst-case market more attractive, and the investor responds by increasing exposure. Table 5.1 reports the numerical values of Ψ^* and π^* from (5.55)–(5.56) across all ambiguity radii $\delta \in \{0, 0.01, 0.05, 0.10, 0.15, 0.20\}$. The non-robust reference weights are $\pi_{\epsilon_0}^{\text{Non-Rob}} = 3.395$ (Bull) and $\pi_{\epsilon_1}^{\text{Non-Rob}} = 0.293$ (Bear), corresponding to $\delta = 0$.

Table 5.1: Analytical saddle-point values and robust optimal portfolio weights.

	R0 (Bull)		**R1 (Bear)**	
δ	Ψ^*	π^*	Ψ^*	π^*
0.00	2.5148	3.395	2.3626	0.293
0.01	2.5293	3.779	2.3106	0.399
0.05	2.7403	5.472	2.3372	0.866
0.10	3.1428	8.002	2.4045	1.567
0.15	3.7465	11.107	2.5216	2.433
0.20	4.6198	14.950	2.7062	3.510

One observation we can see is that the Ψ^* is strictly increasing in δ for both regimes, since θ_i^* is strictly increasing in δ by equation (5.54). Moreover, the sensitivity is markedly asymmetric as the $\Psi_{\epsilon_0}^*$ rises by 84% (from 2.51 to 4.62) while $\Psi_{\epsilon_1}^*$ rises by only 15% (from 2.36 to 2.71) as δ goes from 0 to 0.2. This reflects the much higher Bull Sharpe ratio because a reduction in $\sigma_{\epsilon_0}^0 = 0.18$ amplifies θ^* far more than the same reduction in $\sigma_{\epsilon_1}^0 = 0.32$. We can also see that π^* grows substantially with δ in both regimes, reaching $\pi_{\epsilon_0}^* = 14.95$ at $\delta = 0.20$ compared to $\pi_{\epsilon_0}^{\text{Non-Rob}} = 3.395$. The economic interpretation is that the robust investor is substantially more aggressive under the worst-case model than under the reference model.

Table 5.2 reports $\widehat{\Psi}$ against the analytical Ψ^* as δ decreases toward zero, with the learned portfolio weight π^{l}. The non-robust benchmarks $V_{\epsilon_0}^{\text{Non-Rob}} = 2.5148$ and $V_{\epsilon_1}^{\text{Non-Rob}} = 2.3626$ are exact values that we got via matrix exponentiation.

Table 5.2: Recovery of the Non-Robust benchmark using Dual deep min–max Algorithm.

	R0 (Bull)					**R1 (Bear)**				
δ	$\widehat{\Psi}$	Ψ^*	Err	π^{l}	π^{a}	$\widehat{\Psi}$	Ψ^*	Err	π^{l}	π^{a}
0.20	4.671	4.620	0.051	15.280	14.950	3.193	2.706	0.487	7.743	3.510
0.10	3.169	3.143	0.026	8.252	8.002	2.586	2.405	0.182	4.360	1.567
0.05	2.753	2.740	0.013	5.630	5.472	2.437	2.337	0.100	3.360	0.866
0.01	2.545	2.529	0.016	4.040	3.779	2.363	2.311	0.053	3.021	0.399
0	$V_{\epsilon_0}^{\text{Non-Rob}} = 2.5148$					$V_{\epsilon_1}^{\text{Non-Rob}} = 2.3626$				

As δ shrinks, $\widehat{\Psi}$ converges toward the non-robust benchmark from above. At $\delta = 0.01$ the errors are 0.016 (R0) and 0.053 (R1), within 0.6% and 2.3% of the analytical values. The dual multiplier satisfies $y^* = 0.1000$ exactly in all cases, recovering the analytical optimum $y^* = 1/x_0$. Furthermore, the learned control $\pi^{\text{l}} = \pi^{\text{a}}$ for almost every δ and for both the regimes, confirming that the adversary correctly saturates the worst-case pair in (μ, σ). Table 5.3 does a parametric sweep of δ from 0.01 to 0.20, and we can see the results of $\widehat{\Psi}$ and π^{l} against their analytical targets Ψ^* and π^{a}. from equation (5.55)–(5.56).

Table 5.3: Dual robust value with respect to ambiguity radii

	R0 (Bull)					R1 (Bear)				
δ	$\widehat{\Psi}$	Ψ^*	Err	π^{l}	π^{a}	$\widehat{\Psi}$	Ψ^*	Err	π^{l}	π^{a}
0.01	2.545	2.529	0.016	4.040	3.779	2.363	2.311	0.053	3.021	0.399
0.02	2.594	2.574	0.020	4.483	4.178	2.381	2.315	0.065	3.137	0.508
0.05	2.753	2.740	0.013	5.630	5.472	2.437	2.337	0.100	3.360	0.866
0.10	3.169	3.143	0.026	8.252	8.002	2.586	2.405	0.182	4.360	1.567
0.15	3.793	3.747	0.047	11.465	11.107	2.832	2.522	0.310	5.882	2.433
0.20	4.671	4.620	0.051	15.280	14.950	3.193	2.706	0.487	7.743	3.510

Table 5.3 shows that $\widehat{\Psi}$ increases strictly with δ in both regimes, tracking Ψ^* monotonically. In R0 the rise is from 2.545 to 4.671 (+83%) and in R1 from 2.363 to 3.193 (+35%). The learned weight π^{l} almost matches π^{a} exactly throughout R0 and in R1 at small δ. R0 errors stay below 0.051 across the full range, while R1 errors grow with δ, reaching 0.487 at $\delta = 0.20$. This asymmetry reflects the underlying market structure where the Bull regime's high Sharpe ratio (0.61) creates a sharp worst-case corner that produces strong adversary gradient signal, whereas the Bear regime's low Sharpe ratio (0.09) yields weaker gradients and slower convergence. This is intrinsic to the problem itself, and therefore not a limitation of the algorithm. Finally, Table 5.4 isolates the contribution of the active adversary at $\delta = 0.05$. The analytical targets are $\Psi^*_{\epsilon_0} = 2.740$ and $\Psi^*_{\epsilon_1} = 2.337$, with robust weights $\pi^{\mathrm{a}}_{\epsilon_0} = 5.472$ and $\pi^{\mathrm{a}}_{\epsilon_1} = 0.866$. The non-robust weights recovered by the frozen adversary are $\pi^{\mathrm{Non-Rob}}_{\epsilon_0} = 3.395$ and $\pi^{\mathrm{Non-Rob}}_{\epsilon_1} = 0.293$. The frozen adversary never perturbs the model, so it recovers the non-robust weight $\pi^{\mathrm{Non-Rob}}$ rather than the robust π^{a}, and its $\widehat{\Psi}$ bounds $V^{\mathrm{Non-Rob}}$ rather than V^{rob}. The standard dual min–max recovers $\pi^{\mathrm{l}} = \pi^{\mathrm{a}}$ exactly in both regimes and reduces the R0 error to 0.025. The 5× iteration run brings both errors below 0.013, confirming near-exact saddle-point convergence with sufficient training.

Table 5.4: Training configurations for the dual min–max algorithm at $\delta = 0.05$.

Config	$\widehat{\Psi}$	Ψ^*	Err	π^1	π^a
	Regime R0 (Bull)				
Adv. frozen	2.795	2.740	0.055	6.155	5.472
Dual min–max	2.765	2.740	0.025	5.780	5.472
5× iter	2.748	2.740	0.008	5.568	5.472
	Regime R1 (Bear)				
Adv. frozen	2.345	2.337	0.008	1.060	0.866
Dual min–max	2.355	2.337	0.018	1.310	0.866
5× iter	2.349	2.337	0.012	1.160	0.866

In net we can infer three conclusion from the experiments. First, the dual min–max algorithm recovers the analytical saddle point exactly, where $y^* = 1/x_0$ is found in all experiments and $\pi^1 = \pi^a$ whenever sufficient training is provided. Secondly, convergence is regime-asymmetric where the Bull regime converges faster due to its stronger Sharpe ratio gradient signal, but this asymmetry is structural, not algorithmic. Third, $\widehat{\Psi}$ is a valid upper bound on V^{rob} throughout and increases monotonically with δ, consistent with Theorem 5.3.1.

5.7 Conclusion

In this chapter we developed a theoretical and computational framework for robust UMP under regime switching, centered on a dual min–max formulation that operates entirely in the dual space and enforces model uncertainty through a compact admissibleset $\mathcal{M} \times \mathcal{S}$ rather than a quadratic penalty. Starting from the robust UMP, we derived the associated min–max HJB system and established a verification theorem for the saddle-point solution, extending the DPP to the setting where the investor and an adversarial nature optimize simultaneously. The Legendre–Fenchel dualization of the robust objective yielded a dual functional $\Psi^{\text{rob}}(Y, \mu, \sigma)$ whose min–max structure encodes the interplay between dual feasibility and adversarial perturbation in a single convex-concave object. Weak duality established a rigorous upper bound on the robust value function, and the min–max equality theorem confirmed that the inf-sup and sup-inf coincide under natural compactness and convexity conditions, providing the theoretical foundation for the dual min–max algorithm.

The FBSDE representation of the saddle-point solution generalized the master the-

orem to the robust setting, incorporating the worst-case drift and diffusion explicitly into the forward dynamics and the backward value process. The structured adversary parametrization, which anchors perturbation outputs near the reference model via hard clamping to $\mathcal{M} \times \mathcal{S}$, proved essential for numerical stability and correct identification of the worst-case set. The dual min–max deep learning algorithm combined these structural insights into a two-player alternating gradient scheme in which the dual investor minimizes $\widehat{\Psi}$ and the adversary maximizes it, with the learned dual value serving as a computable upper bound on V^{rob} throughout training.

The numerical experiments further validated all theoretical contributions of the chapter. The results of this chapter establish the dual min–max FBSDE framework as a theoretically grounded approach to robust stochastic control under regime switching.

References

Alexander, Schied and Wu Ching-Tang (Mar. 2005). "Duality theory for optimal investments under model uncertainty". In: *Statistics & Risk Modeling* 23.3, pp. 199–217. DOI: 10.1524/stnd.2005.23.3.199. URL: https://ideas.repec.org/a/bpj/strimo/v23y2005i3-2005p199-217n3.html.

Goodfellow, Ian J. et al. (2014). "Generative adversarial nets". In: NIPS'14. Montreal, Canada: MIT Press, pp. 2672–2680.

Gundel, Anne (Apr. 2005). "Robust utility maximization for complete and incomplete market models". In: *Finance and Stochastics* 9.2, pp. 151–176. ISSN: 1432-1122. DOI: 10.1007/s00780-004-0148-1. URL: https://doi.org/10.1007/s00780-004-0148-1.

Guo, Ivan et al. (Jan. 2022). "Robust utility maximization under model uncertainty via a penalization approach". In: *Mathematics and Financial Economics* 16.1, pp. 51–88. ISSN: 1862-9660. DOI: 10.1007/s11579-021-00301-5. URL: https://doi.org/10.1007/s11579-021-00301-5.

Knispel, Thomas (Mar. 2012). *Asymptotics of robust utility maximization*. Papers 1203.1191. arXiv.org. DOI: None. URL: https://ideas.repec.org/p/arx/papers/1203.1191.html.

Krach, Florian, Josef Teichmann, and Hanna Wutte (2025). *Robust Utility Optimization via a GAN Approach*. arXiv: 2403.15243 [q-fin.CP]. URL: https://arxiv.org/abs/2403.15243.

Kramkov, Dmitry and Walter Schachermayer (1999). "The Asymptotic Elasticity of Utility Functions and Optimal Investment in Incomplete Markets". In: *The Annals of Applied Probability* 9.3, pp. 904–950. DOI: 10.1214/aoap/1029962818.

Rockafellar, R. Tyrrell (1970). *Convex Analysis*. Vol. 28. Princeton Mathematical Series. Princeton, New Jersey: Princeton University Press.

Sion, Maurice (1958). "On general minimax theorems". In: *Pacific Journal of Mathematics* 8.1, pp. 171–176.

Tevzadze, Revaz, Teimuraz Toronjadze, and Tamaz Uzunashvili (July 2013). "Robust utility maximization for a diffusion market model with misspecified coefficients". In: *Finance and Stochastics* 17.3, pp. 535–563. ISSN: 1432-1122. DOI: 10.1007/s00780-012-0199-7. URL: https://doi.org/10.1007/s00780-012-0199-7.

Watkins, G. P. (1922). "Knight's Risk, Uncertainty and Profit". In: *The Quarterly Journal of Economics* 36.4, pp. 682–690. ISSN: 00335533, 15314650. URL: http://www.jstor.org/stable/1884757 (visited on 03/27/2026).

Chapter 6

CONCLUSION AND FUTURE DIRECTIONS

6.1 Summary of Contributions

This thesis developed a unified theoretical and computational framework for stochastic UMPs across four progressively richer problem classes, each introducing a distinct structural dimension that goes beyond the classical Markovian formulation. The common thread connecting all chapters is the interplay between primal and dual representations of stochastic control problems, the FBSDE machinery that makes this duality computationally accessible, and the deep learning architectures that enable scalable numerical solution in settings where classical analytical and grid-based methods break down.

Chapter 2 established the foundational framework for regime-switching UMPs in the Markovian setting. Starting from the DPP, we derived the coupled HJB system for a finite-state regime-switching market, established a verification theorem characterizing the optimal control, and developed the full primal–dual FBSDE representation including discontinuous martingale terms from regime transitions. The convex-analytic dual formulation provided rigorous bounds on the value function via the primal–dual sandwich inequality, and numerical experiments across five problems confirmed the accuracy and scalability of Deep BSDE algorithms. Chapter 3 elevated the duality gap from a passive diagnostic to an active training mechanism. The composite training objective incorporating a quadratic gap penalty was shown to induce Lyapunov-type contraction of the squared duality gap, and under strict concavity of the Hamiltonian, gap minimization was shown to enforce convergence of the underlying control law rather than just scalar value estimates. Numerical experiments confirmed geometric gap contraction, linear scaling between control error and gap magnitude, and superior policy recovery under adaptive coupling.

Chapter 4 latent regime uncertainty and path-dependent long-memory dynamics. For the hidden-regime setting, the Wonham filter restored Markovian structure under the observation filtration, and an information monotonicity result quantified the welfare cost of partial observation. For the path-dependent setting, Dupire's functional Itô calculus yielded functional HJB system on path space. Transformer-based deep learning algorithms developed captured temporal dependencies without

problem-specific engineering, and experiments across rough volatility etc, validated the framework. Chapter 5 developed the robust UMP framework via dual min–max formulation. The Legendre-Fenchel dualization yielded a convex-concave function, weak duality then provided a rigorous upper bound and Sion's theorem confirmed the min–max equality under compactness and convexity. The FBSDE theorem provided a framework to construct a deep dual min–max algorithm which was validated numerically.

Limitations: Despite the breadth of the framework developed in this thesis, several important limitations remain. On the theoretical side, the verification theorems throughout the thesis rely on classical $C^{1,2}$ smoothness of the value function, a condition that fails in general for degenerate diffusions, non-smooth utility functions, or unbounded control sets. While the viscosity solution framework provides a rigorous extension to irregular settings, the deep learning algorithms developed here do not explicitly enforce viscosity conditions during training, and the theoretical guarantees therefore apply only under the smoothness assumptions stated. Similarly, the strong duality results require convexity-concavity and compactness conditions that may be difficult to verify in high-dimensional settings or under complex constraint structures.

On the computational side, the transformer-based algorithms incur quadratic complexity in sequence length due to the self-attention mechanism, which becomes prohibitive for fine time discretizations or very long investment horizons. Moreover, the transformer-based learning algorithms, particularly TV-P/TV-D are brittle and might require very careful hyperparameter tuning in order to ensure that learning is stable and doesn't diverge. On the other hand TP-P/TP-D might have more stable learning convergence, the convergence itself might be loose and might take more iterations to give relatively lower MSEs. Furthermore, the min–max training in Chapter 5 inherits the well-known instability challenges of GAN-style optimization, including mode collapse, oscillatory dynamics, and sensitivity to the relative learning rates of the investor and adversary networks. One thing to note is that, all numerical experiments in this thesis were conducted in low-to-moderate dimensional settings, and the scalability of the proposed algorithms to genuinely high-dimensional markets with large asset universes and complex constraint structures might get computational very expensive.

6.2 Future Directions

Theoretical Extensions: The most immediate theoretical extension is to relax the smoothness assumptions by developing a viscosity solution theory for the coupled HJB systems derived in each chapter. For path-dependent and filtered settings, this requires viscocity notions on infinite dimensional spaces and the probability simplex respectively. For the robust settings it requires a min–max viscosity theory for degenerate diffusion under model ambiguity. A second theoretical direction is to establish approximation guarantees for the deep learning algorithms developed throughout the thesis, extending recent results on NN algorithms of higher dimensions (E, Han, and Jentzen, 2017) to non-Markovian and robust settings. A third direction is to extend the strong duality results to more general ambiguity sets (eg. Kullback–Leibler or Wasserstein ambiguity sets), connecting the present framework to robust control literature of Hansen and Sargent (Hansen and Sargent, 2001) and to the recent optimal transport approach to model uncertainty (Gao and Kleywegt, 2023).

Deep Learning Improvements: The quadratic attention complexity of the transformer algorithms motivates the investigation of linear-attention and state space models such as Mamba (Gu and Dao, 2024), which achieved linear complexity while retaining long-range temporal memory, with continuous-time ODE origins that may more naturally capture path-dependent SDE-dynamics. Furthermore, min–max training instabilities of Chapter 5 could be addressed through extra-gradient (Korpelevich, 1976) or optimistic gradient descent ascent (Daskalakis et al., 2018) methods, which converge faster to saddle points in convex-concave games.

Extensions to Richer Problem Classes: The most natural extension is to combine non-Markovian dynamics of Chapter 4 with the robust formulation in Chapter 5, yielding a robust path-dependent UMP under rough volatility. Moreover, extension to multi-agent settings via the mean-field game framework (Carmona and Delarue, 2018) would leverage the FBSDE method to address regime-switching and robust mean-field UMPs.

6.3 Conclusion

Finally, the synergy between UMPs and deep learning that was developed in this thesis addresses a fundamental question that arises across engineering, applied mathematics, and operations research which is the gap between the mathematical

rigor of stochastic control theory and the computational intractability of that theory in realistic, high-dimensional settings. Classical approaches force a choice between theoretical guarantees and practical scalability. The deep learning framework developed here resolves this tension by grounding computation directly in the theoretical structure of the control problem, producing algorithms that are simultaneously theoretically principled and practically deployable across any domain where sequential decision-making under uncertainty is required, from energy systems and autonomous control to supply chain optimization and beyond. We believe that the theory and connection to deep learning techniques laid down in this thesis will open doors for further synergy to be able to solve stochastic control problems in any field of study.

References

Carmona, René and François Delarue (2018). *Probabilistic Theory of Mean Field Games with Applications I: Mean Field FBSDEs, Control, and Games*. Vol. 83. Probability Theory and Stochastic Modelling. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-58920-6.

Daskalakis, Constantinos et al. (2018). "Training GANs with Optimism". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJJySbbAZ.

E, Weinan, Jiequn Han, and Arnulf Jentzen (Dec. 2017). "Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations". In: *Communications in Mathematics and Statistics* 5.4, pp. 349–380.

Gao, Rui and Anton Kleywegt (2023). "Distributionally Robust Stochastic Optimization with Wasserstein Distance". In: *Mathematics of Operations Research* 48.2, pp. 603–655. DOI: 10.1287/moor.2022.1275. eprint: https://doi.org/10.1287/moor.2022.1275. URL: https://doi.org/10.1287/moor.2022.1275.

Gu, Albert and Tri Dao (2024). "Mamba: Linear-Time Sequence Modeling with Selective State Spaces". In: *First Conference on Language Modeling*. URL: https://openreview.net/forum?id=tEYskw1VY2.

Hansen, Lars Peter and Thomas J. Sargent (2001). "Robust Control and Model Uncertainty". In: *The American Economic Review* 91.2, pp. 60–66. ISSN: 00028282. URL: http://www.jstor.org/stable/2677734 (visited on 03/29/2026).

Korpelevich, G. M. (1976). "The extragradient method for finding saddle points and other problems". In: URL: https://api.semanticscholar.org/CorpusID:118602977.