Optimization-based Statistical Inference: constrained inverse problems, worst-case priors, and kernel regression

Thesis by
Pau Batlle Franch

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 27, 2025

Pau Batlle Franch
ORCID: 0000-0003-4886-058X

To my parents and my sister

“It’s only when you look at an ant through a magnifying glass on a sunny day that you realize how often they burst into flames.” —Harry Hill.

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my advisor, Houman Owhadi, for his help and support, both academically and financially, through all these years and the different roadblocks and successes of the journey. Thanks as well to the professors serving on my candidacy and thesis committees, Andrew Stuart, Franca Hoffmann, Joel Tropp, and Amy Braverman, for their time and helpful feedback.

Many thanks to all of my collaborators throughout the years; every project has been a learning experience, and I am grateful for the time and effort everyone has given to such projects. Special thanks to Mike Stanley, who first introduced me to the field of constrained inverse problems, a field I deeply enjoyed working in for the last three years. My journey would surely not have been the same had I never encountered such research problems, and I am quite proud of the joint effort we put in advancing that field. Thank you as well to Pratik Patil, Mikael Kuusela, and Javier Ruiz-Lupon for their help and collaboration in this project.

Thanks to Bamdad Hosseini, Yifan Chen, and Matthieu Darcy for countless research and casual discussions about the intricacies of Gaussian Processes. Similarly, discussing UQ and Kernel Flows with Jouni Susiluoto, Otto Lamminpaeae, and Amy Braverman from the Uncertainty Quantification group at JPL has been extremely enjoyable and motivating to pursue my research in the area. Thank you to Michael Hecht for hosting me as a visiting researcher in CASUS, Germany, and to Boumediene Hamzi for the opportunity to collaborate with Alpha Lambda Iota.

On a personal level, I would like to thank all of my family for their support, especially my parents, Pere and Anna, and my sister Carla, whose optimism and cheerful energy motivate me to be better and happier every day.

I am extremely grateful to have met wonderful people during my PhD, whether at conferences or at Caltech, whom I am very happy to consider good friends. I would like to give special thanks to Matthieu Darcy, Eitan Levin, and Joe Slote for all the support and help through all the high highs and low lows of this period of my life. It has been an immense pleasure to share all the experiences we shared through the years.

Many thanks as well to many other friends and colleagues who I have shared plenty of good times and experiences with: Theo Bourdais, Edoardo Calvello, Siki Wang, Dohyeon Kim, Chris Yeh, Roy Wang, Margaret Trautner, John Lathrop, Lauren

Conger, Sebastian Lamas, Matt Levine, Francesca Li, Janina Schreiber, Jamiree Harrison, and Nina Fischer, thanks to every one of you for all the unforgettable moments. Thanks as well to my university colleagues Pol Mestres, Adam Teixidó, and Tomàs Ortega for the adventures around Southern California, and to my friends back home Andreu Vergés, Pau Castillo, and Xavier Rubiés, for sweetening my periodic visits through games of padel and late-night dinners.

ABSTRACT

Optimization provides a worst-case framework for quantifying uncertainty in statistical inference, delivering robust and transparent performance guarantees. While this approach provides rigorous bounds, it cannot easily incorporate large-scale data or produce estimates at a prescribed confidence level. To bridge this gap, this thesis develops optimization-based methods that assimilate data while retaining worst-case robustness, exploring three different contexts: Ill-posed inverse problems, Bayesian inference with unknown priors, and Gaussian process regression.

In the first, we introduce a new framework for frequentist, optimization-based intervals that provably achieves desired coverage. The framework unifies many previously proposed optimization-based intervals and disproves a conjecture dating back to 1965. In the second, we introduce data-likelihood constraints in Wald's two-player zero-sum game, which renders the game computationally tractable and provides explicit certificates of minimax optimality. In the third, we develop new Gaussian process (GP) based methods for learning and solving partial differential equations and operator learning. In each setting, our GP algorithms achieve stronger convergence guarantees than existing machine-learning techniques without sacrificing predictive accuracy.

Across these three settings, estimates for the unknown quantity (a finite-dimensional parameter, a prior distribution, or a function, respectively) are obtained as the solution to an optimization problem that characterizes either worst-case or minimax optimality, therefore contributing towards a single optimization-centric view of uncertainty quantification.

PUBLISHED CONTENT AND CONTRIBUTIONS

Stanley, Michael, Pau Batlle, Pratik Patil, Houman Owhadi, and Mikael Kuusela (2025). "Confidence intervals for functionals in constrained inverse problems via data-adaptive sampling-based calibration". In: *arXiv preprint arXiv:2502.02674*.
P.B contributed to conceptualization, proof of the theoretical results, and numerical experiments.

Batlle, Pau, Matthieu Darcy, Bamdad Hosseini, and Houman Owhadi (2024). "Kernel methods are competitive for operator learning". In: *Journal of Computational Physics* 496.
P.B and M.D share equal contribution as main contributors of this work, with contributions to conceptualization, writing, and numerical results., p. 112549. DOI: https://doi.org/10.1016/j.jcp.2023.112549.

Kaveh, Hojjat, Pau Batlle, Mateo Acosta, Pranav Kulkarni, Stephen J Bourne, and Jean Philippe Avouac (2024). "Induced seismicity forecasting with uncertainty quantification: Application to the Groningen gas field". In: *Seismological Research Letters* 95.2A.
P.B contributed to the conceptualization of the Uncertainty Quantification algorithm and writing of its corresponding section, pp. 773–790. DOI: https://doi.org/10.1785/0220230179.

Pandey, Biraj, Bamdad Hosseini, Pau Batlle, and Houman Owhadi (2024). "Diffeomorphic Measure Matching with Kernels for Generative Modeling". In: *SIAM Journal on Mathematics of Data Science (to appear)*.
P.B contributed to conceptualization, early prototyping, and editing.

Schreiber, Janina, Pau Batlle, Damar Wicaksono, and Michael Hecht (2024). "PMBO: Enhancing Black-Box Optimization through Multivariate Polynomial Surrogates". In: *arXiv preprint arXiv:2403.07485*.
P.B contributed to the conceptualization, early prototyping, and editing.

Batlle, Pau, Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart (2023). "Error analysis of kernel/GP methods for nonlinear and parametric PDEs". In: *Journal of Computational Physics*.
P.B contributed to numerical results of Section 4, and writing. DOI: https://doi.org/10.1016/j.jcp.2024.113488.

Batlle, Pau, Pratik Patil, Michael Stanley, Houman Owhadi, and Mikael Kuusela (2023). "Optimization-based frequentist confidence intervals for functionals in constrained inverse problems: Resolving the Burrus conjecture". In: *arXiv preprint arXiv:2310.02461*.
P.B is the main contributor of this work, and contributed to the conceptualization, investigation, proof of the main mathematical results, numerical experiments, and writing.

Bourdais, Théo, Pau Batlle, Xianjin Yang, Ricardo Baptista, Nicolas Rouquette, and Houman Owhadi (2023). "Codiscovering graphical structure and functional relationships within data: A Gaussian Process framework for connecting the dots". In: *Proceedings of the National Academy of Sciences (PNAS)*.
P.B contributed to conceptualization, early prototyping, and writing. DOI: https://doi.org/10.1073/pnas.2403449121.

Dingle, Kamaludin, Pau Batlle, and Houman Owhadi (2023). "Multiclass classification utilising an estimated algorithmic probability prior". In: *Physica D: Nonlinear Phenomena* 448.
P.B contributed to the conceptualization of the Gaussian process-based algorithm and writing revision, p. 133713. DOI: https://doi.org/10.1016/j.physd.2023.133713.

Bajgiran, Hamed Hamze, Pau Batlle, Houman Owhadi, Mostafa Samir, Clint Scovel, Mahdy Shirdel, Michael Stanley, and Peyman Tavallali (2022). "Uncertainty quantification of the 4th kind; optimal posterior accuracy-uncertainty tradeoff with the minimum enclosing ball". In: *Journal of Computational Physics* 471.
P.B contributed to the conceptualization, writing, and numerical experiments in Sections 1 and 6., p. 111608. DOI: https://doi.org/10.1016/j.jcp.2022.111608.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

A crucial part of the scientific process is the combination of observations of real-world data with previously known knowledge about a given physical process, to obtain a theory that can generalize and predict future events. Given the current scale of available computing power and data, classical scientific discovery has been increasingly complemented with computational and data-driven statistical tools. In this context, statistical inference methodologies aim to provide uncertainty estimates in the process of recovering unknowns from a combination of data and previous information.

This thesis explores theory and applications of optimization-based perspectives on statistical inference, extending its original formulation in the context of worst-case Uncertainty Quantification into a more modern setup in which it aims to produce statistical estimates while preserving the certification properties of the optimization-based method.

We consider different statistical inference problems and applications that can be described with the additive noise model $y = F(x^*) + \varepsilon$, in which y is the observed experimental data, F (the *forward model*) is a physical process, dependent on a state of the world or parameter x^*, and $\varepsilon \sim \mu$ is some experimental or measurement noise.

The scientist's knowledge about the different elements of the problem is application-dependent. Statistical techniques vary depending on which elements we aim to infer from which others, and under which assumptions/knowledge of F, x^* and μ, the distribution of ε, we aim to perform such inference. For example, distribution-free methods make no assumptions about the distribution of ε, frequentist methods make distribution assumptions on ε but not on x^*, constrained inference methods further assume $x^* \in \mathcal{X}$, and Bayesian methods assume $x^* \sim \nu$ for a given, known ν. For all the elements present (including the prior distribution ν, whenever included as part of the model), a range of assumptions is possible, including fully unknown, knowledge of a constraint set, and fully known.

There is an inherent tradeoff between accuracy and robustness in Uncertainty Quantification methods, illustrated in this context by the strength of the assumptions any particular method uses: methods with few assumptions are robust, but may overcover

if they are unable to capture all that is known from a problem, while methods with many assumptions can be much more accurate, but might suffer from brittleness and undercover if any of the assumed knowledge is even slightly wrong (H. Owhadi, C. Scovel, and T. Sullivan, 2015b).

In real applications, overcovering leads to the extra economic cost of preparing against scenarios that were not needed, and undercovering leads to taking unnecessary risks. Given these potential consequences of over-assuming and under-assuming in inference problems, and given that many real problems have partial knowledge about the unknowns, this thesis aims to build optimization-based methods that are able to include such partial knowledge, without relying on extra information that is not available to the decision problem.

An example inference problem that will help illustrate the importance of assumptions is the particle unfolding problem (Spano, 2014), illustrated in Figure 1.1. The distribution of particle counts with respect to a quantity of interest (usually, energy) in certain particle physics cannot be observed directly, but is observed through a noisy measurement of a known "blurring" process F that destroys some part of the signal. We refer to this as an ill-posed inverse problem because, even with a noiseless measurement, we would not be able to recover the totality of x^*. It is nevertheless of scientific interest to estimate a function of x^*, referred to as $\varphi(x^*)$ (for example, the number of particles over a given energy threshold), given the data y.

In this problem, we can model F and the distribution of ε as known (we assume that the measurement device can be calibrated in-lab before the experiment), but a prior distribution on x^*, which is a common regularization technique for ill-posedness in inverse problems, might not necessarily be available. Early works dating back to the 1960s proposed using the known constraint $x^* \geq 0$ to obtain meaningful inference without prior regularization (W. R. Burrus, 1965), a problem which we consider in more generality in Chapters 2 and 3. In this setup, not being able to include the constraint $x^* \geq 0$ (under-assuming) can make the given function of interest $\varphi(x^*)$ hard or even impossible to infer, and adding a prior distribution (over-assuming) might lead to intervals without the proper coverage if such a prior is not correct, as observed in a related application (Patil, Kuusela, and Hobbs, 2022). It is hence crucial to develop algorithms that can make use of "what we know, and only what we know".

More precisely, in Chapters 2 and 3, we consider building frequentist confidence intervals for functionals of interest in constrained inverse problems, which corre-

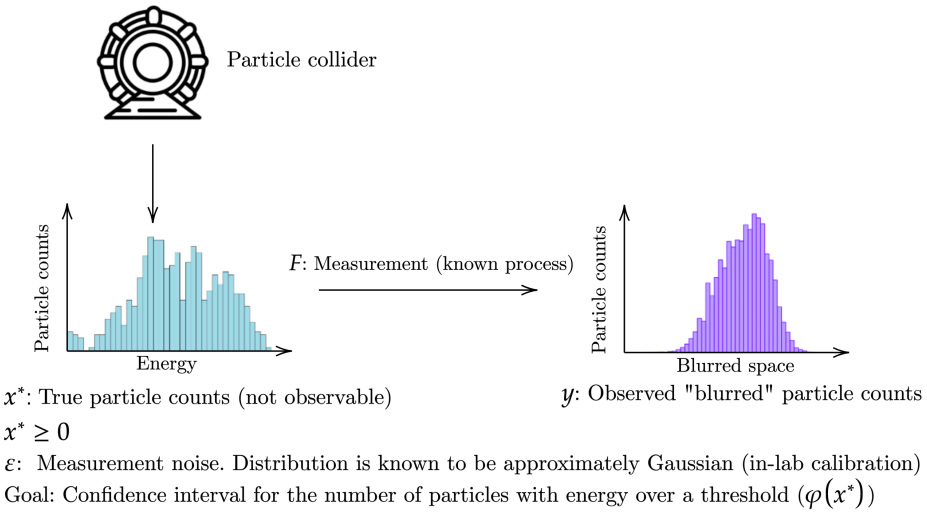


Figure 1.1: Setup of the particle unfolding inference problem. In it, we are aiming to recover a function φ of a discrete distribution of particle counts x^*, given a distribution y of blurred particle counts obtained by a noisy measurement $y = F(x^*) + \varepsilon$.

sponds to using the data y and knowledge of F, the distribution of ε and some constraints $x^* \in \mathcal{X}$ to calibrate in a frequentist sense a confidence interval for a quantity of interest $\varphi(x^*)$. In this problem, standard worst-case analysis shows that before observing any data, it holds that

$$\varphi(x^*) \in \left[\inf_{x \in \mathcal{X}} \varphi(x), \sup_{x \in \mathcal{X}} \varphi(x) \right] \quad \forall x^* \in \mathcal{X}. \tag{1.1}$$

If we were to observe the data $y = F(x)$ without noise, defining $C(y) := \{x : F(x) = y\}$ we would be able to certify that

$$\varphi(x^*) \in \left[\inf_{x \in \mathcal{X} \cap C(y)} \varphi(x), \sup_{x \in \mathcal{X} \cap C(y)} \varphi(x) \right] \quad \forall x^* \in \mathcal{X}. \tag{1.2}$$

In the noisy case, in which we observe $y = F(x) + \varepsilon$, with ε sampled from a noise distribution, the main problem addressed in Chapters 2 and 3 is how to build an equivalent set $C(y; \alpha)$ such that the property (1.2) holds with probability $1 - \alpha$ i.e.

$$\mathbb{P}_{y \sim P_{x^*}} \left(\varphi(x^*) \in \left[\inf_{x \in \mathcal{X} \cap C(y;\alpha)} \varphi(x), \sup_{x \in \mathcal{X} \cap C(y;\alpha)} \varphi(x) \right] \right) \geq 1 - \alpha \quad \forall x^* \in \mathcal{X}. \tag{1.3}$$

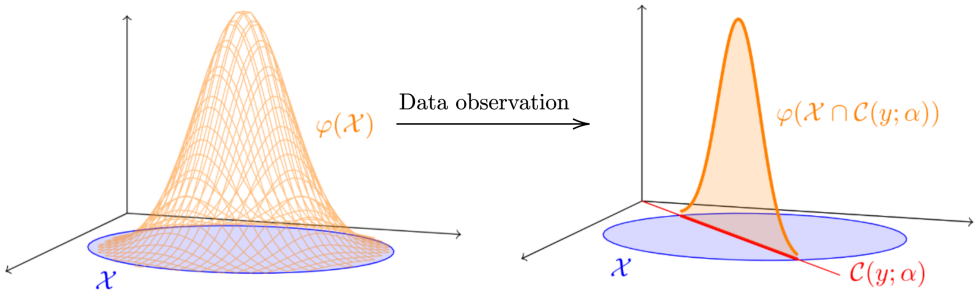


Figure 1.2: In optimization-based methods, the observation of the data can be used to shrink the constraint set of the optimization programs considered

while having the resulting interval be as small as possible. We note that we treat x^* as a fixed unknown and the probability is only taken with respect to the randomness induced by ε. Figure 1.2 illustrates how the process of data observation leads to a shrinkage of the set we are optimizing over, from $\mathcal{X}$ to $\mathcal{X} \cap C(y;\alpha)$, leading to potentially tighter Uncertainty Quantification. Chapter 2, published as (Batlle, Stanley, et al., 2023), focuses on the theoretical background of this problem, illustrating the connections with hypothesis test inversion of many different previously used techniques (W. R. Burrus, 1965; Philip B. Stark, 1992a; Philip B. Stark, 1994) and disproving a conjecture originally stated in (W. R. Burrus, 1965) claiming that a particular construction of $C(y;\alpha)$ has correct $1-\alpha$ frequentist coverage. Chapter 3, published as (Stanley, Batlle, et al., 2025), focuses on building a sampling algorithm that scales to higher-dimensional problems, by leveraging quantile regression techniques and maximization of p-values over confidence sets, a theoretical technique pioneered originally by (Roger L. Berger and Boos, 1994).

In Chapter 4, published as (Bajgiran et al., 2022a), inverse problems are considered from a Bayesian and decision-theoretical perspective, with optimization and certification being performed with respect to an unknown prior distribution. This is, we consider that, before sampling our data, $x^* \sim \mu$ for some unknown distribution μ, and we aim to find certificates similar to (1.3).

As opposed to the worst-case approach with respect to measures developed in (H. Owhadi, C. Scovel, T. J. Sullivan, et al., 2013b), we aim to find minimax optimal certificates in the setting of the classical decision-theoretical framework of (A. Wald, 1945), in which a prior distribution is chosen by nature adversarially to the decision-maker as a randomized strategy for a zero-sum game. In such a game, we introduce likelihood constraints after seeing the data. This corresponds to a novel perspective in Uncertainty Quantification, which we name *of the fourth kind* (in

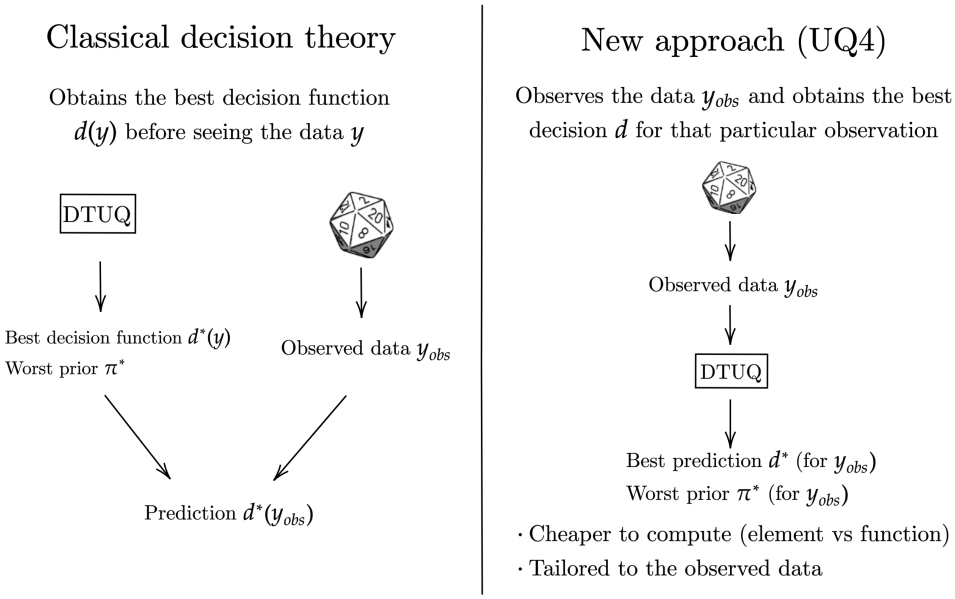


Figure 1.3: In Optimization-based methods, the observation of the data can be used to shrink the constraint set of the optimization programs considered

reference to the three previous families: worst-case, Bayesian, and the aforementioned Wald's decision-theoretical), which consists of formulating and solving a two-player zero-sum data-dependent game subject to likelihood constraints. Figure 1.3 illustrates the main differences between the classical decision-theoretical UQ and our new approach. This algorithm has been later used in (Kaveh et al., 2024) for the quantification of uncertainty in the number of earthquakes induced by reservoir operations for gas extraction in the Groningen region in the Netherlands.

In Chapters 5, 6, and 7, published as (Batlle, Darcy, et al., 2023; Batlle, Yifan Chen, et al., 2023; Bourdais et al., 2023) respectively, three different *functional regression* settings are considered, in which we aim to recover the forward model F based on pairs of $\{x_i, y_i = F(x_i) + \varepsilon_i\}_{i\in\mathcal{I}}$ data. In all the situations, we consider kernel methods, which emerge in the context of optimization-based statistical inference as solutions of an optimal recovery game in a Banach space $\mathcal{B}$. In particular, letting an unknown $u \in \mathcal{B}$ we consider the problem of recovering it from linear measurements $\Phi(u) := ([\phi_1, u], \ldots, [\phi_m, u]) \in \mathbb{R}^m$, with an estimator $\Psi(\Phi(u)) \in \mathcal{B}$. If we pose this as a zero-sum game with loss function $\mathcal{E}(u, \Psi) = \dfrac{\|u - \Psi(\Phi(u))\|_{\mathcal{B}}}{\|u\|_{\mathcal{B}}}$ (which the decision maker aims to minimize by choosing Ψ, and nature aims to minimize by adversarially choosing u), Gaussian processes conditioned on the data, equivalently thought of as kernel methods by seeing kernels as Gaussian process covariance

functions, arise as minimax optimal solutions of the randomized version of the optimal recovery game (Houman Owhadi and Clint Scovel, 2019a).

Chapter 5 considers using kernel methods for operator learning, the task of learning an operator between Banach spaces of functions. In Chapter 6, error estimates for the use of kernel methods to solve partial differential equations are developed, and in Chapter 7, we use Gaussian processes as a sensing tool to discover functional relationships between data. This task can be thought of as recovering particular properties about the unknown f, in cases in which recovering the full f is unfeasible due to the quantity or quality of the available data. In the sequel, each chapter defines distinct notation within its content.

Chapter 2

OPTIMIZATION-BASED FREQUENTIST CONFIDENCE INTERVALS FOR FUNCTIONALS IN CONSTRAINED INVERSE PROBLEMS: RESOLVING THE BURRUS CONJECTURE

We present an optimization-based framework to construct confidence intervals for functionals in constrained inverse problems, ensuring valid one-at-a-time frequentist coverage guarantees. Our approach builds upon the now-called strict bounds intervals, originally pioneered by (W. R. Burrus, 1965; Burt W. Rust and Walter R. Burrus, 1972), which offer ways to directly incorporate any side information about the parameters during inference without introducing external biases. This family of methods allows for uncertainty quantification in ill-posed inverse problems without needing to select a regularizing prior. By tying optimization-based intervals to an inversion of a constrained likelihood ratio test, we translate interval coverage guarantees into type I error control and characterize the resulting interval via solutions to optimization problems. Along the way, we refute the Burrus conjecture, which posited that, for possibly rank-deficient linear Gaussian models with positivity constraints, a correction based on the quantile of the chi-squared distribution with one degree of freedom suffices to shorten intervals while maintaining frequentist coverage guarantees. Our framework provides a novel approach to analyzing the conjecture, and we construct a counterexample employing a stochastic dominance argument, which we also use to disprove a general form of the conjecture. We illustrate our framework with several numerical examples and provide directions for extensions beyond the Rust–Burrus method for nonlinear, non-Gaussian settings with general constraints.

2.1 Introduction

Advances in data collection and computational power in recent years have led to an increase in the prevalence of high-dimensional, ill-posed inverse problems, especially within the physical sciences. These challenges are particularly evident in domains such as remote sensing and data assimilation, where uncertainty quantification (UQ) in inverse problems is of paramount importance. Many of these inverse problems also come with inherent physical constraints on their parameters. This paper focuses on constrained inverse problems for which the noise model is

known, and the forward model, defined on a finite-dimensional parameter space, can be computationally evaluated. Our primary objective is to construct a confidence interval for a functional of the forward model parameters.

Formally, we consider statistical models of the form $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$, where $\boldsymbol{y} \in \mathbb{R}^m$ is sampled according to a parametric probability distribution. Here $\boldsymbol{x}^* \in \mathbb{R}^p$ is a fixed unknown parameter, which we know a priori lies within the set $\mathcal{X}$; see Figure 2.1 for an illustration. Our goal is to construct confidence intervals for a known one-dimensional functional $\varphi(\boldsymbol{x}^*) \in \mathbb{R}$. Ideally, we want the length of these intervals to be as small as possible, while still maintaining a nonasymptotic frequentist coverage guarantee. In other words, given a prescribed coverage level $1-\alpha$ for some $\alpha \in (0, 1)$, we want to construct functions of the data $I^-(\boldsymbol{y})$ and $I^+(\boldsymbol{y})$ such that the following coverage guarantee holds in finite sample[1]:

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}}}\big(\varphi(\boldsymbol{x}) \in [I^-(\boldsymbol{y}), I^+(\boldsymbol{y})]\big) \geq 1 - \alpha. \tag{2.1}$$

While the requirement (2.1) requires that we maintain at least $1-\alpha$ coverage, we also want it to be approximately accurate by minimizing the slack in the inequality. Ensuring such *proper calibration*, namely, confidence intervals that do not *undercover* (fail to meet the $1 - \alpha$ guarantee for some $\boldsymbol{x} \in \mathcal{X}$) or *overcover* (are too large and therefore exceed the required coverage) is paramount in practical applications. This is especially true in contexts that require stringent safety and certification standards. Intervals that undercover yield unreliable inferences that may expose the system to unforeseen risks. Conversely, intervals that overcover might lead to excessive economic costs by needing to guard against scenarios that are unlikely to occur.

In many applied contexts, Bayesian methods constitute a primary set of techniques for uncertainty quantification. These methods leverage a prior for regularization, derived either from the intrinsic details of the problem or introduced externally. A key advantage of this regularization approach is the natural UQ that emerges from the Bayesian statistical framework. Specifically, the combination of a predefined prior and data likelihood results in a posterior distribution via Bayes' theorem. This distribution can subsequently be used to derive the intended posterior UQ. However, there is a caveat: Bayesian methods can offer *marginal coverage* (probability over $\boldsymbol{x}$ and $\boldsymbol{y}$) if the prior is correctly specified. They do not necessarily provide *conditional coverage* (probability over $\boldsymbol{y}$ given $\boldsymbol{x}$). The former notion of coverage is weaker (and,

[1] This form of "simple" interval is only for expositional simplicity. One can consider more general forms of confidence sets $\mathcal{I}(\boldsymbol{y})$ beyond simple intervals, which we will do when describing the general framework in Section 2.3.

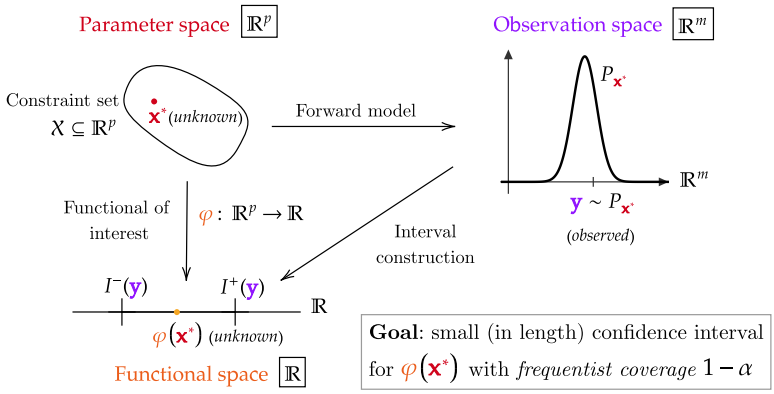


Figure 2.1: Illustration of the problem setup. We seek to construct confidence intervals $[I^-(\boldsymbol{y}), I^+(\boldsymbol{y})] \subseteq \mathbb{R}$ for $\varphi(\boldsymbol{x}^*) \in \mathbb{R}$ from an observation $\boldsymbol{y} \in \mathbb{R}^m$ sampled from $P_{\boldsymbol{x}^*}$ that satisfies a frequentist coverage guarantee in finite sample while being as small (in length) as possible.

in particular, the latter implies the former, but the converse may not be true), as it replaces the infimum in the coverage requirement (2.1) with a probability distribution over $\boldsymbol{x}$. Generally, Bayesian methods may not align with the analyst's expectations due to inherent bias (Kuusela, 2016; Patil, Kuusela, and Hobbs, 2022). While, in theory, priors present an effective mechanism to incorporate scientific knowledge into UQ, they can inadvertently introduce extraneous information (Philip B. Stark, 2015) and a lack of robustness in the resulting estimates (Houman Owhadi, Clint Scovel, and Tim Sullivan, 2015a; Houman Owhadi, Clint Scovel, and Tim Sullivan, 2015b; H. Owhadi and C. Scovel, 2017b).

On the other hand, we could consider a basic worst-case approach that is rooted in the simple observation that

$$\varphi(\boldsymbol{x}^*) \in \left[\inf_{\boldsymbol{x} \in \mathcal{X}} \varphi(\boldsymbol{x}),\ \sup_{\boldsymbol{x} \in \mathcal{X}} \varphi(\boldsymbol{x}) \right]. \tag{2.2}$$

Of course, this method is inherently conservative given the absence of assumptions and any specific knowledge regarding data generation. More importantly, the method does not use observations y in any way to calibrate the confidence set. This means that the sets cannot be fine-tuned to approximately achieve the desired $1-\alpha$ coverage level. Nevertheless, they illustrate the essential idea of constructing a confidence interval based on the outcomes of two boundary optimization problems, an approach that the more sophisticated methods that we will study in this paper build on. We

shall henceforth refer to such intervals with the notation:

$$\begin{array}{ll} \inf_{\boldsymbol{x}}/\sup_{\boldsymbol{x}} & \varphi(\boldsymbol{x}) \\ \text{st} & \boldsymbol{x} \in \mathcal{X} \end{array} := \left[\inf_{\boldsymbol{x} \in \mathcal{X}} \varphi(\boldsymbol{x}),\ \sup_{\boldsymbol{x} \in \mathcal{X}} \varphi(\boldsymbol{x}) \right].$$

An example of such a more sophisticated method is the so-called "simultaneous" approach (Philip B. Stark, 1992a; Philip B. Stark, 1994), which provides intervals with at least $1 - \alpha$ frequentist coverage for the functional of interest $\varphi(\boldsymbol{x}^*)$ from confidence sets for the parameter $\boldsymbol{x}^*$. The approach can be summarized in three steps (see Figure 2.2 for an illustration):

Step 1. Construct a set $C(\boldsymbol{y}) \in \mathbb{R}^p$ that is a $1 - \alpha$ confidence set for $\boldsymbol{x}^*$.

Step 2. Intersect this set $C(\boldsymbol{y})$ with the constraint set $\mathcal{X}$.

Step 3. Project this intersection through the functional of interest φ.

The term "simultaneous" refers to Steps 1 and 2 being independent of the quantity of interest φ, so the resulting set from Step 2 can be simultaneously projected to different quantities of interest. Under mild assumptions, the resulting intervals can be equivalently written as

$$\mathcal{I}_{\mathsf{SSB}}(\boldsymbol{y}) := \left[\inf_{\boldsymbol{x} \in \mathcal{X} \cap C(\boldsymbol{y})} \varphi(\boldsymbol{x}),\ \sup_{\boldsymbol{x} \in \mathcal{X} \cap C(\boldsymbol{y})} \varphi(\boldsymbol{x}) \right] = \begin{array}{ll} \inf_{\boldsymbol{x}}/\sup_{\boldsymbol{x}} & \varphi(\boldsymbol{x}) \\ \text{st} & \boldsymbol{x} \in \mathcal{X} \cap C(\boldsymbol{y}). \end{array} \tag{2.3}$$

This illustrates how the simultaneous approach is a refinement of the basic worst-case method (2.2): the observation of the data $\boldsymbol{y}$ shrinks the "pre-data set" $\mathcal{X}$ into a smaller "post-data set" $\mathcal{X} \cap C(\boldsymbol{y})$, which is then projected through φ in a worst-case manner. Given that this simultaneous framework is broadly encapsulated in (Philip B. Stark, 1992a) as "strict bounds," we label these intervals as "simultaneous strict bounds" or SSB intervals, for short.

Unlike methods that rely on explicit regularization through a prior, the techniques outlined above leverage only the physical constraints and the functional of interest to address the underlying ill-posedness of the inverse problem. This approach allows for uncertainty quantification without the need to assume a prior distribution, circumventing potential biases and miscalibrated coverage issues previously mentioned.

Although the interval (2.3) has guaranteed coverage for $\varphi(\boldsymbol{x}^*)$ inherited from the coverage of $C(\boldsymbol{y})$, this method generally suffers from overcoverage, especially when

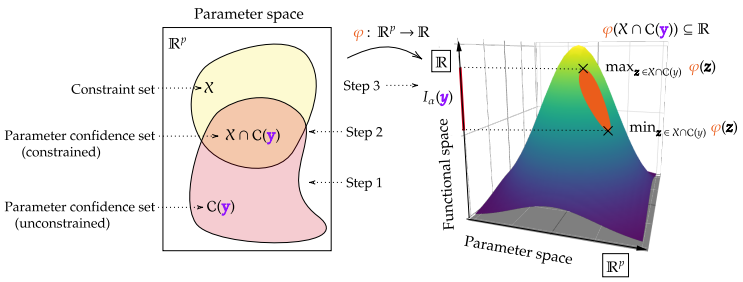


Figure 2.2: Illustration of the simultaneous approach for confidence interval building, which works generically for any $\mathcal{X}$, φ, and P. The intersection of $\mathcal{X}$ and $C(\boldsymbol{y})$ occurs in the original parameter space $\mathbb{R}^p$, and is then projected via the functional of interest function into the real line. The confidence interval is then constructed using the minimum and maximum of the quantity of interest φ over the intersection $\mathcal{X} \cap C(\boldsymbol{y})$.

the dimension of $\mathcal{X}$ is large (Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022; Kuusela and Philip B. Stark, 2017a). This happens due to two main factors: (i) its generality cannot account for the specific structure of P, φ, and $\mathcal{X}$; and (ii) while the set $C(\boldsymbol{y})$ being a $1 - \alpha$ confidence set is a sufficient condition, it is not necessary for (2.3) to ensure accurate coverage, which implies that smaller sets might also produce valid confidence intervals. Consequently, an important research direction has been constructing confidence intervals that are shorter than the simultaneous approach, but still maintain nominal coverage for a given φ. Sometimes, this is achieved by assuming that P, φ, and $\mathcal{X}$ come from a particular class (Philip B. Stark, 1994; Burt W. Rust and Walter R. Burrus, 1972; Tenorio, Fleck, and Moses, 2007; Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022). In the sequel, we discuss one such special class.

The Burrus conjecture

The Gaussian linear forward model with nonnegativity constraints and a linear functional of interest is a setting that has attracted significant attention, going back to the works of (W. R. Burrus, 1965; Burt W. Rust and Walter R. Burrus, 1972). These foundational studies consider the applied problem of unfolding gamma-ray and neutron spectra from pulse-height distributions under rank-deficient linear systems. They demonstrated that incorporating the nonnegativity physical constraint allowed for the computation of nontrivial (i.e., finite length) intervals for linear functionals of the parameters. In order to describe the construction of these intervals, consider the canonical form of the Gaussian linear model with nonnegativity constraints, along

with a linear functional of interest:

$$\underbrace{\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)}_{\text{model}}, \quad \text{with} \quad \underbrace{\boldsymbol{x}^* \geq \boldsymbol{0}}_{\text{constraints}} \quad \text{and} \quad \underbrace{\varphi(\boldsymbol{x}^*) = \boldsymbol{h}^\top \boldsymbol{x}^*}_{\text{functional}}. \tag{2.4}$$

Here, $\boldsymbol{K} \in \mathbb{R}^{m \times p}$ is the forward operator[2], $\boldsymbol{x}^* \in \mathbb{R}^p$ is the true parameter vector, and $\boldsymbol{h} \in \mathbb{R}^p$ contains weights for the functional of interest. In this setting, (W. R. Burrus, 1965; Burt W. Rust and Walter R. Burrus, 1972) posed that the following interval construction yields valid $1 - \alpha$ confidence intervals, a result now known as the *Burrus conjecture* (Bert W. Rust and O'Leary, 1994):

$$\mathcal{I}_{\mathsf{OSB}}(\boldsymbol{y}) := \begin{array}{lll} \min\limits_{\boldsymbol{x}} / \max\limits_{\boldsymbol{x}} & \boldsymbol{h}^\top \boldsymbol{x} & \\ \text{st} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 \leq \psi_\alpha^2(\boldsymbol{y}), & \\ & \boldsymbol{x} \geq \boldsymbol{0}, & \end{array} \tag{2.5}$$

where $\psi_\alpha^2 = z_{\alpha/2}^2 + s^2(\boldsymbol{y})$. Here z_α is the upper quantile of standard normal such that $\mathbb{P}(Z > z_\alpha) = \alpha$ for $Z \sim \mathcal{N}(0, 1)$, and $s^2(\boldsymbol{y})$ is defined through an optimization problem as follows:

$$s^2(\boldsymbol{y}) := \begin{cases} \min\limits_{\boldsymbol{x}} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 \\ \text{st} & \boldsymbol{x} \geq \boldsymbol{0}. \end{cases} \tag{2.6}$$

Comparison of (2.5) with (2.3) shows that Rust and Burrus proposed a "simultaneous-like" construction. In this construction, the set $\{\boldsymbol{x} : \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 \leq \psi_\alpha^2(\boldsymbol{y})\}$ plays the role of $C(\boldsymbol{y})$. It typically does not represent a $1-\alpha$ confidence set for $\boldsymbol{x}^*$, thus relaxing the stringent assumption of the SSB interval construction. Furthermore, a possible simultaneous interval for this setting can be built by observing that $\|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}^*\|_2^2 \sim \chi_m^2$. This yields the following valid $1 - \alpha$ interval:

$$\begin{array}{lll} \min\limits_{\boldsymbol{x}} / \max\limits_{\boldsymbol{x}} & \boldsymbol{h}^\top \boldsymbol{x} & \\ \text{st} & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 \leq Q_{\chi_m^2}(1 - \alpha) & \\ & \boldsymbol{x} \geq \boldsymbol{0}. & \end{array} \tag{2.7}$$

Here, $Q_{\chi_m^2}$ is the quantile function of a χ_m^2 distribution. It should be noted that the data-dependent term $\psi_\alpha^2(\boldsymbol{y})$ in (2.5) could be considerably smaller than $Q_{\chi_m^2}(1 - \alpha)$, especially when m is large and α is small. So if the Burrus conjecture were true,

[2]Note that the forward operator $\boldsymbol{K}$ is allowed to be column rank deficient and the overparameterized setting when $p > m$ is allowed.

it would provide a significant reduction in the length of the interval for problems in the class (2.4). For instance, assuming $\alpha = 0.05$ (so that we are after a 95% coverage level), (Stanley, Patil, and Kuusela, 2022) observe an expected length reduction of about a factor of two across a variety of functionals in a particle unfolding application. The gain in the interval length originates from the fact that these intervals take into account that we are only required to guarantee coverage for *one specific* functional. Given that intervals of the form (2.5) are designed to provide coverage for one functional at a time, following the nomenclature of (Stanley, Patil, and Kuusela, 2022), we refer to these intervals as "one-at-a-time strict bounds" or OSB intervals, for short[3].

(Burt W. Rust and Walter R. Burrus, 1972) and subsequently (Bert W. Rust and O'Leary, 1994) investigated the conjecture posed in (W. R. Burrus, 1965), with (Bert W. Rust and O'Leary, 1994) purpoting to have found definitive proof for the conjecture's validity. However, this claim was later refuted by (Tenorio, Fleck, and Moses, 2007) through a two-dimensional counterexample. In this work, we demonstrate that, in fact, this two-dimensional counterexample proposed in (Tenorio, Fleck, and Moses, 2007) is not a valid counterexample. However, we present and prove another counterexample that refutes the conjecture and we propose ways to fix the previous faulty results by reinterpreting the conjecture. We achieve this through a novel hypothesis test-based framework that not only revisits but also broadens the scope beyond the linear Gaussian setting paired with positivity constraints in which the conjecture was originally proposed.

Summary and outline

In this paper, we frame the problem of confidence interval construction for functionals in constrained, ill-posed problems through the inversion of a particular likelihood ratio test. This perspective allows us to reinterpret the interval coverage guarantee in terms of type-I error control associated with the test and, subsequently, the distribution of the log-likelihood ratio under the null hypothesis. We also establish connections between different fields of hypothesis testing with likelihood ratio tests, optimization-based confidence intervals, and chance-constrained optimization. A detailed summary of contributions in this paper along with an outline for the paper

[3] (Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022) also extend the setting and the conjecture to encompass linear constraints of the form $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$. Such constraints are of interest in practical applications such as X_{CO2} retrieval and particle unfolding. For simplicity, we present only the positivity constraint case here. However, our counterexample based on positivity constraints in Section 2.4 will also be sufficient to disprove the conjecture in this general case.

is given below.

(1) **Strict bounds intervals from test inversion.** In Section 2.2, we present a general framework to construct strict bounds intervals through test inversion, resulting in two optimization problems for the interval endpoints. This approach generalizes the Rust–Burrus-type interval technique to potentially nonlinear and non-Gaussian settings. Our main result in Theorem 2.2.4 proves coverage of the test inversion construction and Proposition 2.2.5 provides sufficient conditions under which the coverage is tight. Examples in Section 2.2 provide straightforward but concrete analytical illustrations of our framework.

(2) **General interval construction methodology.** In Section 2.3, we present a general methodology for computing confidence intervals that builds on the framework in Section 2.2. We outline the methodology in Algorithm 1 and discuss two key components: the chance-constrained optimization problem and the stochastic dominance argument in Section 2.3 and Section 2.3, respectively. The chance-constrained optimization problem allows us to obtain optimal decision values for the proposed framework, while the stochastic dominance argument provides a theoretical tool to find provable upper bounds.

(3) **Refuting the Burrus conjecture.** In Section 2.4, we demonstrate that our method successfully recovers previously proposed OSB intervals for the linear Gaussian setting. In Theorem 2.4.1, we leverage this novel interpretation to disprove the Burrus conjecture (Burt W. Rust and Walter R. Burrus, 1972; Bert W. Rust and O'Leary, 1994) in the general case, by refuting a previously proposed counterexample and providing a new, provably correct counterexample in Lemma 2.4.5. Furthermore, we provide a negative result that disproves a natural generalization of the original conjecture in Proposition 2.4.6. Our proof technique provides a method to detect when the Rust–Burrus approach is effective and when it falls short and introduces a means to rectify the earlier erroneous examples.

(4) **Illustrative numerical examples.** In Section 2.5, we elucidate our findings through a suite of numerical illustrations. These span various scenarios, including the counterexample to the Burrus conjecture. We show that test inversion-based intervals, which have provable guarantees, achieve better coverage calibration than previous approaches.

Other related work

Given the effectiveness of the strict bounds methodology in high-dimensional ill-posed inverse problems, this paper seeks to deepen our understanding of these intervals and provide related perspectives by connecting them with the broader statistical literature. Specifically, we relate these intervals to the well-developed areas of likelihood ratio tests, test inversions, and constrained inference, enabling us to make rigorous statements about their properties and generalize the methodology beyond its earlier confines. We provide a brief overview of earlier work in this area below.

Confidence intervals in penalized inverse problems Various optimization-based strategies exist for constructing confidence intervals for functionals in linear inverse problems with constraints. A first connection between optimization and inference in inverse problems is given by classical approaches seeking to optimize an objective function to balance data misfit with regularization, while adhering to prior constraints (Hansen, 1992; Hansen and O'Leary, 1993). It is then common to use the variability of the minimizer to quantify uncertainty. Another closely related strategy employs Bayesian methods to estimate the posterior distribution of model parameters given a regularizing prior and subsequently constructs credible intervals from marginal distributions (Tarantola, 2005; Andrew M. Stuart, 2010). Since these methods effectively quantify uncertainty around the expectation of the regularized estimator, their coverage is highly dependent on the precision of prior information and strength of regularization (Patil, Kuusela, and Hobbs, 2022; Kuusela, 2016; Kuusela and Victor M Panaretos, 2015a). A recent line of work starting with (Javanmard and Montanari, 2014) attempts to improve the coverage of confidence intervals derived from penalized estimators by "de-biasing" the regularized estimators; however, in practice, guarantees can only be obtained asymptotically and finite-sample performance depends on the choice of tuning parameters.

An alternative line of work in optimization-based confidence intervals focuses on ensuring correct frequentist finite-sample coverage. This approach is more resistant to the aforementioned challenges associated with relying heavily on prior assumptions or de-biasing and offers a robust framework for uncertainty quantification. In the following section, we will describe these optimization-based methods in more detail.

Optimization-based confidence intervals and the Burrus conjecture This paper is largely motivated by the literature (Burt W. Rust and Walter R. Burrus, 1972; O'Leary and Bert W. Rust, 1986; Bert W. Rust and O'Leary, 1994; Tenorio, Fleck, and Moses, 2007; Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022) surrounding the Burrus conjecture (see Section 2.4 for further discussion), which makes a claim about how to set a calibration parameter in an optimization-based confidence interval construction so that the resulting interval has a desired level of coverage for a single functional. For intervals with a simultaneous coverage guarantee for an arbitrary collection of functionals, (Philip B. Stark, 1992b) provides the most general optimization-based confidence interval construction. While these intervals provide the desired coverage, they are overly conservative in terms of length when compared to intervals calibrated for a specific functional (Stanley, Patil, and Kuusela, 2022). These prior works consider only the Gaussian linear inverse problem and can thus be seen as a particular instance of the more general optimization-based confidence intervals treated in this paper.

Inverting likelihood ratio tests and constrained inference Traditionally, optimization-based confidence interval constructions in inverse problems have developed somewhat independently of the broader statistical literature, often overlooking the duality between confidence intervals and hypothesis testing (George Casella and Roger L. Berger, 2002; Wasserman, 2004; Lehmann and Romano, 2008). Our work reinterprets these optimization-based confidence intervals from the inverse problem literature as inverted hypothesis tests and situates them within the realm of constrained testing and inference; see, e.g., (Gouriéroux, Holly, and Monfort, 1982; Wolak, 1987; Robertson, F. T. Wright, and Dykstra, 1988; Alexander Shapiro, 1988; Wolak, 1989; Molenberghs and Verbeke, 2007), among others.

The constrained inference literature often employs the $\bar{\chi}^2$ distribution, a convex combination of χ^2 distributions with different degrees of freedom, dictated by the problem constraints. Recent work in (M. Yu, Gupta, and Kolar, 2019) has extended these constrained testing frameworks to high-dimensional settings with linear inequality constraints, examining both sparse and non-sparse scenarios. Although such tests can be more powerful than their unconstrained counterparts, their definitions typically limit the null hypothesis to linear subspaces, complicating their use in test inversion scenarios (Silvapulle and Sen, 2011).

Although there have been applications of constrained test inversion (Feldman and

Cousins, 1998), these are limited in scope due to grid-based inversion approaches. The statistics literature contains other approaches to inverting likelihood ratio tests (LRTs), which center around sampling procedures (Cash, 1979; Venzon and Moolgavkar, 1988; Garthwaite and Buckland, 1992; Murphy, 1995; Neale and Miller, 1997; Schweiger et al., 2018). Alternatively, one can sample from the parameter space and the forward model to generate training data for a quantile regression, which can then be used to invert an LRT (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Masserano, Dorigo, et al., 2023; Niccolò Dalmasso, Masserano, et al., 2023; Fisher, Schweiger, and Rosset, 2020). Since these latter approaches require sampling points in the parameter space, they are practically limited to compact parameter spaces and may encounter difficulties with high-dimensional parameters. In scenarios where the data can be split, approaches such as Universal Inference (Wasserman, Ramdas, and Balakrishnan, 2020) offer a way to obtain confidence sets for irregular likelihoods with nonasymptotic coverage.

Worst-case and likelihood-free methods Most of the approaches and methods referenced and described above make strong assumptions about the underlying data-generating distribution (e.g., linear forward model and Gaussian noise). To generalize these assumptions, one can either take a worst-case approach within the model class (e.g., (Donoho, 1994) which looks at worst-case confidence intervals for linear inverse problems) or remove distribution assumptions altogether. For example, Optimal Uncertainty Quantification (OUQ) (see, e.g., (H. Owhadi, C. Scovel, T. J. Sullivan, et al., 2013a)) does not assume a particular likelihood function to perform statistical inference by relying instead on worst-case bounds. If one is willing to make boundedness assumptions on the parameter space, simulation-based inference approaches such as (Gutmann and Corander, 2015; O. Thomas et al., 2022; Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Masserano, Dorigo, et al., 2023; Niccolò Dalmasso, Masserano, et al., 2023; Cranmer, Brehmer, and Louppe, 2020; Cranmer, Pavez, and Louppe, 2016) explore the use of sampling-only access to the likelihood, typically through a simulator, which has found particular relevance in the physical sciences. While these likelihood-free methods are advantageous in contexts where the likelihood is uncertain, unknown, or accessible only through a simulator, they tend to yield conservative estimates when a well-defined likelihood is available.

2.2 Strict bounds intervals from test inversion

Suppose that we observe data $\boldsymbol{y} \in \mathbb{R}^m$ according to a data-generating process $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$. Here, $P_{\boldsymbol{x}^*}$ is a distribution that depends on a fixed but unknown parameter $\boldsymbol{x}^* \in \mathbb{R}^p$. Furthermore, suppose that we have prior knowledge that this parameter $\boldsymbol{x}^*$ lies in a constraint set $\mathcal{X} \subseteq \mathbb{R}^p$, namely $\boldsymbol{x}^* \in \mathcal{X}$. Given a nominal coverage level $1-\alpha$, where $\alpha \in (0,1)$, this paper investigates methods for constructing a $1-\alpha$ confidence interval for $\varphi(\boldsymbol{x}^*)$, where $\varphi : \mathbb{R}^p \to \mathbb{R}$ is a known one-dimensional quantity of interest[4] (we will also refer to φ as a functional of interest). More precisely, we are interested in constructing an interval $\mathcal{I}_\alpha(\boldsymbol{y}) \subseteq \mathbb{R}$ for $\varphi(\boldsymbol{x}^*)$ that satisfies the following coverage requirement:

$$\mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}^*}}\big(\varphi(\boldsymbol{x}^*) \in \mathcal{I}_\alpha(\boldsymbol{y})\big) \geq 1-\alpha, \quad \text{for all} \quad \boldsymbol{x}^* \in \mathcal{X}. \tag{2.8}$$

Our primary focus lies in intervals that (i) effectively utilize the information that $\boldsymbol{x}^* \in \mathcal{X}$, (ii) are valid (i.e., satisfying the coverage requirement in (2.8)) in the finite data and noisy regimes (rather than, e.g., in the large system or noiseless limits), (iii) do not make overly restrictive assumptions (e.g., identifiability) about the structure of the parametric model $P_{\boldsymbol{x}^*}$, and (iv) are short in length[5]. We view the observation vector $\boldsymbol{y}$ as a single observation in $\mathbb{R}^m$ drawn from a multivariate distribution $P_{\boldsymbol{x}^*}$. This may include the case of repeated sampling (i.i.d. or not) from an experiment and aggregating the samples in a vector. In this case, $P_{\boldsymbol{x}^*}$ is then defined as the measure that accounts for all the observations.[6]

Review: classical test inversion for simple null hypotheses

We briefly review the concept of test inversion and the duality between hypothesis testing and confidence sets upon which the subsequent subsections will be built. After observing $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$, two classical statistical tasks emerge: (i) determining whether $\boldsymbol{x}^* = \boldsymbol{x}$ for a particular $\boldsymbol{x} \in \mathcal{X}$ at a significance level α (hypothesis testing), and (ii) constructing a subset of $\mathcal{X}$ that contains $\boldsymbol{x}^*$ with a coverage level $1-\alpha$ (confidence set building). In hypothesis testing, for a given parameter $\boldsymbol{x}$, one can

[4] Confidence sets of several functionals of interest with guarantees can be constructed by using the proposed method with, e.g., Bonferroni correction, but studying the performance of that approach is beyond the scope of this work

[5] Note that length will, in general, depend on the unknown parameter $\boldsymbol{x}^*$. There are several notions for the "optimality" of the method with respect to length, such as minimax length (Donoho, 1994; Schafer and Philip B. Stark, 2009) or expected length (Stanley, Patil, and Kuusela, 2022), among others.

[6] For example, in the typical case where a d dimensional vector is observed a total of n times, we aggregate the results in an $m = n \times d$ dimensional vector. Throughout, we use m to denote the total dimensionality of the observation vector.

consider the hypothesis test:

$$H_0 : \boldsymbol{x}^* = \boldsymbol{x} \quad \text{versus} \quad H_1 : \boldsymbol{x}^* \neq \boldsymbol{x}. \tag{2.9}$$

We then build an acceptance region $A(\boldsymbol{x})$ in the data space (the space in which the observations $\boldsymbol{y}$ live) corresponding to the observations that would not reject H_0, with the condition that H_0 is rejected with probability at most α when it is true. In confidence set building, one builds a subset in parameter space (the space in which the parameters $\boldsymbol{x}$ live) as a function of the data, $C(\boldsymbol{y})$, such that it contains $\boldsymbol{x}^*$ with probability at least $1 - \alpha$ (over repeated samples of $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$).

Lifting to the product space of the data and parameter spaces (see Figure 2.3 for an illustration), both tasks amount to the construction of a compatibility region $\mathcal{S}$. For a fixed observation $\boldsymbol{y}$, a confidence set is given by $C(\boldsymbol{y}) = \{\boldsymbol{x} : (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{S}\}$, and for a fixed parameter $\boldsymbol{x}$, the acceptance region is given by $\mathcal{A}(\boldsymbol{x}) = \{\boldsymbol{y} : (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{S}\}$. Observe that $\mathbb{P}(\boldsymbol{y} \in \mathcal{A}(\boldsymbol{x})) = \mathbb{P}(\boldsymbol{x} \in C(\boldsymbol{y}))$. Therefore, a procedure that forms confidence sets with coverage $1 - \alpha$ for any possible data $\boldsymbol{y}$ also creates a procedure that yields valid hypothesis tests at the level α for any possible parameter value $\boldsymbol{x}$, and vice versa. This observation can be used to create confidence sets as the set of parameter values that would not be rejected by a hypothesis test, a construction known as test inversion (see, e.g., Chapter 7 of (George Casella and Roger L. Berger, 2002) or Chapter 5 of (Victor M. Panaretos, 2016)).

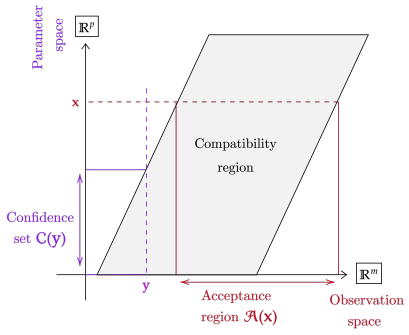


Figure 2.3: Illustration of the classical duality between hypothesis testing and confidence set building as seen in the product space of the data and parameter spaces. Pairs of the dual hypothesis test and the confidence set can be viewed as a set $\mathcal{S}$ in the product space (the compatibility region). For fixed data $\boldsymbol{y}$, a confidence set is given by $C(\boldsymbol{y}) = \{\boldsymbol{x} : (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{S}\}$, and for a fixed parameter $\boldsymbol{x}$, the acceptance region is given by $\mathcal{A}(\boldsymbol{x}) = \{\boldsymbol{y} : (\boldsymbol{y}, \boldsymbol{x}) \in \mathcal{S}\}$.

Formulation and inversion of constrained likelihood ratio tests

The starting point of this work is the inversion of specific hypothesis tests that can incorporate the constraint information $\mathcal{X}$ and the functional of interest φ. We will establish that the test inversion can be achieved by solving two endpoint optimization problems. We note that unlike the simple null versus composite alternative tests (2.9) described in Section 2.2, the tests we will consider have composite nulls. We focus on the continuous case and assume that the Lebesgue measure dominates the set of distributions $\mathcal{P} := \{P_{\boldsymbol{x}} \mid \boldsymbol{x} \in \mathcal{X}\}$. However, a discrete analog can also be constructed using a similar approach as in (Feldman and Cousins, 1998). Let $L_{\boldsymbol{x}}$ be the density of $P_{\boldsymbol{x}}$, and let $\ell_{\boldsymbol{x}} := \log L_{\boldsymbol{x}}$. For any $\mu \in \mathbb{R}$, denote the level sets of the quantity of interest φ by Φ_μ. These are defined as follows:

$$\Phi_\mu := \{\boldsymbol{x} \colon \varphi(\boldsymbol{x}) = \mu\} \subseteq \mathbb{R}^p. \tag{2.10}$$

Subsequently, define a hypothesis test T_μ as follows:

$$H_0 \colon \boldsymbol{x}^* \in \Phi_\mu \cap \mathcal{X} \quad \text{versus} \quad H_1 \colon \boldsymbol{x}^* \in \mathcal{X} \setminus \Phi_\mu. \tag{2.11}$$

We can test hypothesis (2.11) (for a fixed μ) with a Likelihood Ratio (LR) test statistic defined as the following function of the observed data $\boldsymbol{y}$:

$$\Lambda(\mu, \boldsymbol{y}) := \frac{\sup\limits_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} L_{\boldsymbol{x}}(\boldsymbol{y})}{\sup\limits_{\boldsymbol{x} \in \mathcal{X}} L_{\boldsymbol{x}}(\boldsymbol{y})}. \tag{2.12}$$

The corresponding log-likelihood ratio (LLR) statistic $\lambda(\mu, \boldsymbol{y})$ is given by

$$\begin{aligned} \lambda(\mu, \boldsymbol{y}) = -2 \log \Lambda(\mu, \boldsymbol{y}) &= -2 \left\{ \sup_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} \ell_{\boldsymbol{x}}(\boldsymbol{y}) - \sup_{\boldsymbol{x} \in \mathcal{X}} \ell_{\boldsymbol{x}}(\boldsymbol{y}) \right\} \\ &= \inf_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x} \in \mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}). \end{aligned} \tag{2.13}$$

As is standard (see, e.g., (George Casella and Roger L. Berger, 2002; Wasserman, 2004)), we use the supremum over all $\mathcal{X}$ in the denominator of (2.12), instead of over $\mathcal{X} \setminus \Phi_\mu$[7]. The factor of -2 helps connect with the standard likelihood ratio test in the context of Wilks' theorem and is needed, together with the optimization being over the whole space, to reinterpret the previous constrained inference intervals as coming from the inversion of this test (see Section 2.4).

[7](Schervish, 1995) provides conditions for the equality of both test statistics.

Motivation behind the choice of test and test statistic In addition to the reinterpretation of the previous constrained inference intervals as a result of inverting this test, there are other theoretical and practical reasons that make it a reasonable choice for this work. Theoretically, the LR emerges as the optimal test statistic (resulting in the most powerful level-α test) in the simple versus simple hypothesis testing setting via the Neyman–Pearson Lemma (George Casella and Roger L. Berger, 2002; Lehmann and Romano, 2008). Although uniformly most powerful tests do not exist in general, LR tests have been effective in several contexts. For example, (Abraham Wald, 1943) provides some optimality properties for the likelihood ratio test in terms of its asymptotic average power. Although our test of interest does not fall under the simple versus simple paradigm and we are interested in nonasymptotic properties, these two properties support the sensibility of adopting the LR-based test. Furthermore, the literature on constrained inference (Robertson, F. T. Wright, and Dykstra, 1988; Silvapulle and Sen, 2011) extensively uses the LR, deriving both asymptotic and nonasymptotic log-likelihood ratio (LLR) distributions in various scenarios, often leading to the $\bar{\chi}^2$ distribution. These characterizations indicate that, in very specific situations, it is possible to obtain the distribution of the test statistic under the null hypothesis, either exactly or in an asymptotic sense. Our setting extends well beyond those situations because our setup is general, and we do not make particular assumptions on the likelihood model or the constraint set.

The distribution of the LLR and test inversion In hypothesis testing, we reject the null hypothesis when the values of $\lambda(\mu, \boldsymbol{y})$ exceed a threshold. This indicates that there is substantial evidence against the data being generated by a distribution in the composite null defined by μ. To choose a rejection region, we next study the distribution of the LLR, denoted as $\lambda(\mu, \boldsymbol{y})$, in the context where $\mu = \varphi(\boldsymbol{x})$ (pertaining to the null hypothesis) and $\boldsymbol{y} \sim P_{\boldsymbol{x}}$, a data sampling model, across various values of $\boldsymbol{x} \in \mathcal{X}$. Let $F_{\boldsymbol{x}}$ denote the distribution of $\lambda(\varphi(\boldsymbol{x}), \boldsymbol{y})$ for any $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{y} \sim P_{\boldsymbol{x}}$. To simplify the notation, we will write $\lambda \sim F_{\boldsymbol{x}}$ to indicate that an LLR is sampled following the procedure described above.

To ensure an α-level test for test inversion, we need to control the distribution of the test statistic under the null hypothesis. Since the null is composite, the false positive rate must hold for any parameter under the null hypothesis H_0.

Suppose that we are conducting a test T_μ to determine whether $\mu^* = \varphi(\boldsymbol{x}^*)$ equals a particular $\mu \in \varphi(\mathcal{X}) \subseteq \mathbb{R}$, that is, $\boldsymbol{x}^* \in \Phi_\mu \cap \mathcal{X}$. We use $\lambda > q_\alpha$ as the rejection

region, where q_α is a predetermined decision threshold. Under the null hypothesis, if the decision threshold satisfies

$$\sup_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} \mathbb{P}_{\lambda\sim F_{\boldsymbol{x}}}\left(\lambda > q_\alpha\right) \leq \alpha \tag{2.14}$$

for all $\alpha \in (0,1)$, then we say T_μ is a *level-α test*.[8]

Inverting the test with respect to μ will require choosing an appropriate q_α for all μ; henceforth we will denote it as $q_\alpha(\mu)$.

We seek to invert this test using a methodology similar to that outlined in Section 2.2, but adapted to accommodate the composite null hypothesis. The acceptance region is formally defined as:

$$\mathcal{A}_\alpha(\mu) := \{\boldsymbol{y} : \lambda(\mu, \boldsymbol{y}) \leq q_\alpha(\mu)\}\,. \tag{2.15}$$

Subsequently, we define the proposed confidence set for $\mu^* = \varphi(\boldsymbol{x}^*) \in \mathbb{R}$ through test inversion as follows:

$$C_\alpha(\boldsymbol{y}) := \{\mu \colon \lambda(\mu, \boldsymbol{y}) \leq q_\alpha(\mu)\}. \tag{2.16}$$

We prove in Lemma 2.2.1 that if (2.14) is satisfied for $\mu^* := \varphi(\boldsymbol{x}^*)$ (that is, T_{μ^*} is a level-α test), the resulting confidence set will have the desired $1-\alpha$ coverage, thus extending the classical test inversion framework to our specific case.

Lemma 2.2.1 (Coverage of the inverted test)**.** *Let $\alpha \in (0,1)$. Let $\boldsymbol{x}^*$ be the true parameter value and μ^* its image under φ. If T_{μ^*} is a level-α test, then*

$$\mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}^*}}\left(\mu^* \in C_\alpha(\boldsymbol{y})\right) \geq 1-\alpha.$$

Proof sketch. The proof is based on a straightforward test inversion argument. For a detailed proof, see Section 2.7. □

To ensure that condition (2.14) holds in practice, when $\boldsymbol{x}^*$ and therefore μ^* are both unknown, we need to satisfy this condition for all possible null hypotheses.

[8]Here q_α is the decision value corresponding to intervals with a coverage probability of 1-α, aligning with classical textbook notation (see, e.g., (George Casella and Roger L. Berger, 2002), (Wasserman, 2004)). For any random variable Z, we will denote with the subscript α the cutoff points that satisfy $\mathbb{P}(Z > z_\alpha) = \alpha$.

Specifically, we choose an appropriate $q_\alpha(\mu)$ for each μ to ensure that all hypothesis tests T_μ are level-α. Formally, this is expressed as:

$$\sup_{\mu\in\varphi(\mathcal{X})}\sup_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}}\mathbb{P}_{\lambda\sim F_{\boldsymbol{x}}}\left(\lambda > q_\alpha(\mu)\right)\leq\alpha. \tag{2.17}$$

This condition is equivalent[9] to

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\mathbb{P}_{\lambda\sim F_{\boldsymbol{x}}}\left(\lambda > q_\alpha(\varphi(\boldsymbol{x}))\right)\leq\alpha. \tag{2.18}$$

Although (2.18) lacks the interpretation of (2.17) of having hypothesis tests for each different $\mu\in\varphi(\mathcal{X})$, it simplifies the calculations. We refer to a set of values $q_\alpha(\mu)$ that satisfy (2.18) (or equivalently (2.17)) as *valid values*. Since μ in (2.17) is equal to $\varphi(\boldsymbol{x})$ as $\boldsymbol{x}\in\Phi_\mu$, we can use $q_\alpha(\mu)$ and $q_\alpha(\varphi(\boldsymbol{x}))$ interchangeably.

From Lemma 2.2.1, we know that valid values can be used in (2.16) to construct a confidence set for μ^* with the correct $1-\alpha$ coverage. Moreover, as argued in the proof of Lemma 2.2.1, the probability that the set (2.16) covers the unknown μ^* is given by

$$\mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}^*}}\left(\mu^*\in C_\alpha(\boldsymbol{y})\right) = 1-\mathbb{P}_{\lambda\sim F_{\boldsymbol{x}^*}}\left(\lambda > q_\alpha(\mu^*)\right), \tag{2.19}$$

which is guaranteed to be at least $1-\alpha$ by the condition (2.17). To obtain intervals with the smallest possible size while maintaining coverage, we aim to find the optimal decision values $q_\alpha(\mu)$, which are solutions to optimization problems involving the quantiles $Q_{F_{\boldsymbol{x}}}:[0,1]\to\mathbb{R}$ of the distributions of the family $\{F_{\boldsymbol{x}},\boldsymbol{x}\in\mathcal{X}\}$.

Lemma 2.2.2 (Optimal decision values)**.** *The optimal (smallest valid) value of $q_\alpha(\mu)$ is given by the maximum quantile (MQ) optimization problem:*

$$Q^{\max}_{\mu,1-\alpha} := \sup_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha). \tag{2.20}$$

Furthermore, if one wants to choose a single q_α for all μ, then the optimal value is given by

$$Q^{\max}_{1-\alpha} := \sup_{\mu\in\varphi(\mathcal{X})} Q^{\max}_{\mu,1-\alpha} = \sup_{\boldsymbol{x}\in\mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha). \tag{2.21}$$

Proof sketch. It can be directly checked that the proposed quantities are valid and that using any smaller decision value leads to intervals with undercoverage for at least one point. See Section 2.7 for more details. □

[9] Note that every $\boldsymbol{x}\in\mathcal{X}$ is accounted for in (2.17) since $\mu=\varphi(\boldsymbol{x})$.

In the event that only an upper bound on the quantities defined in Lemma 2.2.2 can be obtained, those can also be used as valid decision values, as stated precisely in the following corollary.

Corollary 2.2.3. *For every $\mu \in \mathbb{R}$, $\nu(\mu) \geq Q^{\max}_{\mu,1-\alpha}$ (as defined in Lemma 2.2.2) if and only if $\nu(\mu)$ is a valid decision value for T_μ for a particular α. In addition, $\nu \geq Q^{\max}_{1-\alpha}$ if and only if ν is a valid decision value for T_μ for all $\mu \in \mathbb{R}$ for a particular α.*

The result follows immediately from substituting the proposed ν values into the probability statements in Lemma 2.2.2. Theoretical and computational methods for obtaining valid $q_\alpha(\mu)$ are discussed in Section 2.3, and we investigate the computation of $C_\alpha(\boldsymbol{y})$ via optimization techniques in the next subsection, assuming valid $q_\alpha(\mu)$ are known.

Characterizing the inverted confidence set via optimization problems

The set defined in (2.16) produces a random collection of real numbers that contains the true functional value with a probability of at least $1-\alpha$. Although this set is not necessarily an interval, it is contained within an interval whose (possibly infinite) extremes are computable through optimization techniques.

Given a valid $q_\alpha(\mu)$, which satisfies either (2.17) or (2.18), let us define the following sets:

$$\mathcal{D}(\boldsymbol{y}) := \{\boldsymbol{x} : \ -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\varphi(\boldsymbol{x})) + \inf_{\boldsymbol{x}' \in \mathcal{X}} -2\ell_{\boldsymbol{x}'}(\boldsymbol{y})\} \subseteq \mathbb{R}^p, \tag{2.22}$$

$$\bar{\mathcal{X}}_\alpha(\boldsymbol{y}) := \mathcal{X} \cap \mathcal{D}(\boldsymbol{y}). \tag{2.23}$$

If $\bar{\mathcal{X}}_\alpha(\boldsymbol{y}) \neq \emptyset$, we further define:

$$\mathcal{I}_\alpha(\boldsymbol{y}) := \left[\inf_{\boldsymbol{x} \in \bar{\mathcal{X}}_\alpha(\boldsymbol{y})} \varphi(\boldsymbol{x}), \ \sup_{\boldsymbol{x} \in \bar{\mathcal{X}}_\alpha(\boldsymbol{y})} \varphi(\boldsymbol{x}) \right]. \tag{2.24}$$

If $\bar{\mathcal{X}}_\alpha(\boldsymbol{y}) = \emptyset$, let $\mathcal{I}_\alpha(\boldsymbol{y})$ be the empty interval.

Theorem 2.2.4 (From test inversion to optimization-based intervals)**.** *For any $\alpha \in (0,1)$, and for any $\boldsymbol{x} \in \mathcal{X}$, let $\mathcal{I}_\alpha(\boldsymbol{y})$ be the interval constructed according to* (2.24)*. It holds that*

$$\mathbb{P}_{\boldsymbol{y} \sim P_{\boldsymbol{x}}} \left(\varphi(\boldsymbol{x}) \in \mathcal{I}_\alpha(\boldsymbol{y}) \right) \geq 1 - \alpha.$$

In other words, $\mathcal{I}_\alpha(\boldsymbol{y})$ is a valid $1-\alpha$ confidence interval for $\varphi(\boldsymbol{x}^*)$.

Proof sketch. We prove that

$$\left[\inf_{\lambda(\mu,\boldsymbol{y})\le q_\alpha(\mu)} \mu,\ \sup_{\lambda(\mu,\boldsymbol{y})\le q_\alpha(\mu)} \mu\right] = \mathcal{I}_\alpha(\boldsymbol{y}). \tag{2.25}$$

The object on the left-hand side is defined by enclosing $C_\alpha(\boldsymbol{y})$ within the smallest possible interval that contains it, and therefore, it has guaranteed coverage. The equality arises from the equivalence between the optimization problems under consideration. For a complete proof, see Section 2.7. □

Remark 1 (Comparison with the simultaneous strict bound intervals)**.** Observe that the construction of $\mathcal{I}_\alpha(\boldsymbol{y})$ follows the form outlined in (2.3) for the simultaneous strict bound intervals. However, a key distinction lies in not requiring that $\mathcal{D}(\boldsymbol{y}) \subseteq \mathcal{X}$ serves as a $1-\alpha$ confidence set for $\boldsymbol{x}$. This relaxation will translate into shorter intervals when q_α is chosen appropriately.

Remark 2 (Handling empty constrained sets)**.** If α is chosen such that $1-\alpha$ becomes too small, the set $\bar{\mathcal{X}}_\alpha(y)$ can be empty. In that case, we default to the empty interval, under the interpretation that there are no parameter values that simultaneously agree with the constraint and the observed data (at a particular level α). However, the actual interval produced under this circumstance does not compromise the $1-\alpha$ coverage level provided by the theorem. If a point estimate inside the constraint region is desired, an option is to choose the closest point from $\mathcal{X}$ to $\mathcal{D}$. This point specifically ensures the continuity of the interval with respect to α in many standard scenarios. Generally, an empty set $\bar{\mathcal{X}}_\alpha(\boldsymbol{y})$ should inform one of three possibilities: either (i) an outlier event has been observed, or (ii) the initial assumption that $\boldsymbol{x} \in \mathcal{X}$ is flawed, or (iii) the forward model $P_{\boldsymbol{x}}$ is misspecified. Here, the definition of an "outlier" is intrinsically linked to the choice of α. A larger α will make such events more frequent, as it broadens the range of data considered as outliers.

We also present a partial converse result, stating that the interval coverage implies the validity of q_α, subject to appropriate assumptions on φ, P, and $\mathcal{X}$. This result will be instrumental in refuting the coverage claims of the Rust–Burrus intervals, and consequently, the Burrus conjecture, as discussed in Section 2.4.

Proposition 2.2.5 (Coverage implies validity of quantile levels)**.** *Assume that $\mathcal{X}$ forms a convex cone, $\ell_{\boldsymbol{x}}(\boldsymbol{y})$ is a concave function, and $\varphi(\boldsymbol{x})$ is linear. Define $\mathcal{I}_\alpha(\boldsymbol{y})$ as in Theorem 2.2.4, for a particular choice of $q_\alpha(\mu)$. If $\mathcal{I}_\alpha(\boldsymbol{y})$ is a valid $1-\alpha$ confidence interval for all $\boldsymbol{x}$, then the values of $q_\alpha(\mu)$ are valid.*

Proof sketch. Generally, the values of $q_\alpha(\mu)$ are valid if and only if $C_\alpha(\boldsymbol{y})$ constitutes a $1-\alpha$ set. Since $\mathcal{I}_\alpha(\boldsymbol{y})$ is the smallest interval that contains $C_\alpha(\boldsymbol{y})$, if $C_\alpha(\boldsymbol{y})$ is already an interval, then the result holds. The assumptions on $\mathcal{X}, \ell_{\boldsymbol{x}}(\boldsymbol{y})$, and φ ensure that this is the case by the convexity of the function

$$\mu \mapsto \inf_{\substack{\varphi(\boldsymbol{x})=\mu \\ \boldsymbol{x}\in\mathcal{X}}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y})$$

for any $\boldsymbol{y}$. For a detailed proof, see Section 2.7. □

Finally, we remark that the construction presented in this paper provides an approach to uncertainty quantification that does not rely on a specific point estimator, distinguishing it from many other UQ procedures. However, if one wishes to obtain a point estimator, it is worth noting that the midpoint of the interval can be justified from a decision-theoretic perspective. This idea has been discussed in previous works by (Micchelli and Rivlin, 1977; Bajgiran et al., 2022b), among others.

Illustrative examples

To elucidate the general methodology outlined in Theorem 2.2.4, we offer two simple illustrative examples where the LLR and its distribution are explicitly computable: a one-dimensional constrained Gaussian scenario and an unconstrained linear Gaussian case.

Constrained Gaussian in one dimension As a tangible example, consider the following one-dimensional model:

$$\underbrace{y = x^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)}_{\text{model}} \quad \text{with} \quad \underbrace{x^* \geq 0}_{\text{constraints}} \quad \text{and} \quad \underbrace{\varphi(x^*) = x^*}_{\text{functional}}. \tag{2.26}$$

In this case, the distribution of the LLR is precisely known. Hence, a confidence interval can be constructed without resorting to the techniques introduced in Section 2.3, which are otherwise necessary when such information is not available.

The form of the hypothesis test T_μ, as given in (2.11), is as follows:

$$H_0 : x^* = \mu \quad \text{versus} \quad H_1 : x^* \neq \mu \text{ and } x^* \geq 0. \tag{2.27}$$

The LLR as defined in (2.13) for the test (2.27) is given by

$$\begin{aligned} \lambda(\mu, y) &= \inf_{x=\mu, x\geq 0} (y-x)^2 - \inf_{x\geq 0} (y-x)^2 \\ &= \begin{cases} (y-\mu)^2, & y \geq 0, \\ (y-\mu)^2 - y^2, & y < 0. \end{cases} \end{aligned} \tag{2.28}$$

We can also derive its distribution under the null hypothesis (i.e., when $x^* = \mu$, leading to $y = \mu + \varepsilon$) for any $\mu \in [0, \infty)$, as formalized below.

Example 2.2.6 (Distribution of the LLR statistic for a constrained Gaussian in one dimension)**.** *For $\lambda(\mu, y)$ as defined in* (2.28) *with $\mu \geq 0$, when $y \sim \mathcal{N}(\mu, 1)$ (null hypothesis), for all $c > 0$, we have*

$$\mathbb{P}(\lambda(\mu, y) \leq c) = \begin{cases} \chi_1^2(c) + \frac{1}{2}, & \mu = 0 \\ \chi_1^2(c) \cdot \mathbf{1}\{c < \mu^2\} + \{\Phi(\sqrt{c}) - \Phi((-\mu^2 - c)/(2\mu))\} \cdot \mathbf{1}\{c \geq \mu^2\}, & \mu > 0, \end{cases}$$

where χ_1^2 and Φ are the CDFs of a χ_1^2 and a standard Gaussian, respectively.

Proof. See Section 2.7. □

The expression for $\lambda(\mu, y)$, with the appropriately scaled log transformation, is equivalent to Equation (4.3) in (Feldman and Cousins, 1998) where the Neyman confidence interval construction for the same problem is considered. (Feldman and Cousins, 1998) characterizes this quantity as a likelihood ordering for determining an acceptance region.

By virtue of the previous result and Lemma 2.2.2, we can take $Q_\mu(1 - \alpha)$, where Q_μ is the quantile of the distribution of $\lambda(\mu, y)$ when μ is fixed, as $q_\alpha(\mu)$ satisfying (2.14). A direct computation shows

$$q_\alpha(\mu) = Q_\mu(1 - \alpha) = \begin{cases} Q_{\chi_1^2}(1 - \alpha), & 1 - \alpha < \chi_1^2(\mu^2), \\ r_{\mu,\alpha}, & 1 - \alpha \geq \chi_1^2(\mu^2), \end{cases} \tag{2.29}$$

where $r_{\mu,\alpha}$ is the unique nonnegative root of the function $x \mapsto \Phi(\sqrt{x}) - \Phi((-\mu^2 - x)/(2\mu)) - (1-\alpha)$, which can be found using numerical methods. Therefore, $\mathcal{D} = \{x : (y-x)^2 \leq q_\alpha(x) + \min_{x' \geq 0}(y - x')^2\}$ and the final form of the confidence interval becomes

$$\mathcal{I}_\alpha(y) = \left[\min_{\substack{x \in \mathcal{D} \\ x \geq 0}} x, \ \max_{\substack{x \in \mathcal{D} \\ x \geq 0}} x \right].$$

For a numerical comparison of this interval with alternative methods, we refer the reader to Section 2.5.

Unconstrained Gaussian linear model Consider the following problem setup:

$$\underbrace{\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)}_{\text{model}} \quad \text{and} \quad \underbrace{\varphi(\boldsymbol{x}^*) = \boldsymbol{h}^\top \boldsymbol{x}^*}_{\text{functional}}. \tag{2.30}$$

Assume $\boldsymbol{K} \in \mathbb{R}^{m\times p}$ has full column rank. The assumption $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{I}_m$ is without loss of generality as it is equivalent to assuming a known positive definite covariance for $\boldsymbol{y}$ and performing a basis change with the Cholesky factor. Note that this setup is the same as (2.4) but the parameter space is not constrained, that is, $\mathcal{X} = \mathbb{R}^p$, and the forward model $\boldsymbol{K}$ is assumed to be full rank.

Using the framework established in Section 2.2, our aim is to invert the following family of hypothesis tests:

$$H_0 : \boldsymbol{h}^\top \boldsymbol{x}^* = \mu \quad \text{versus} \quad H_1 : \boldsymbol{h}^\top \boldsymbol{x}^* \neq \mu. \tag{2.31}$$

The LLR as defined in (2.13) for the test (2.31) takes the form

$$\lambda(\mu, \boldsymbol{y}) \;:=\; \min_{\boldsymbol{x}\,:\,\boldsymbol{h}^\top \boldsymbol{x} = \mu} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2. \tag{2.32}$$

In this particular scenario, the LLR admits a closed-form expression and has a straightforward distribution, as formalized below:

Example 2.2.7 (Distribution of the LLR statistic for the unconstrained Gaussian linear model)**.** *$\lambda(\mu, \boldsymbol{y})$ for the unconstrained full column rank Gaussian linear model* (2.32) *can be expressed in closed form as*

$$\lambda(\mu, \boldsymbol{y}) = \frac{(\boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1} \boldsymbol{K}^\top \boldsymbol{y} - \mu)^2}{\boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1} \boldsymbol{h}}. \tag{2.33}$$

Furthermore, for any $\boldsymbol{x}^*$, whenever $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}^*, \boldsymbol{I}_m)$, $\lambda(\boldsymbol{h}^\top \boldsymbol{x}^*, \boldsymbol{y})$ is distributed as a chi-squared distribution with 1 *degree of freedom.*

Proof. See Section 2.7. □

Using the above results, we can set $q_\alpha(\mu) = Q_{\chi_1^2}(1-\alpha)$ for all values of μ. Here, $Q_{\chi_1^2}$ represents the quantile function of a chi-squared distribution with 1 degree of freedom. Consequently, we can express the interval in (2.24) as:

$$\mathcal{I}_\alpha(\boldsymbol{y}) = \left[\min_{\boldsymbol{x}\in\mathcal{D}(\boldsymbol{y})} \boldsymbol{h}^\top \boldsymbol{x},\; \max_{\boldsymbol{x}\in\mathcal{D}(\boldsymbol{y})} \boldsymbol{h}^\top \boldsymbol{x} \right], \tag{2.34}$$

where we define $\mathcal{D}(\boldsymbol{y}) := \{\boldsymbol{x} : \|\boldsymbol{y}-\boldsymbol{K}\boldsymbol{x}\|_2^2 \leq Q_{\chi_1^2}(1-\alpha) + \min_{\boldsymbol{x}'} \|\boldsymbol{y}-\boldsymbol{K}\boldsymbol{x}'\|_2^2\}$. Similarly, let us define $z_\alpha = \Phi^{-1}(1-\alpha)$, where Φ is the cumulative distribution function of the standard normal distribution. Using the equivalence $z_{\alpha/2}^2 = Q_{\chi_1^2}(1-\alpha)$, we can rewrite the expression in terms of the standard normal. Moreover, as shown

in Appendix A of (Patil, Kuusela, and Hobbs, 2022), the endpoints of the above interval can be calculated in closed form and are given by

$$\mathcal{I}_\alpha(\boldsymbol{y}) = \left[\boldsymbol{h}^\top \widehat{\boldsymbol{x}} - z_{\alpha/2}\sqrt{\boldsymbol{h}^\top \left(\boldsymbol{K}^\top \boldsymbol{K}\right)^{-1} \boldsymbol{h}},\ \boldsymbol{h}^\top \widehat{\boldsymbol{x}} + z_{\alpha/2}\sqrt{\boldsymbol{h}^\top \left(\boldsymbol{K}^\top \boldsymbol{K}\right)^{-1} \boldsymbol{h}}\right], \tag{2.35}$$

where we define the least-squares estimator $\widehat{\boldsymbol{x}} = (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{K}^\top \boldsymbol{y}$. This interval is equivalent to the one derived from observing that $\widehat{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{x}^*, (\boldsymbol{K}^\top \boldsymbol{K})^{-1})$. Therefore, we have $\boldsymbol{h}^\top \widehat{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{h}^\top \boldsymbol{x}^*, \boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1} \boldsymbol{h})$. The interval in (2.35) is thus a standard construction of a Gaussian $1 - \alpha$ confidence interval. Our construction therefore coincides with the classical interval in this case where a guaranteed-coverage interval can be obtained with standard manipulations; however, our framework remains valid in constrained, rank-deficient, non-Gaussian and/or nonlinear problems where few alternative approaches are available.

2.3 General interval construction methodology

In this section, we outline the core practical methodology for constructing intervals derived from Theorem 2.2.4. To summarize the preceding, Lemma 2.2.1 asserts that if we know $q_\alpha(\mu)$ satisfying (2.18), we can invert the hypothesis test with a composite null hypothesis defined in (2.11) to yield a valid $1-\alpha$ confidence interval. Lemma 2.2.2 poses two optimization problems that, if solved, yield valid decision values. The optimization problems in Lemma 2.2.2 give rise to two approaches of increasing complexity: (i) finding a single q_α that is valid for any μ and (ii) finding valid $q_\alpha(\mu)$ dependent on $\mu \in \varphi(\mathcal{X})$. While approach (ii) can lead to tighter intervals, it is usually at the cost of more complex theoretical analysis and computations, including the computational complexity of solving the optimization problems in (2.24). In particular, when we reinterpret previously proposed optimization-based methods in Section 2.4, we will observe that these previously proposed methods are of type (i), which this work expands to accommodate type (ii) generalizations. In this section, we briefly analyze the hardness of the optimization problems (2.20) and (2.21) by connecting them to the chance-constrained optimization literature. In case solving these problems is impractical, in Section 2.3, we describe using stochastic dominance as a theoretical tool that can be used to create and analyze provable upper bounds to the optimization problems. Stochastic dominance will also be used in Section 2.4 as the main technique to disprove coverage of the previously proposed Rust–Burrus intervals. This section is summarized in a meta-algorithm, detailed in Section 2.3.

Maximum quantile problems as chance-constrained optimization

Lemma 2.2.2 presents optimization problems for finding the maximum quantiles, which are crucial for the proposed hypothesis test inversion procedure. We show that these problems, of the form $\sup_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha)$ and $\sup_{\boldsymbol{x}\in\mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha)$, can be formulated as chance-constrained optimization (CCO) problems (see (Geng and Xie, 2019a) for a review of theory and applications of CCO).

Lemma 2.3.1 (Chance constrained characterization of the max quantile problems)**.** *Let $\mathcal{S}\subseteq\mathcal{X}$. Then the max quantile optimization problem* $\sup_{\boldsymbol{x}\in\mathcal{S}} Q_{F_{\boldsymbol{x}}}(1-\alpha)$ *can be equivalently written as the chance constrained optimization problem:*

$$\begin{aligned} \sup_{q,\boldsymbol{x}} \quad & q \\ \text{st} \quad & \boldsymbol{x}\in\mathcal{S} \\ & q\in\mathbb{R} \\ & \mathbb{P}_{u\sim\mathcal{U}([0,1])}(\mathcal{F}(\boldsymbol{x},u)\le q)\le 1-\alpha, \end{aligned} \tag{2.36}$$

where $\mathcal{F}(\boldsymbol{x},u)=F_{\boldsymbol{x}}^{-1}(u)$, with $F_{\boldsymbol{x}}^{-1}$ being the (possibly generalized) inverse CDF of $F_{\boldsymbol{x}}$

Proof. By using the definition of $(1-\alpha)$-quantile of X as the maximum q such that $\mathbb{P}(X\le q)\le 1-\alpha$, and $\lambda\sim F_{\boldsymbol{x}} \overset{d}{=} \mathcal{F}(\boldsymbol{x},u\sim\mathcal{U}([0,1])$ □

Note that Lemma 2.3.1 applies to both (2.20) and (2.21) by choosing appropriate $\mathcal{S}$. In general, CCO problems are known to be strongly NP-hard (Geng and Xie, 2019a), even with convexity assumptions for $\mathcal{F}$. Although various algorithms exist for general chance-constrained optimization, we leave the development of an algorithm specific to this problem and comparison with the aforementioned algorithms for future work. In our numerical examples (see Section 2.5), we solve these problems using gradient-free optimizers that do not exploit the chance-constrained structure but instead see the quantile function as a noisy black-box function to be optimized, with evaluations performed by estimating quantiles from large amount of samples. In higher dimensional scenarios, more advanced techniques tailored to the chance-constrained structure might be required. One can also write optimization problem (2.36) and the interval optimization problem (2.24) jointly as one chance-constrained optimization problem; see Section 2.8.

Analytical ways to obtain quantile levels via stochastic dominance

In this subsection, we develop an analytical tool to find valid q_α that allows for a straightforward evaluation for any confidence level $1-\alpha \in (0,1)$. We first consider the case where we aim to choose a valid q_α for all μ. We propose taking $q_\alpha = Q_X(1-\alpha)$, where Q_X is the quantile function of a random variable X with a known, easy-to-compute distribution. We establish that for the resulting confidence interval to maintain a $1-\alpha$ coverage guarantee for any α, X must stochastically dominate the random variable with distribution $F_{\boldsymbol{x}^*}$, i.e., $\lambda(\mu^*, \boldsymbol{y})$ where $\boldsymbol{y}$ is a random variable with distribution $P_{\boldsymbol{x}^*}$. This is denoted as $X \succeq \lambda(\mu^*, \boldsymbol{y})$ or, with slight abuse of notation, as $X \succeq F_{\boldsymbol{x}^*}$. Following the classical definition of stochastic dominance for real-valued random variables (see, e.g., (Shaked and Shanthikumar, 2007)), we say that $X \succeq Y$ if and only if $\mathbb{P}(X \geq z) \geq \mathbb{P}(Y \geq z)$ [10] for all $z \in \mathbb{R}$

Lemma 2.3.2 (Valid quantile level via stochastic dominance)**.** *$Q_X(1-\alpha)$ serves as a valid (in the sense of* (2.18)*) choice for q_α for all α if and only if $X \succeq \lambda(\mu^*, \boldsymbol{y})$, where $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$.*

Proof. See Section 2.8. □

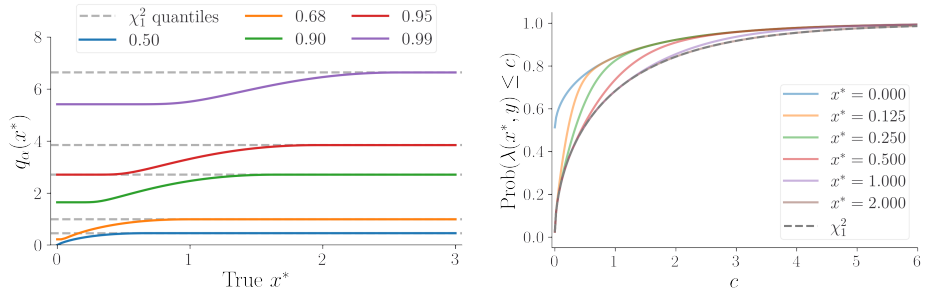
Remark 3 (Partial validity of quantile levels)**.** If X does not stochastically dominate $\lambda(\mu^*, \boldsymbol{y})$, a valid q_α can still be identified for specific α levels, provided that certain conditions are met. Specifically, z can serve as a valid q_α where $\alpha = 1 - F_X(z)$ and F_X being the cumulative distribution function of X, if and only if $\mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z)$ for some value of z.

Remark 4 (Support restriction)**.** Candidates for X can be restricted to the range $[0, \infty)$ without loss of generality, as $\lambda(\mu^*, \boldsymbol{y})$ is supported on this range by moving the mass a candidate X might have in $(-\infty, 0)$ to 0.

An economic interpretation of our result is that agents with nondecreasing utility functions would prefer a reward drawn from X over one from $\lambda(\mu^*, \boldsymbol{y})$. In practical scenarios where the true parameter $\boldsymbol{x}^*$ is unknown, it is required to establish stochastic dominance for the entire family of distributions $F_{\boldsymbol{x}}$, where $\boldsymbol{x} \in \mathcal{X}$.

Although all stochastically dominant distributions provide correct coverage when used to obtain q_α, a larger stochastic dominance gap provides more conservative

[10] One can equivalently define stochastic dominance with strict inequalities $X > z$ and $Y > z$; see (Shaked and Shanthikumar, 2007)



(a) (b)

Figure 2.4: Comparison of quantile functions and CDFs for LLRs with different true parameter values. The **left** panel provides the true values of the quantile function as a function of x^* across different confidence levels. As proven in Example 2.3.3, the quantile of χ^2_1 is greater than the true quantile for all x^* and all levels. The **right** panel shows the CDFs for LLRs under different values of the true parameters, x^*. From Example 2.3.3, as the true parameter increases, the CDF is increasingly dominated by its χ^2_1 component, so it follows that as x^* increases, the CDF approaches the χ^2_1 CDF. This figure also provides a visual explanation of why using the true quantile or the true quantile function to compute the interval in (2.24) produces shorter intervals compared to those computed with the χ^2_1 quantile.

bounds. Furthermore, if X_1, X_2 both stochastically dominate the family $F_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$, we can take the pointwise minimum $q_\alpha = \min\{Q_{X_1}(1-\alpha), Q_{X_2}(1-\alpha)\}$ which will be no worse than using either X_1 or X_2.

The perspective of stochastic dominance also enables the use of coupling arguments to identify stochastically dominating distributions. For instance, one approach to find stochastically dominating distributions to a given $F_{\boldsymbol{x}}$ is finding a function $g(\varphi(\boldsymbol{x}), \boldsymbol{y})$ such that for all z,

$$\mathbb{P}(g(\varphi(\boldsymbol{x}), \boldsymbol{y}) \geq z) \geq \mathbb{P}(\lambda(\varphi(\boldsymbol{x}), \boldsymbol{y}) \geq z),$$

where the randomness is from $\boldsymbol{y} \sim P_{\boldsymbol{x}}$. A particular case is that of nonrandom bounds. If $g(\varphi(\boldsymbol{x}), \boldsymbol{y}) \geq \lambda(\varphi(\boldsymbol{x}), \boldsymbol{y})$ almost surely in $\boldsymbol{y}$ (as opposed to when $\boldsymbol{y} \sim P_{\boldsymbol{x}}$), then this implies a coupling of random variables once $\boldsymbol{y}$ is sampled that implies stochastic dominance (see e.g. Theorem 4.2.3 in (Roch, 2024)).

This technique can be generalized to find $q_\alpha(\mu)$. Instead of finding a stochastic dominant variable X such that $X \succeq F_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$, we aim to find a distribution X_μ for each μ, such that $X_\mu \succeq F_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}$, and then set $q_\alpha(\mu) = Q_{X_\mu}(1-\alpha)$. This ensures that $Q_{X_{\mu^*}} \succeq F_{\boldsymbol{x}^*}$, providing the desired coverage guarantees.

As an illustration, we revisit the one-dimensional constrained example discussed in Section 2.2. We consider the model $y = x^* + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$, $x^* \geq 0$, and $\varphi(x) = x$. We recall that we have $\lambda(\mu, y) = (y - \mu)^2 - \mathbf{1}(y < 0)y^2$ and an analytical solution for the quantile of the distribution of $\lambda(\mu, y)$ for every μ, which we can use as valid q_α. We extend the results in Example 2.2.6 below to prove that the distribution of $\lambda(\mu, y)$ is stochastically dominated by a χ^2_1. See Figure 2.4 for an illustration.

Example 2.3.3 (Stochastic dominance for LLR for constrained Gaussian in one dimension)**.** *For the LLR $\lambda(\mu, y)$, when $y \sim \mathcal{N}(\mu, 1)$ under the null hypothesis, we have that, for $Z \sim \chi^2_1$, $Z \succeq \lambda(\mu, y)$ for all $\mu \geq 0$.*

Proof. See Section 2.8. □

For this example, given the stochastic dominance result, we can define $1 - \alpha$ confidence intervals using $\chi^2_{1,1-\alpha}$ instead of using $q_\alpha(\mu)$. This produces larger intervals than using the true quantile, but the true quantile in the closed form will generally be unavailable in more complex examples, while the presented stochastic dominance tools can still be used. The intervals using the χ^2_1 quantile are

$$\mathcal{I}_\alpha(y) := \begin{array}{rll} \min_x / \max_x & x & \\ \text{st} & x \geq 0 & \\ & (x - y)^2 \leq \chi^2_{1,1-\alpha} + \min_{x' \geq 0} (x' - y)^2. & \end{array} \tag{2.37}$$

General confidence interval construction

In this section, we present our meta-algorithm that uses the methodologies described in the preceding sections. The goal of this meta-algorithm is to construct a $1 - \alpha$ confidence interval for a given quantity of interest $\varphi(\boldsymbol{x}^*)$. The algorithmic steps are outlined in Algorithm 1.

It is worth noting that the optimization problems defined in (2.39) and (2.40) may not always be convex or straightforward to solve. However, their dual formulations can be constructed, offering provably valid confidence intervals for any feasible dual solution (Philip B. Stark, 1992b). We defer the exploration of specialized optimization techniques specifically tailored to solve (2.39) and (2.40) to future work.

Algorithm 1 Meta-algorithm for confidence interval construction

Input: Observed data $\boldsymbol{y}$, log-likelihood model $\ell_{\boldsymbol{x}}(\boldsymbol{y})$, quantity of interest functional φ, constraint set $\mathcal{X}$, miscoverage level α.

1: **Test statistic**: Write down the LLR test statistic

$$\lambda(\mu, \boldsymbol{y}) = \inf_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x} \in \mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}). \tag{2.38}$$

2: **Distribution control**: Control $F_{\boldsymbol{x}}$, the distribution of $\lambda(\varphi(\boldsymbol{x}), \boldsymbol{y})$ where $\boldsymbol{y} \sim P_{\boldsymbol{x}}$, for all $\boldsymbol{x} \in \mathcal{X}$, by either:

A. *Explicit solution*: Obtain $F_{\boldsymbol{x}}$ explicitly, and let $q_\alpha(\mu) := \sup_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha)$.

B. *Computational way to directly find valid* q_α (Section 2.3): Solve $\sup_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha)$ (to set $q_\alpha(\mu)$) or $\sup_{\boldsymbol{x} \in \mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha)$ (to set q_α) numerically.

C. *Analytical way using stochastic dominance* (Section 2.3): Construct a distribution X that stochastically dominates $F_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$, and let $q_\alpha := Q_X(1-\alpha)$, or construct distributions X_μ that stochastically dominate $F_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}$ and let $q_\alpha(\mu) := Q_{X_\mu}(1-\alpha)$.

3: **Confidence interval calculation**: Obtain the confidence intervals by solving the pair of optimization problems that is easier in the particular case:

I. Parameter space formulation:

$$\begin{aligned} \min_{\boldsymbol{x}} / \max_{\boldsymbol{x}} \quad & \varphi(\boldsymbol{x}) \\ \text{st} \quad & \boldsymbol{x} \in \mathcal{X} \\ & -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \le q_\alpha(\varphi(\boldsymbol{x})) + \inf_{\boldsymbol{x}' \in \mathcal{X}} -2\ell_{\boldsymbol{x}'}(\boldsymbol{y}). \end{aligned} \tag{2.39}$$

II. Functional space formulation:

$$\begin{aligned} \min_{\mu} / \max_{\mu} \quad & \mu \\ \text{st} \quad & \mu \in \varphi(\mathcal{X}) \subseteq \mathbb{R} \\ & \inf_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x} \in \mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \le q_\alpha(\mu). \end{aligned} \tag{2.40}$$

Output: Confidence interval with coverage $1-\alpha$.

2.4 Refuting the Burrus conjecture

As discussed in Section 2.1, the family of constrained problems that has received the most attention is the positivity-constrained version of the problem as described

in Section 2.2. To recap, the model is defined as follows:

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}^*, \boldsymbol{I}_m) \quad \text{with} \quad \mathcal{X} = \{\boldsymbol{x} : \boldsymbol{x} \geq \boldsymbol{0}\} \quad \text{and} \quad \varphi(\boldsymbol{x}^*) = \boldsymbol{h}^\top \boldsymbol{x}^*. \tag{2.41}$$

Here $\boldsymbol{K} \in \mathbb{R}^{m \times p}$ is the forward linear operator. We again emphasize here that $\boldsymbol{K}$ need not have full column rank, so we can for example have $p > m$. It was initially conjectured in (W. R. Burrus, 1965; Burt W. Rust and Walter R. Burrus, 1972) that a valid $1 - \alpha$ confidence interval could be obtained as

$$\begin{aligned} \min_{\boldsymbol{x}}/\max_{\boldsymbol{x}} \quad & \boldsymbol{h}^\top \boldsymbol{x} \\ \text{st} \quad & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 \leq \psi_\alpha^2 \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned} \tag{2.42}$$

Here, $\psi_\alpha^2 = z_{\alpha/2}^2 + s^2(\boldsymbol{y})$, with $z_{\alpha/2}$ being the previously defined standard Gaussian quantile, and $s^2(\boldsymbol{y})$ is defined as the optimal value of

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 \\ \text{st} \quad & \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned}$$

Although initially believed to be proved in (Bert W. Rust and O'Leary, 1994), an error in the proof was later identified in (Tenorio, Fleck, and Moses, 2007), along with a counterexample. However, we demonstrate that this counterexample actually satisfies the conjecture, leaving the conjecture unresolved until now prior to our work, to the best of our knowledge.

The main result of this section is the construction of a new valid counterexample using the test inversion perspective developed in Section 2.2 and the stochastic dominance approach of Section 2.3, disproving the conjecture.

Theorem 2.4.1 (Refutation of the Burrus conjecture)**.** *The Burrus conjecture is false in general. The two-dimensional example previously proposed of a particular instance of* (2.41) *in (Tenorio, Fleck, and Moses, 2007),*

$$\boldsymbol{K} = \boldsymbol{I}_2 \quad \textit{and} \quad \boldsymbol{h} = (1, -1)^\top \quad \textit{with} \quad \boldsymbol{x}^* = (a, a)^\top \textit{ such that } a \geq 0,$$

does not constitute a valid counterexample to the Burrus conjecture. However, the following constitutes a valid counterexample for the Burrus conjecture:

$$\boldsymbol{K} = \boldsymbol{I}_3 \quad \textit{and} \quad \boldsymbol{h} = (1, 1, -1)^\top \quad \textit{with} \quad \boldsymbol{x}^* = (0, 0, 1)^\top.$$

The main idea of the proof is to first connect the conjecture to our framework, identifying the conjectured intervals as a particular case of our construction with a particular choice of q_α. We then apply Proposition 2.2.5 to show that coverage is equivalent to a valid choice of q_α. Finally, we present a counterexample to prove that the proposed q_α is not universally valid. The proof is divided into several lemmas for clarity.

Our approach is novel in that it diverges from previous geometric perspectives on the Gaussian likelihood (Burt W. Rust and Walter R. Burrus, 1972; Bert W. Rust and O'Leary, 1994; O'Leary and Bert W. Rust, 1986), instead leveraging the test inversion and stochastic dominance perspectives developed in Section 2.2 and Section 2.3.

Proof outline of Theorem 2.4.1

This subsection provides a structured outline of the proof for Theorem 2.4.1, which refutes the Burrus conjecture. We break down the proof into several key lemmas.

Lemma 2.4.2 (Framing the Burrus conjecture as test inversion)**.** *The construction of intervals in* (2.42) *for a particular instance of the problem* $(\boldsymbol{x}^*, \boldsymbol{K}, \boldsymbol{h})$ *is equivalent to the general construction in Theorem 2.2.4 for the model* $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}^*, \boldsymbol{I}_m)$*, with* $\boldsymbol{x}^* \geq \boldsymbol{0}$ *component wise, and* $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x}$*, using the threshold* $q_\alpha(\mu) = z^2_{\alpha/2}$ *independent of* μ*. Therefore, it is equivalent to inverting a hypothesis test* $H_0 : \boldsymbol{h}^\top \boldsymbol{x} = \mu$ *versus* $H_1 : \boldsymbol{h}^\top \boldsymbol{x} \neq \mu$ *with LLR*

$$\lambda(\mu, \boldsymbol{y}) \;:=\; \min_{\substack{\boldsymbol{h}^\top \boldsymbol{x} = \mu \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2. \tag{2.43}$$

Furthermore, the interval has correct coverage if and only if $q_\alpha = z^2_{\alpha/2}$ *is valid in the sense of satisfying the false positive guarantee* (2.17)*.*

Proof. See Section 2.9. □

Lemma 2.4.3 (Reducing the Burrus conjecture to stochastic dominance)**.** *The construction of intervals in* (2.42) *has the right coverage for any* α *(and hence the conjecture holds) for a particular instance of the problem* $(\boldsymbol{x}^*, \boldsymbol{K}, \boldsymbol{h})$ *if and only if the log-likelihood ratio test statistic*

$$\lambda(\mu = \boldsymbol{h}^\top \boldsymbol{x}^*, \boldsymbol{y}) \;:=\; \min_{\substack{\boldsymbol{h}^\top \boldsymbol{x} = \boldsymbol{h}^\top \boldsymbol{x}^* \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2$$

is stochastically dominated by a χ^2_1 *distribution whenever* $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}^*, \boldsymbol{I}_m)$*.*

Proof. See Section 2.9. □

As an example, the constrained one-dimensional example considered in Section 2.2 satisfies the stochastic dominance result and hence the conjecture. Furthermore, using Example 2.2.7, an alternative characterization of the conjecture is the stochastic dominance of the unconstrained LLR test statistic $\min_{\boldsymbol{h}^\top \boldsymbol{x} = \boldsymbol{h}^\top \boldsymbol{x}^*} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2$ over the constrained test statistic $\min_{\substack{\boldsymbol{h}^\top \boldsymbol{x} = \boldsymbol{h}^\top \boldsymbol{x}^* \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2$. We use Lemma 2.4.3 to prove both that the example in (Tenorio, Fleck, and Moses, 2007) obeys the conjecture and that our new counterexample does not.

Invalidity of a previous counterexample in two dimensions The previously proposed counterexample from (Tenorio, Fleck, and Moses, 2007) is a two-dimensional problem with $\boldsymbol{K} = \boldsymbol{I}_2$, $\boldsymbol{x}^* = (a, a)^\top$ with $a \geq 0$, $\boldsymbol{h} = (1, -1)^\top$ (and therefore $\mu^* = \boldsymbol{h}^\top \boldsymbol{x}^* = 0$). The LLR test statistic is

$$\lambda(\mu^* = 0, \boldsymbol{y}) = \min_{\substack{x_1 = x_2 \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

which, after solving the optimization problems, is equal to

$$\lambda(\mu^*, \boldsymbol{y}) = \begin{cases} y_1^2 + y_2^2 - (y_1 - \max(y_1, 0))^2 - (y_2 - \max(y_2, 0))^2, & y_1 + y_2 < 0, \\ \frac{1}{2}(y_1 - y_2)^2 - (y_1 - \max(y_1, 0))^2 - (y_2 - \max(y_2, 0))^2, & y_1 + y_2 \geq 0, \end{cases}$$

which we can equivalently write as

$$\begin{aligned} \lambda(\mu^*, \boldsymbol{y}) = (y_1^2 + y_2^2)\mathbb{1}\{y_1 + y_2 < 0\} + \frac{1}{2}(y_1 - y_2)^2 \mathbb{1}\{y_1 + y_2 \geq 0\} \\ - y_1^2 \mathbb{1}\{y_1 < 0\} - y_2^2 \mathbb{1}\{y_2 < 0\}. \end{aligned} \tag{2.44}$$

Lemma 2.4.4 (Invalidity of a previous counterexample)**.** *The LLR statistic $\lambda(\mu^*, \boldsymbol{y})$ in* (2.44) *is stochastically dominated by a χ_1^2 random variable whenever $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}^*, \boldsymbol{I}_2)$, $\boldsymbol{x}^* = (a, a)^\top$ for $a \geq 0$, and $\boldsymbol{h} = (1, -1)^\top$. Therefore, it does not constitute a valid counterexample to the conjecture.*

Proof sketch. The proof follows from a coupling argument between the LLR and a χ_1^2 random variable. See Section 2.9 for proof details. □

In summary, we used Lemma 2.4.3 to demonstrate that the previously proposed counterexample actually satisfies the conjecture.

A new provably valid counterexample in three dimensions We now present a new counterexample in $\mathbb{R}^3$ to refute the Burrus conjecture. Specifically, we consider $\boldsymbol{K} = \boldsymbol{I}_3$, $\boldsymbol{x}^* = (0, 0, 1)^\top$, and $\boldsymbol{h} = (1, 1, -1)^\top$, yielding $\mu^* = -1$. We prove that χ^2_1 does not stochastically dominate $\lambda(\mu^*, \boldsymbol{y})$, which in this case is

$$\lambda(\mu^* = -1, \boldsymbol{y}) = \min_{\substack{x_1+x_2-x_3=-1 \\ \boldsymbol{x}\geq\boldsymbol{0}}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \min_{\boldsymbol{x}\geq\boldsymbol{0}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{2.45}$$

We prove that $\mathbb{E}[\lambda(\mu^*, \boldsymbol{y})] > \mathbb{E}[\chi^2_1] = 1$. Here, the expectation is taken with respect to $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}^*, \boldsymbol{I}_3)$, and the inequality is a general sufficient condition to refute stochastic dominance and hence for the conjecture to break.

Lemma 2.4.5 (Validity of a new counterexample)**.** *$\lambda(\mu^*, \boldsymbol{y})$ in* (2.45) *is* not *stochastically dominated by a χ^2_1 random variable whenever $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}^*, \boldsymbol{I}_3)$ with $\boldsymbol{x}^* = (0, 0, 1)^\top$. Therefore, it constitutes a valid counterexample to the general conjecture.*

Proof sketch. We compute the expected value and show that it is greater than 1 (the expected value of a χ^2_1), therefore proving stochastic dominance. See Section 2.9 for the proof details. □

Remark 5 (A more general counterexample)**.** The validity of the counterexample does not hinge on $\boldsymbol{x}^*$ being on the boundary of the constraint set. In fact, the example remains valid for $\boldsymbol{x}^* = (\varepsilon, \varepsilon, 1)^\top$ with $\varepsilon > 0$ sufficiently small. We choose $\varepsilon = 0$ for the simplicity of the proof. See Figure 2.10 for numerical evidence, where the quantiles over the dashed line correspond to valid counterexamples.

Figure 2.5 shows the difference between the two examples. By plotting the difference between the CDF of λ (obtained numerically with $N = 10^6$ samples) and the CDF of a χ^2_1 distribution, we observe stochastic dominance for the two-dimensional example in Figure 2.5 (left panel) and no stochastic dominance (hence breaking of the conjecture) for the three-dimensional example in Figure 2.5 (right panel). Section 2.5 contains numerical coverage studies for both scenarios agreeing with the observation made here.

A negative result in high dimensions

After establishing that the χ^2_1 distribution fails to stochastically dominate the constrained log-likelihood ratio, a natural question arises: Is there another distribution, possibly within the χ^2_k family, that can stochastically dominate the constrained LLR? If such a distribution exists, it would allow us to redefine ψ^2_α in (2.42) as $s^2 + Q_X(1-\alpha)$,

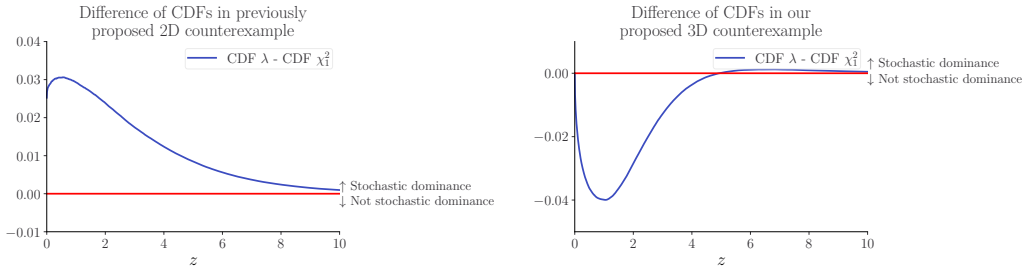


Figure 2.5: Difference of cumulative distribution functions between the LLR test statistic and χ^2_1 distribution for the statistics defined in (2.44) **(left)** and (2.45) **(right)**. Stochastic dominance, which is equivalent to the Burrus conjecture, is broken in the right example only. There is a direct correspondence between the points at which the CDF difference is negative and confidence levels $1 - \alpha$ that fail to hold (see Remark 3).

making the Q_X term in the optimization problem dimension independent, leading to intervals with shorter length in large dimensions. It is worth noting that in the unconstrained scenario, the LLR distribution is precisely χ^2_1, regardless of the dimensionality of the problem. However, the following proposition shows that no such dimension-independent distribution exists for the constrained case.

Proposition 2.4.6 (A negative result in high dimensions)**.** *The family of constrained LLRs for general $\boldsymbol{K}, \boldsymbol{h}$ in arbitrary dimensions, defined as*

$$\lambda(\mu = \boldsymbol{h}^\top \boldsymbol{x}^*, \boldsymbol{y}) \;=\; \min_{\substack{\boldsymbol{h}^\top \boldsymbol{x} = \boldsymbol{h}^\top \boldsymbol{x}^* \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2,$$

cannot be stochastically dominated a dimension-independent way by any finite-mean distribution (including all χ^2_k for $k \geq 1$).

Proof sketch. We construct a sequence of examples with increasing dimensions and demonstrate that the expected value of the constrained LLR grows unbounded as the dimension increases. This result negates the possibility of stochastic dominance by any finite-mean distribution. For a detailed proof, see Section 2.9. □

2.5 Numerical examples

In this section, we provide numerical illustrations for the procedures and theoretical results described above. In particular, we analyze coverage properties of four types of intervals. The first two intervals come from previous works: $\mathcal{I}_{\mathsf{SSB}}$ (2.3), which has provably correct coverage but is known to overcover when inferring a single functional, and $\mathcal{I}_{\mathsf{OSB}}$ (2.5), which comes from the Burrus conjecture and, as we

proved in this work, does not correctly cover in general. The last two are the interval constructions described in this work by solving the max quantile problems (2.20) and (2.21) to find $q_\alpha(\mu)$ (Section 2.3): $\mathcal{I}_{\mathsf{MQ}}$ (where q_α does not depend on μ) and the more refined $\mathcal{I}_{\mathsf{MQ}\mu}$ (where q_α depends on μ). Both of these intervals have provable coverage.

We analyze four settings in which the Burrus conjecture applies, including a one-dimensional example in Section 2.5, the previously proposed invalid counterexample in Section 2.5 (for which we prove in Lemma 2.4.4 that $\mathcal{I}_{\mathsf{OSB}}$ correctly covers), our proposed counterexample in Section 2.5 (for which we prove in Lemma 2.4.5 that $\mathcal{I}_{\mathsf{OSB}}$ does not cover for at least some level $1-\alpha$ and a certain $\boldsymbol{x}^*$), and a setting with a bounded constraint set[11] $\mathcal{X}$ in Section 2.5.

Throughout the examples, the observed coverage (or lack thereof) agrees with the developed theoretical results, and we observe that our interval $\mathcal{I}_{\mathsf{MQ}\mu}$ consistently fixes the miscalibration of the other interval types: when $\mathcal{I}_{\mathsf{OSB}}$ undercovers, $\mathcal{I}_{\mathsf{MQ}\mu}$ is on average longer than $\mathcal{I}_{\mathsf{OSB}}$ to obtain coverage, and when $\mathcal{I}_{\mathsf{OSB}}$ overcovers, $\mathcal{I}_{\mathsf{MQ}\mu}$ is on average shorter than $\mathcal{I}_{\mathsf{OSB}}$ with coverage closer to the prescribed level $1-\alpha$.

Constrained Gaussian in one dimension

We revisit the constrained Gaussian model in one dimension (2.26) described in Section 2.2, $y = x^* + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$, $x^* \geq 0$, and $\varphi(x) = x$. We perform a simulation experiment using six true parameter settings of $x^* \in \{0, 2^{-3}, 2^{-2}, 2^{-1}, 2^0, 2^1\}$. We focus on settings closer to the boundary since that is where the biggest differences between the considered intervals exist. For each of these settings, we simulate 10^5 observations according to the model (2.26) and compute three different 95% confidence intervals for each sample: interval (2.24) using the actual quantile function given in (2.29) ($\mathcal{I}_{\mathsf{MQ}_\mu}$[12]), interval (2.24) using the stochastically dominating $Q_{\chi^2_1}(1-\alpha)$ quantile ($\mathcal{I}_{\mathsf{OSB}}$, which in this problem is equal to $\mathcal{I}_{\mathsf{MQ}}$), and the standard Truncated Gaussian interval, which equals the SSB interval in this case ($\mathcal{I}_{\mathsf{SSB}}$). The

[11] Strictly speaking, this setting is not included in the original Burrus conjecture but is later extended in (Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022).

[12] Since this is a one-dimensional problem with $\varphi(x) = x$, this is equivalent to being able to solve all the $\mu-$dependant quantile optimization problems

intervals computed with the true quantile function are characterized by:

$$\begin{aligned} \mathcal{I}_{\mathsf{MQ}_\mu}(y) := \min_x/\max_x \quad & x \\ \text{st} \quad & x \geq 0 \\ & (x-y)^2 \leq q_\alpha(x) + \min_{x \geq 0}(x-y)^2, \end{aligned} \tag{2.46}$$

where $q_\alpha(x)$ is given by (2.29). For the stochastically dominating $Q_{\chi^2_1}(1-\alpha)$, the interval in (2.24) becomes:

$$\begin{aligned} \mathcal{I}_{\mathsf{OSB}}(y) = \min_x/\max_x \quad & x \\ \text{st} \quad & x \geq 0 \\ & (x-y)^2 \leq Q_{\chi^2_1}(1-\alpha) + \min_{x \geq 0}(x-y)^2. \end{aligned} \tag{2.47}$$

Finally, the truncated Gaussian interval, which is shown below to be equivalent to the SSB interval in this case, is defined as:

$$\mathcal{I}_{\mathsf{SSB}}(y) := \left[y - z_{\alpha/2},\ y + z_{\alpha/2}\right] \cap \mathbb{R}_{\geq 0}. \tag{2.48}$$

Observe that (2.47) admits an explicit solution:

$$\mathcal{I}_{\mathsf{OSB}}(y) = \begin{cases} [y - \sqrt{Q_{\chi^2_1}(1-\alpha)},\ y + \sqrt{Q_{\chi^2_1}(1-\alpha)}] \cap \mathbb{R}_{\geq 0}, & y \geq 0 \\ [y - \sqrt{Q_{\chi^2_1}(1-\alpha) + y^2},\ y + \sqrt{Q_{\chi^2_1}(1-\alpha) + y^2}] \cap \mathbb{R}_{\geq 0}, & y < 0. \end{cases} \tag{2.49}$$

Furthermore, note that $\sqrt{Q_{\chi^2_1}(1-\alpha)} = z_{\alpha/2}$, so that (2.49) is always larger than or equal to (2.48). Conversely, we can express (2.48) as the solution to optimization problems, illustrating that the truncated Gaussian interval is equivalent to the SSB interval for this case:

$$\begin{aligned} \mathcal{I}_{\mathsf{SSB}}(y) = \min_x/\max_x \quad & x \\ \text{st} \quad & x \geq 0 \\ & (x-y)^2 \leq z^2_{\alpha/2}. \end{aligned} \tag{2.50}$$

To empirically estimate coverage, for each x^* setting and each interval type, we compute 10^5 intervals and keep track of their coverage of the true parameter. The left panel in Figure 2.6 shows how $\mathcal{I}_{\mathsf{OSB}}$ based on $Q_{\chi^2_1}(1-\alpha)$ over-covers when the true parameter is on the boundary, which makes sense as this setting of q_α holds for all x^*, and therefore is a conservative quantile. As expected, the interval

computed with $q_\alpha(x)$ maintains the nominal 95% coverage over all considered x^* values. This shows that knowing the quantile function means that we can compute an interval with exact nominal coverage that is adaptive to the unknown true parameter value. Additionally, we note that as x^* grows, the estimated coverage values across these methods converge, illustrating the intuition that when x^* gets sufficiently far from the constraint boundary, the problem is essentially unconstrained, and all considered methods produce nearly identical results. The right panel in Figure 2.6 shows each interval's expected length as a function of x^*. Again, we observe the tightness of the interval (2.24) constructed with the true quantile function $q_\alpha(x)$ compared to the interval constructed with the stochastically dominating quantile, $Q_{\chi^2_1}(1-\alpha)$. Similarly to coverage, as x^* grows, the expected interval lengths of $\mathcal{I}_{\mathsf{MQ}_\mu}$ and $\mathcal{I}_{\mathsf{OSB}}$ converge, and the methods become indistinguishable. Also, observe that the truncated Gaussian intervals have a smaller expected length compared to the intervals computed with $q_\alpha(x)$. We note that this length observation is particular to this one-dimensional example, as OSB intervals have been shown to be shorter than SSB intervals in higher dimensional problems (O'Leary and Bert W. Rust, 1986; Burt W. Rust and Walter R. Burrus, 1972; Stanley, Patil, and Kuusela, 2022).

Constrained Gaussian in two dimensions

We consider the Gaussian linear model in (2.4) with $\boldsymbol{K} = \boldsymbol{I}_2$, $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x} = x_1 - x_2$ and $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} \geq 0\}$. The work (Tenorio, Fleck, and Moses, 2007) proposes this scenario as a counterexample to the Burrus conjecture, but as shown in Lemma 2.4.4, it is in fact a case where the χ^2_1 distribution stochastically dominates the LLR for all true $\boldsymbol{x}^* \in \{\boldsymbol{x} : \boldsymbol{h}^\top \boldsymbol{x} = 0, \boldsymbol{x} \geq 0\}$, so the OSB intervals proposed by the conjecture have provably correct coverage for $\boldsymbol{x}^*$ in this set.

We estimate interval coverage with $1 - \alpha = 0.95$ for the four types of intervals ($\mathcal{I}_{\mathsf{SSB}}$, $\mathcal{I}_{\mathsf{OSB}}$, $\mathcal{I}_{\mathsf{MQ}}$ and $\mathcal{I}_{\mathsf{MQ}_\mu}$) for three true parameter values, two of them inside the $\{\boldsymbol{x} : \boldsymbol{h}^\top \boldsymbol{x} = 0, \boldsymbol{x} \geq 0\}$ region in which Lemma 2.4.4 applies, and one outside. In this and all the examples that follow, we solve the max quantile optimization problems (2.20), (2.21) using the Bayesian Optimization package BayesOpt (Martinez-Cantin, 2014) (see fig. 2.8 for an illustration of the quantile function being optimized in this particular example). We solve the outer optimization problems with the convex optimization package CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018) (for the convex problems in $\mathcal{I}_{\mathsf{SSB}}$, $\mathcal{I}_{\mathsf{OSB}}$ and $\mathcal{I}_{\mathsf{MQ}}$) or root-finding numerical algorithms (in the case of $\mathcal{I}_{\mathsf{MQ}_\mu}$, where we use the functional space formulation in algorithm 1 and look for those μ satisfying the constraint as equality).

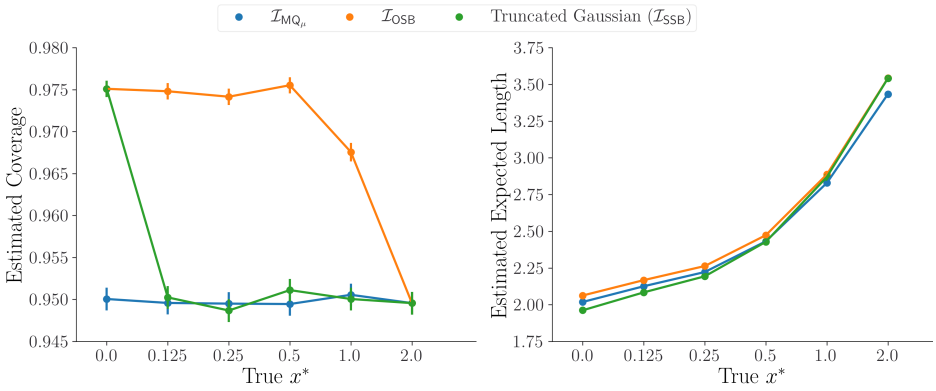


Figure 2.6: The **(left)** figure shows estimated coverage for each 95% confidence interval and each true x^* for the one-dimensional constrained Gaussian model. Both the Truncated Gaussian (SSB) and OSB intervals overcover when $x^* = 0$ while the MQμ interval predictably achieves nominal coverage. All the coverage values converge as x^* gets larger, as the problem moves toward the unconstrained problem where all the intervals are effectively the same. The intervals surrounding the estimated values are 95% Clopper–Pearson intervals, expressing the Monte Carlo uncertainty of each coverage estimate. The **(right)** figure shows an estimate of the expected interval length for each method (with 95% confidence intervals that are nearly length zero since the standard error of each estimate is nearly zero with 10^5 realizations each). Similarly to the coverage results in the left panel, as x^* gets larger, the expected OSB and MQμ interval lengths converge while the SSB intervals remain slightly larger.

For each parameter value and all intervals, coverage and expected length are estimated by drawing 5×10^4 observations from the data generating process, computing all interval types for each generated observation, and then checking coverage and length. The coverage confidence intervals are 95% Clopper–Pearson intervals for a binomial parameter, whereas the length confidence intervals are standard asymptotic Gaussian intervals using sample means and standard errors.

The results are shown in fig. 2.7. We observe correct coverage for the OSB intervals, in agreement with Lemma 2.4.4 (which applies for $\boldsymbol{x}^* = (0, 0)^\top$ and $\boldsymbol{x}^* = (0.33, 0.33)^\top$). We observe that the SSB intervals are the longest on average and tend to overcover, and that the MQ and MQμ intervals have nearly identical properties to the OSB intervals. This is because, for this problem setting, solving the optimization problems (2.20) and (2.21) recovers the χ^2_1 quantile (up to the numerical precision of the optimization solvers) of the Burrus conjecture as the maximum quantile (both for all $\mathcal{X}$ and for $\boldsymbol{h}^\top \boldsymbol{x} = \mu$ for any μ), so both of those intervals actually recover the OSB intervals in this case.

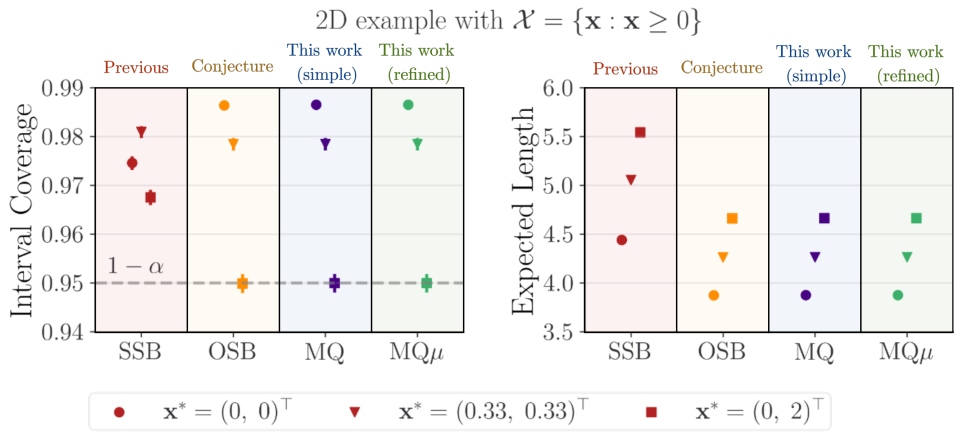


Figure 2.7: Estimated interval coverage **(left)** and expected lengths **(right)** for 95% intervals resulting from the SSB, OSB, MQ, and MQμ methods for the Gaussian linear model in (2.4) with $\boldsymbol{K} = \boldsymbol{I}_2$, $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x} = x_1 - x_2$, and $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{x} \geq 0\}$.

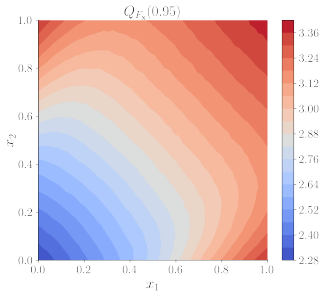


Figure 2.8: Estimated 95th quantiles from the LLR test statistic null distributions in the region $[0, 1]^2 \subset \mathbb{R}^2_+$ where color shows the estimated quantiles. In contrast to the unconstrained case, the quantile is dependent on the true parameter, and the quantile surface is non-trivial.

Constrained Gaussian in three dimensions

We use the three-dimensional counterexample of the Burrus conjecture from Section 2.4 to numerically show that the $\mathcal{I}_{\mathsf{OSB}}$ intervals undercover in this example and that the max-quantile intervals are able to fix the undercoverage. Concretely, we consider the Gaussian linear model in (2.4) with $\boldsymbol{K} = \boldsymbol{I}_3$, $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x} = x_1 + x_2 - x_3$ and $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{x} \geq 0\}$. We repeat the same experimental setup as in Section 2.5, comparing the four interval types for $\boldsymbol{x}^* = (0, 0, 0)^\top$ and $\boldsymbol{x}^* = (0, 0, 1)^\top$; this last parameter value is the one analyzed in Lemma 2.4.5 and for which we know the OSB interval can undercover for some α. Figure 2.9 shows the results for $1 - \alpha = 0.68$ (one

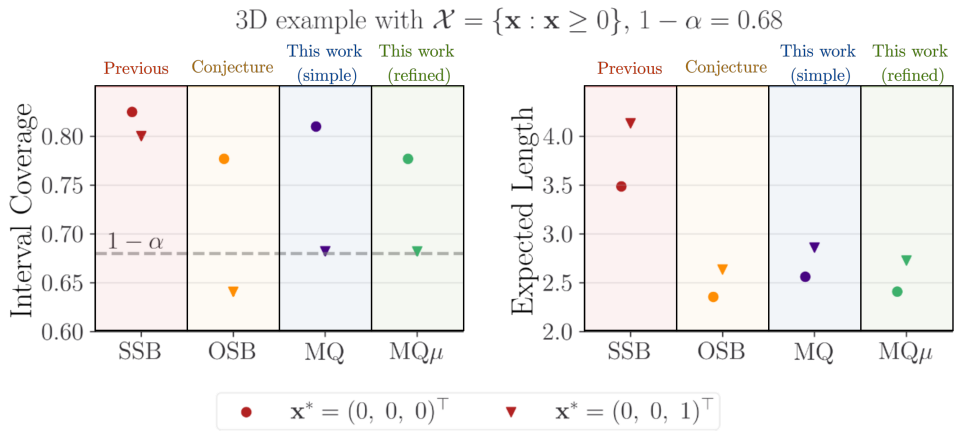


Figure 2.9: Estimated interval coverage **(left)** and expected lengths **(right)** for 68% intervals resulting from the SSB, OSB, MQ, and MQμ methods for the Gaussian linear model in (2.4) with $\boldsymbol{K} = \boldsymbol{I}_3$, $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x} = x_1 + x_2 - x_3$, and $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{x} \geq 0\}$.

sigma interval coverage), and we include in Section 2.10 the results for $1 - \alpha = 0.95$, which lead to the same conclusions. Furthermore, to illustrate that the conjecture breaks in an area around the studied point $(0, 0, 1)^\top$ in Figure 2.10, we plot the numerically estimated 68% LLR test statistic quantiles for parameters of the form $(t, t, 1)$, showing that there is a range of points with a larger quantile than the χ^2_1 quantile, implying undercoverage.

The results agree with our theoretical findings in Lemma 2.4.5, as we observe that the OSB intervals undercover both at 95% and 68% confidence levels, thus invalidating the Burrus conjecture. In contrast, the SSB, MQ and MQμ intervals all have provable coverage in this scenario, which is reflected in the estimated coverage values. Notably, the MQμ intervals are not much longer than the OSB intervals, but they obtain the required coverage, enlarging the conjectured intervals just enough. The simpler MQ intervals overcover a bit more, which illustrates the benefit of solving (2.21) over (2.20) when computationally feasible. Nevertheless, their length is not much larger than MQμ and significantly smaller than for SSB, the other simple method with coverage guarantees.

Bounded constraint set in two dimensions

As a last case study, we consider a modification of the example in Section 2.5 in which the constraint set is chosen to be the bounded set $\mathcal{X} = [0, 1]^2$. While not in the original scope of the Burrus conjecture, which only considers $\mathcal{X} = \{\boldsymbol{x} : \boldsymbol{x} \geq 0\}$,

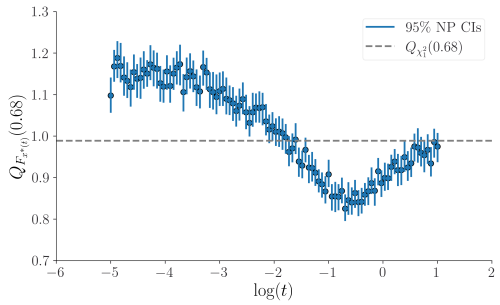


Figure 2.10: For the numerical example in Section 2.5, considering $\boldsymbol{x}^* \in \{\boldsymbol{x}^*(t) = (t, t, 1)^\top : 0 < t \leq e\} \subset \mathbb{R}^3$, we estimate the 68% LLR test statistic quantiles along with 95% nonparametric (NP) confidence intervals for percentiles (Hahn and Meeker, 1991). The test statistic quantiles exceeding $\chi^2_{1,0.32}$ correspond to the Burrus conjecture failing in this scenario. We note that for this example, the Burrus conjecture fails close to the constraint boundary while the $Q_{\chi^2_1}(0.68)$ quantile becomes valid once sufficiently far from the boundary.

as mentioned in Section 2.1, the $\mathcal{I}_{\mathsf{OSB}}$ interval construction has since then been used with constraints of the form $\mathbf{A}\boldsymbol{x} \leq \mathbf{b}$ by replacing $\boldsymbol{x} \geq 0$ with $\mathbf{A}\boldsymbol{x} \leq \mathbf{b}$ in the optimization problems (2.5) and (2.6) (Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022). We use the same experimental setup as in Section 2.5, taking into account that the change in $\mathcal{X}$ affects both the interval optimization problem and the optimizations required for $q_\alpha(\mu)$ (since the LLR statistic changes as well).

The results are shown in Figure 2.11. We observe that in this setting, as opposed to the previous three-dimensional problem, the OSB intervals *overcover*, because the maximum quantile of the LLR test statistic over the constraint set is *smaller* than the χ^2_1 quantile. Furthermore, our intervals MQ, and especially MQμ, are able to exploit this fact to obtain shorter intervals than OSB with coverage closer to $1 - \alpha$ by using the actual max quantiles over $\mathcal{X}$ instead of the χ^2_1 quantile used by OSB.

2.6 Discussion

This paper presents a framework for constructing confidence intervals with guaranteed frequentist coverage for a given functional of forward model parameters in the presence of constraints. For the specific case of the Gaussian linear forward model with nonnegativity constraints, we refute the Burrus conjecture (W. R. Burrus, 1965) by providing a counterexample and propose a more general approach for interval construction. Our approach hinges on the inversion of a specific like-

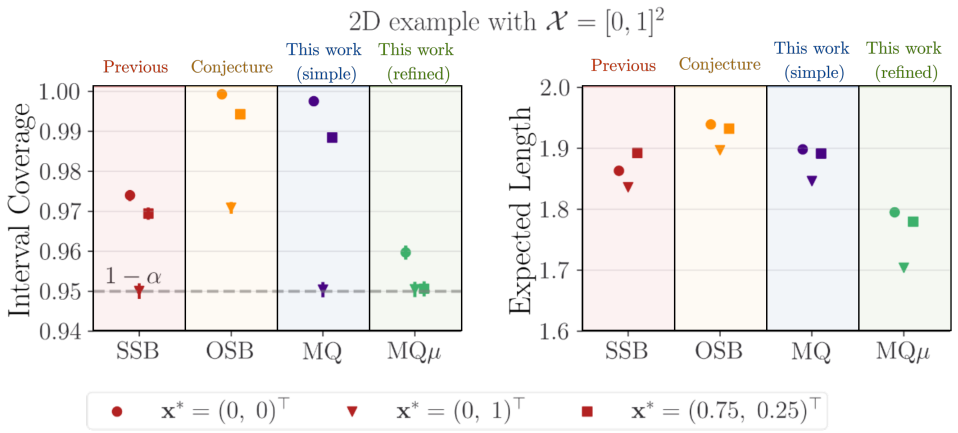


Figure 2.11: Estimated interval coverage **(left)** and expected lengths **(right)** for 95% intervals resulting from the SSB, OSB, MQ, and MQμ methods for the Gaussian linear model in (2.4) with $\boldsymbol{K} = \boldsymbol{I}_2$, $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x} = x_1 - x_2$, and $\mathcal{X} = [0, 1]^2 \subset \mathbb{R}^2$.

lihood ratio test, and we offer theoretical and practical insights into the properties of the constructed intervals via illustrative examples. Our framework is versatile, accommodating potentially nonlinear, non-Gaussian, and rank-deficient settings.

At a high level, the practical effectiveness of UQ methods depends on the (sometimes implicit) assumptions of the method. Different methods come into play depending on what we assume or know, be it the likelihood, the constraints, or the prior in Bayesian settings. In classical statistics, confidence intervals serve as a valuable tool for UQ, especially for one-dimensional quantities of interest. These intervals are constructed to offer guaranteed coverage under repeated sampling, aligning with frequentist principles. While frequentist coverage guarantees are a useful criterion, especially in contexts where repeatability is essential, we acknowledge that the "best" UQ method is often context-dependent. For example, this frequentist approach is the most natural in applications like remote sensing (Patil, Kuusela, and Hobbs, 2022), where repeatability is a key requirement. Conversely, when it is natural to think of the parameter as arising from a prior distribution, Bayesian methods are well-motivated and have desirable properties.

A key aim of this paper is to serve as a basis for the future development of these UQ procedures. We conclude this paper by discussing a few possible directions for future work:

- **Data-adaptive calibration procedure.** We saw in Section 2.5 that MQ is valid where OSB is not and can leverage smaller quantiles to produce tighter intervals.

In the event that one does not have the assumed true parameter bounding box, it may be possible to create a data-generated one and adjust the error budget accordingly. Such a procedure would enable us to expand the use of the MQ intervals to scenarios with unbounded parameter constraints. We are currently investigating this approach, which will be the subject of a future follow-up paper.

- **Exploration of high-dimensional problems.** A key benefit of the optimization-based interval construction is that it promises to provide a solution to uncertainty quantification in high-dimensional and ultra-high-dimensional problems, where most alternative approaches (including sampling-based ones) are infeasible. For example, in (Stanley, Patil, and Kuusela, 2022), we applied a precursor of the methods presented here in a problem where $p = 80$, and we are currently exploring the use of these methods in a data-assimilation setting where $p \approx 10^4$. Indeed, optimization is one of the only computational techniques known to work in ultra-high-dimensional problems, such as those in 4DVar data assimilation (Kalmikov and Heimbach, 2014; Forget et al., 2015; J. Liu, Bowman, Meemong, et al., 2016). While optimization has been successfully used for point estimation in such problems, the approaches developed here may enable modifying the existing programs to obtain confidence intervals in addition to point estimates.

- **Joint confidence sets for multiple functionals.** Since our framework is devised for UQ of a single functional, its application to collections of functionals (a higher-dimensional quantity of interest), would be a natural and desirable extension. Trivially, given a collection of K functionals, one could apply this methodology K times and use the Bonferroni correction to adjust the confidence levels so that they all cover at the desired coverage level. Although this approach might be practically reasonable when K is small, it becomes markedly inefficient as K becomes large. Furthermore, this approach would create a K-dimensional hyper-rectangle for the quantity of interest, which may not be the optimal geometry for bounding the quantity of interest. As such, extending the framework of Section 2.3 to simultaneously consider the K functionals of interest would be the first step to creating a more nuanced approach. One way this can be achieved is by appropriately adjusting the definition of H_0 in the hypothesis test in (2.11).

- **Choice of test statistics beyond LLR.** The log-likelihood ratio test statistic considered in this work connects with the Rust–Burrus intervals and is observed to perform well in practice, but other choices can be explored in future work. While the LLR is a natural choice for the generic problem, im-

proving the interval length on particular families of problems with different test statistics might be possible. Since the main theoretical machinery comes from the test inversion framework, which is independent of the actual form of the test statistic, alternate versions of Theorem 2.2.4 can be constructed as long as the test statistic constructs valid level-α hypothesis tests; the resulting intervals of which could be explored theoretically and numerically. For instance, the construction of confidence intervals in Section 2.2, originally written for $\lambda(\mu, \boldsymbol{y}) = \inf_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y})$, readily generalizes to test statistics of the form $\lambda_{f,g}(\mu, \boldsymbol{y}) = \inf_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} f(\boldsymbol{x}, \boldsymbol{y}) - g(\boldsymbol{y})$. In that case, (2.22) becomes $\{\boldsymbol{x} : f(\boldsymbol{x}, \boldsymbol{y}) \leq q_\alpha(\varphi(\boldsymbol{x})) + g(\boldsymbol{y})\}$, where q_α must be valid for the particular choice of f and g, and can be obtained by analyzing the distribution of $\lambda_{f,g}$.

- **Generalization to simulation-based problems.** An extension of our methodology to settings in which the likelihood is not exactly known can be considered, ranging from only partial knowledge of the form of the likelihood to full simulation-based (likelihood-free) settings where the likelihood is not known explicitly but can be sampled from. A possible avenue is to develop robust worst-case approaches with respect to possible likelihoods. In a fully likelihood-free setting, approaches such as (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Heinrich, 2022; Masserano, Dorigo, et al., 2023; Niccolò Dalmasso, Masserano, et al., 2023) provide ways to invert hypothesis tests to obtain confidence sets in these scenarios in multidimensional parameter spaces. Projections of these sets could produce confidence intervals for a functional of the model parameters, as seen for the SSB intervals. However, as we have explored, orienting the hypothesis test to the given functional of interest can have dramatic length benefits for the resulting confidence interval (as seen for the OSB and MQ intervals). Since the log-likelihood plays a key role in the definition of our intervals, extensions providing ways to relax that dependence would be a necessary first step.

2.7 Proofs in Section 2.2

Proof of Lemma 2.2.1

To prove the lemma, we need to show that the probability of μ^* being in the confidence set $C_\alpha(\boldsymbol{y})$ is at least $1-\alpha$. Towards this end, observe that

$$\begin{aligned}\mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}^*}}(\mu^* \in C_\alpha(\boldsymbol{y})) &= \mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}^*}}(\boldsymbol{y} \in A_\alpha(\mu^*)) \\ &= 1 - \mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}^*}}(\boldsymbol{y} \notin A_\alpha(\mu^*)) \\ &\geq 1 - \sup_{\boldsymbol{x}\in\Phi_{\mu^*}\cap\mathcal{X}} \mathbb{P}_{\boldsymbol{y}\sim P_{\boldsymbol{x}}}(\boldsymbol{y} \notin A_\alpha(\mu^*)) \\ &\geq 1-\alpha,\end{aligned}$$

as desired. This completes the proof.

Proof of Lemma 2.2.2

The value $Q_\mu^{\max}$ is a valid decision value for any given μ, since for all $\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}$ it holds that

$$\mathbb{P}_{\lambda\sim F_{\boldsymbol{x}}}\left(\lambda > \sup_{\boldsymbol{x}'\in\Phi_\mu\cap\mathcal{X}} Q_{F_{\boldsymbol{x}'}}(1-\alpha)\right) \leq \mathbb{P}_{\lambda\sim F_{\boldsymbol{x}}}\left(\lambda > Q_{F_{\boldsymbol{x}}}(1-\alpha)\right) = \alpha. \tag{2.51}$$

For any $v < Q_\mu^{\max}$ that one could use as a decision value, there exists $\widetilde{\boldsymbol{x}} \in \Phi_\mu \cap \mathcal{X}$ such that $Q_{F_{\widetilde{\boldsymbol{x}}}}(1-\alpha) > v$. Therefore, $\mathbb{P}_{\lambda\sim F_{\widetilde{\boldsymbol{x}}}}(\lambda > v) > \alpha$, and thus v is not a valid decision value. $Q^{\max}$ is clearly valid for all μ, and a similar argument shows that choosing any smaller v would make it not valid, as there exists a $\widetilde{\boldsymbol{x}} \in \mathcal{X}$ with a larger quantile than v, making v invalid as a decision value for $\mu = \varphi(\widetilde{\boldsymbol{x}})$. The equality between the two formulations comes from the same argument that shows (2.17) is equivalent to (2.18).

Proof of Theorem 2.2.4

Assume $\bar{\mathcal{X}}_\alpha(\boldsymbol{y})$ is nonempty and write as shorthand $\inf_{\boldsymbol{x}\in\mathcal{X}}/\sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})$ for the interval

$$\left[\inf_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}),\ \sup_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x})\right].$$

Observe that

$$C_\alpha(\boldsymbol{y}) \subseteq \inf_{\mu\in C_\alpha(\boldsymbol{y})} / \sup_{\mu\in C_\alpha(\boldsymbol{y})} \mu. \tag{2.52}$$

From Lemma 2.2.1, $C_\alpha(\boldsymbol{y}) \subseteq \mathcal{I}_\alpha(\boldsymbol{y})$ implies that $\mathcal{I}_\alpha(\boldsymbol{y})$ is also a $1-\alpha$ confidence interval. We prove this interval exactly equals the defined $\mathcal{I}_\alpha(\boldsymbol{y})$ in (2.24). Unpacking

the definition of $C_\alpha(\boldsymbol{y})$, we write the interval

$$\begin{aligned} \inf_{\mu}/\sup_{\mu} \quad & \mu \\ \text{st} \quad & \mu \in \mathbb{R} \\ & -2\log\Lambda(\mu, \boldsymbol{y}) \leq q_\alpha(\mu). \end{aligned} \tag{2.53}$$

We can write different optimization problems which are equivalent to the optimization problem (2.53). First, we use the definition of Λ to write:

$$\begin{aligned} \inf_{\mu}/\sup_{\mu} \quad & \mu \\ \text{st} \quad & \mu \in \mathbb{R} \\ & \inf_{\varphi(\boldsymbol{x})=\mu, \boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\mu). \end{aligned} \tag{2.54}$$

Notice that we can rewrite the feasibility condition of μ as follows:

$$\inf_{\varphi(\boldsymbol{x})=\mu, \boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\mu) + \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y})$$

as there exists $\boldsymbol{x} \in \mathcal{X}$ such that $\varphi(\boldsymbol{x}) = \mu$ and

$$-2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\mu) + \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}).$$

Therefore, the optimization problem can be rewritten with $\boldsymbol{x}$ and μ as the optimization variables:

$$\begin{aligned} \inf_{\mu,\boldsymbol{x}}/\sup_{\mu,\boldsymbol{x}} \quad & \mu \\ \text{st} \quad & \boldsymbol{x} \in \mathcal{X},\ \mu \in \mathbb{R} \\ & \varphi(\boldsymbol{x}) = \mu \\ & -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\mu) + \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}). \end{aligned} \tag{2.55}$$

And μ can be eliminated using the constraint, yielding

$$\begin{aligned} \inf_{\boldsymbol{x}}/\sup_{\boldsymbol{x}} \quad & \varphi(\boldsymbol{x}) \\ \text{st} \quad & \boldsymbol{x} \in \mathcal{X} \\ & -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\varphi(\boldsymbol{x})) + \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}), \end{aligned} \tag{2.56}$$

that is, $\inf_{\boldsymbol{x}\in\bar{\mathcal{X}}_\alpha(\boldsymbol{y})}/\sup_{\boldsymbol{x}\in\bar{\mathcal{X}}_\alpha(\boldsymbol{y})} \varphi(\boldsymbol{x})$. The choice when $\bar{\mathcal{X}}_\alpha(\boldsymbol{y})$ is empty does not affect coverage properties. An alternative proof comes by observing the set equality $\varphi(\{\boldsymbol{x} \in \mathcal{X} : -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\varphi(\boldsymbol{x}))+\inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y})\}) = \{\mu \in \varphi(\mathcal{X}) : \inf_{\varphi(\boldsymbol{x})=\mu,\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y})- \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\mu)\}$, and the result follows. This finishes the proof.

Proof of Proposition 2.2.5

We have, by definition and test inversion, that $q_\alpha(\mu)$ are valid if and only if

$$C_\alpha(\boldsymbol{y}) := \{\mu \colon \lambda(\mu, \boldsymbol{y}) \leq q_\alpha(\mu)\}$$

is a valid $1-\alpha$ confidence interval for any $\boldsymbol{x} \in \mathcal{X}$. Since $\mathcal{I}_\alpha(\boldsymbol{y})$ is the smallest interval that contains $C_\alpha(\boldsymbol{y})$, we aim to prove that $C_\alpha(\boldsymbol{y})$ is already an interval (including singletons or empty sets), so that $C_\alpha(\boldsymbol{y}) = \mathcal{I}_\alpha(\boldsymbol{y})$ and the result holds. Define the function:

$$\mu \mapsto \mathcal{F}(\mu) = \inf_{\substack{\varphi(\boldsymbol{x})=\mu \\ \boldsymbol{x}\in\mathcal{X}}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \tag{2.57}$$

for a given $\boldsymbol{y}$, supported in all μ such that $\Phi_\mu \cap \mathcal{X} \neq \emptyset$. Write $C_\alpha(\boldsymbol{y})$ explicitly using (2.13), we get

$$C_\alpha(\boldsymbol{y}) := \left\{\mu \colon \inf_{\substack{\varphi(\boldsymbol{x})=\mu \\ \boldsymbol{x}\in\mathcal{X}}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha\right\}. \tag{2.58}$$

The second term on the left-hand side does not depend on μ, so it is enough to prove that any set of the form $\{\mu \colon \mathcal{F}(\mu) \leq z\}$ is an interval, which is implied by the function $\mathcal{F}(\mu)$ being convex in μ (for a fixed $\boldsymbol{y}$). Indeed, if the set is not an interval, we have $\mu^- < \mu < \mu^+$ with $\mu^-, \mu^+ \in C_\alpha(\boldsymbol{y})$ and $\mu \notin C_\alpha(\boldsymbol{y})$ which contradicts convexity, since

$$\mathcal{F}(\mu) \geq z > \gamma\mathcal{F}(\mu^-) + (1-\gamma)\mathcal{F}(\mu^+).$$

To see convexity and finish the proof, let $\mu_1 \neq \mu_2$ and let $\mathcal{G}(\boldsymbol{x}) := -2\ell_{\boldsymbol{x}}(\boldsymbol{y})$, a convex function by assumption. Write for $i = 1, 2$:

$$x_i \in \operatorname{argmin}_{\substack{\varphi(\boldsymbol{x})=\mu \\ \boldsymbol{x}\in\mathcal{X}}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}),$$

with x_i being any possible element in the set of minimizers, so that $\mathcal{F}(\mu_i) = \mathcal{G}(\boldsymbol{x}_i)$. For any $0 < \gamma < 1$, $\gamma\boldsymbol{x}_1 + (1-\gamma)\boldsymbol{x}_2 \in \mathcal{X}$ since $\mathcal{X}$ is a convex cone and

$$\varphi(\gamma\boldsymbol{x}_1 + (1-\gamma)\boldsymbol{x}_2) = \gamma\mu_1 + (1-\gamma)\mu_2,$$

since φ is linear, so $\gamma\boldsymbol{x}_1 + (1-\gamma)\boldsymbol{x}_2$ is a feasible point of the optimization problem

$$\inf_{\substack{\varphi(\boldsymbol{x})=\gamma\mu_1+(1-\gamma)\mu_2 \\ \boldsymbol{x}\in\mathcal{X}}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}),$$

that has optimal value $\mathcal{F}(\gamma\mu_1 + (1-\gamma)\mu_2)$. Therefore, by convexity of $\mathcal{G}$ and definition of the $\boldsymbol{x}_i$, we have that:

$$\mathcal{F}(\gamma\mu_1+(1-\gamma)\mu_2) \leq \mathcal{G}(\gamma\boldsymbol{x}_1+(1-\gamma)\boldsymbol{x}_2) \leq \gamma\mathcal{G}(\boldsymbol{x}_1)+(1-\gamma)\mathcal{G}(\boldsymbol{x}_2) = \gamma\mathcal{F}(\mu_1)+(1-\gamma)\mathcal{F}(\mu_2). \tag{2.59}$$

This completes the proof.

Proof of Example 2.2.6

Since the case when $\mu^* = 0$ is of particular interest, we show the result in this specific case and then generalize to the case of $\mu^* > 0$.

Case of $\mu^* = 0$ When $\mu^* = 0$, we can argue from symmetry of the standard Gaussian about the origin to write down the CDF in closed form. For $c \geq 0$, we have

$$\begin{aligned}\mathbb{P}_{\mu_0}(\ell_0 \leq c) &= \mathbb{P}_{\mu_0}(\ell_0 \leq c, y < 0) + \mathbb{P}_{\mu_0}(\ell_0 \leq c, y \geq 0) \\ &= \mathbb{P}_{\mu_0}(\ell_0 \leq c \mid y < 0)\,\mathbb{P}_{\mu_0}(y < 0) + \mathbb{P}_{\mu_0}(\ell_0 \leq c \mid y \geq 0)\,\mathbb{P}_{\mu_0}(y \geq 0).\end{aligned} \tag{2.60}$$

By definition, $\mathbb{P}_{\mu_0}(y < 0) = \mathbb{P}_{\mu_0}(y \geq 0) = \frac{1}{2}$, so only the conditional probabilities remain. By (2.28), we have

$$\begin{aligned}\mathbb{P}_{\mu_0}(\ell_0 \leq c \mid y < 0) &= \mathbb{P}_{\mu_0}(0 \leq c \mid y < 0) = 1 \\ \mathbb{P}_{\mu_0}(\ell_0 \leq c \mid y \geq 0) &= \mathbb{P}_{\mu_0}\left(y^2 \leq c \mid y \geq 0\right).\end{aligned} \tag{2.61}$$

In (2.61), we immediately observe that

$$\mathbb{P}_{\mu_0}\left(y^2 \leq c \mid y \geq 0\right) = \mathbb{P}_{\mu_0}\left(y^2 \leq c, y \geq 0\right)\mathbb{P}_{\mu_0}(y \geq 0)^{-1} = 2\mathbb{P}_{\mu_0}\left(0 \leq y \leq \sqrt{c}\right) = 2\Phi(\sqrt{c}) - 1. \tag{2.62}$$

But we also have that

$$\mathbb{P}_{\mu_0}\left(y^2 \leq c\right) = \mathbb{P}_{\mu_0}\left(-\sqrt{c} \leq y \leq \sqrt{c}\right) = 2\Phi(\sqrt{c}) - 1.$$

So we have

$$\mathbb{P}_{\mu_0}\left(y^2 \leq c \mid y \geq 0\right) = \mathbb{P}_{\mu_0}\left(y^2 \leq c\right).$$

Hence, we obtain

$$\mathbb{P}_{\mu_0}(\ell_0 \leq c \mid y \geq 0) = \chi^2_1(c).$$

Note this independence on the sign of y means that the magnitude of y is statistically independent of its direction. Thus, when $\mu_0 = 0$, the log-likelihood ratio has the following distribution:

$$\ell_0 \sim \frac{1}{2} + \frac{1}{2}\chi^2_1. \tag{2.63}$$

This completes the case when $\mu^* = 0$.

Case of $\mu^* > 0$ When $\mu > 0$, the closed-form solution to the CDF of ℓ_0 becomes more complicated, as we can no longer use symmetry around the origin. Picking up at (2.60), we first note that when $y \sim \mathcal{N}(\mu_0, 1)$, we have

$$\mathbb{P}_{\mu_0}(y < 0) = \Phi(-\mu_0) \quad \text{and} \quad \mathbb{P}_{\mu_0}(y \geq 0) = \Phi(\mu_0).$$

Next, we must find the conditional probabilities. Starting with the case when $\{y < 0\}$, we obtain

$$\begin{aligned}
&\mathbb{P}_{\mu_0}\left((y-\mu_0)^2 - y^2 \leq c \mid y < 0\right) \qquad (2.64)\\
&= \mathbb{P}_{\mu_0}\left(-2y\mu_0 + \mu_0^2 \leq c \mid y < 0\right)\\
&= \mathbb{P}_{\mu_0}\left(y \geq \frac{\mu_0^2 - c}{2\mu_0} \mid y < 0\right)\\
&= \Phi(-\mu_0)^{-1}\mathbb{P}_{\mu_0}\left(y \geq \frac{\mu_0^2 - c}{2\mu_0}, y < 0\right)\\
&= \Phi(-\mu_0)^{-1}\left\{0 \cdot \mathbf{1}\{c \leq \mu_0^2\} + \mathbb{P}_{\mu_0}\left(\frac{\mu_0^2 - c}{2\mu_0} \leq y \leq 0\right)\mathbf{1}\{c > \mu_0^2\}\right\}\\
&= \Phi(-\mu_0)^{-1}\mathbb{P}_{\mu_0}\left(\frac{-\mu_0^2 - c}{2\mu_0} \leq y - \mu_0 \leq -\mu_0\right)\mathbf{1}\{c > \mu_0^2\}\\
&= \Phi(-\mu_0)^{-1}\left\{\Phi(-\mu_0) - \Phi\left(\frac{-\mu_0^2 - c}{2\mu_0}\right)\right\}\mathbf{1}\{c > \mu_0^2\}. \qquad (2.65)
\end{aligned}$$

Then, when $\{y \geq 0\}$, we have

$$\begin{aligned}
&\mathbb{P}_{\mu_0}\left((y-\mu_0)^2 \leq c \mid y \geq 0\right) \qquad (2.66)\\
&= \mathbb{P}_{\mu_0}\left(-\sqrt{c} \leq y - \mu_0 \leq \sqrt{c} \mid y \geq 0\right)\\
&= \Phi(\mu_0)^{-1}\mathbb{P}_{\mu_0}\left(-\sqrt{c} \leq y - \mu_0 \leq \sqrt{c}, y \geq 0\right)\\
&= \Phi(\mu_0)^{-1}\mathbb{P}_{\mu_0}\left(0 \leq y \leq \sqrt{c} + \mu_0\right)\mathbf{1}\{-\sqrt{c} + \mu_0 \leq 0\}\\
&\quad + \Phi(\mu_0)^{-1}\mathbb{P}_{\mu_0}\left(-\sqrt{c} + \mu_0 \leq y \leq \sqrt{c} + \mu_0\right)\mathbf{1}\{-\sqrt{c} + \mu_0 > 0\}\\
&= \Phi(\mu_0)^{-1}\left\{\left(\Phi(\sqrt{c}) - \Phi(-\mu_0)\right)\mathbf{1}\{c \geq \mu_0^2\} + (2\Phi(\sqrt{c}) - 1)\mathbf{1}\{c < \mu_0^2\}\right\}. \qquad (2.67)
\end{aligned}$$

Putting together (2.65) and (2.67), we obtain the following CDF:

$$\mathbb{P}_{\mu_0}(\ell_0 \leq c) = \chi_1^2(c) \cdot \mathbf{1}\{c < \mu_0^2\} + \left\{\Phi(\sqrt{c}) - \Phi\left(\frac{-\mu_0^2 - c}{2\mu_0}\right)\right\} \cdot \mathbf{1}\{c \geq \mu_0^2\}. \quad (2.68)$$

This completes the case of $\mu^* > 0$.

Proof of Example 2.2.7

We derive this result using a duality argument inspired by (Gouriéroux, Holly, and Monfort, 1982). By definition, we have

$$\lambda(\mu^*, \boldsymbol{y}) = \min_{\boldsymbol{x}:\varphi(\boldsymbol{x})=\mu^*} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2. \tag{2.69}$$

For ease of notation, let $\widehat{\boldsymbol{x}}^* = \underset{\boldsymbol{x}\,:\,\boldsymbol{h}^\top\boldsymbol{x}=\mu^*}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2$. Consider the Lagrangian for the first optimization in (2.69):

$$L(\boldsymbol{x}, \lambda) = \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 + \lambda(\boldsymbol{h}^\top\boldsymbol{x} - \mu^*). \tag{2.70}$$

First-order optimality allows solving for $\widehat{\boldsymbol{x}}^*$ as a function of the dual variable λ:

$$\begin{aligned}
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = -2\boldsymbol{K}^\top(\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}) + \lambda\boldsymbol{h} &= 0 \\
\implies -2\boldsymbol{K}^\top\boldsymbol{y} + 2\boldsymbol{K}^\top\boldsymbol{K}\boldsymbol{x} + \lambda\boldsymbol{h} &= 0 \\
\implies \widehat{\boldsymbol{x}}^* = (\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{K}^\top\boldsymbol{y} - \frac{1}{2}\lambda(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h} \\
\implies \widehat{\boldsymbol{x}}^* = \widehat{\boldsymbol{x}} - \frac{1}{2}\lambda(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h}.
\end{aligned}$$

Substituting back into the LLR, we obtain

$$\begin{aligned}
\lambda(\mu^*, \boldsymbol{y}) &= \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}}^*\|_2^2 - \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}}\|_2^2 \\
&= \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}} + \frac{1}{2}\lambda\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h}\|_2^2 - \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}}\|_2^2.
\end{aligned} \tag{2.71}$$

Performing some algebra, we note that

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}} + \frac{1}{2}\lambda\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}}\|_2^2 \\
+ \lambda(\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}})^\top\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h} + \frac{1}{4}\lambda^2\boldsymbol{h}^\top(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\lambda(\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}})^\top\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h} &= \lambda\boldsymbol{y}^\top\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h} - \lambda\widehat{\boldsymbol{x}}^\top\boldsymbol{K}^\top\boldsymbol{K}(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h} \\
&= \lambda\widehat{\boldsymbol{x}}^\top\boldsymbol{h} - \lambda\widehat{\boldsymbol{x}}^\top\boldsymbol{h} \\
&= 0.
\end{aligned}$$

So the substitution in (2.71) can be further simplified such that:

$$\lambda(\mu^*, \boldsymbol{y}) = \frac{1}{4}\lambda^2\boldsymbol{h}^\top(\boldsymbol{K}^\top\boldsymbol{K})^{-1}\boldsymbol{h}. \tag{2.72}$$

We now turn our attention to finding λ. Note that this optimization defining the Lagrangian (2.70) is convex with an affine equality constraint. Therefore, strong duality holds. We then define the dual function as follows:

$$\begin{aligned} g(\lambda) &= \min_{\boldsymbol{x}} L(\boldsymbol{x}, \lambda) = L(\widehat{\boldsymbol{x}}^*, \lambda) \\ &= \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}}^*\|_2^2 + \lambda(\boldsymbol{h}^\top \widehat{\boldsymbol{x}}^* - \mu^*) \\ &= \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}} + \frac{1}{2}\lambda \boldsymbol{K}(\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h}\|_2^2 + \lambda \left(\boldsymbol{h}^\top \widehat{\boldsymbol{x}} - \frac{1}{2}\lambda \boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h} - \mu^* \right). \end{aligned} \tag{2.73}$$

We note that we can make many of the same simplifications above to arrive at the simplified dual function:

$$g(\lambda) = \|\boldsymbol{y} - \boldsymbol{K}\widehat{\boldsymbol{x}}\|_2^2 + \lambda \boldsymbol{h}^\top - \frac{1}{4}\lambda^2 \boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h} + \lambda \boldsymbol{h}^\top \widehat{\boldsymbol{x}} - \lambda \mu^*. \tag{2.74}$$

To maximize $g(\lambda)$, we again use the following first order optimality condition:

$$\begin{aligned} \frac{dg}{d\lambda} &= -\frac{1}{2}\lambda \boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h} + \boldsymbol{h}^\top \widehat{\boldsymbol{x}} - \mu^* = 0 \\ \implies \widehat{\lambda} &= \frac{2\left(\boldsymbol{h}^\top \widehat{\boldsymbol{x}} - \mu^*\right)}{\boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h}}. \end{aligned} \tag{2.75}$$

Substituting (2.75) back into (2.72), we obtain

$$\begin{aligned} \lambda(\mu^*, \boldsymbol{y}) &= \frac{1}{4}\left(\frac{2\left(\boldsymbol{h}^\top \widehat{\boldsymbol{x}} - \mu^*\right)}{\boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h}} \right)^2 \boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h} \\ &= \frac{\left(\boldsymbol{h}^\top \widehat{\boldsymbol{x}} - \mu^*\right)^2}{\boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h}}. \end{aligned}$$

For the second part, observe that when $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}^*, \boldsymbol{I}_m)$, we have

$$\boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{K}^\top \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{h}^\top \boldsymbol{x}^*, \boldsymbol{h}^\top (\boldsymbol{K}^\top \boldsymbol{K})^{-1}\boldsymbol{h}),$$

hence (2.33) is the square of a one-dimensional standard Gaussian distribution. This finishes the proof.

2.8 Proofs in Section 2.3

Proof of Lemma 2.3.2

Let $Y := \lambda(\boldsymbol{y}, \mu^*)$. Recall the validity of q_α can be written as $\mathbb{P}(Y \le q_\alpha) \ge 1 - \alpha$ from (2.14) as:

$$\begin{aligned} X \succeq Y &\iff \mathbb{P}(X \ge \gamma) \ge \mathbb{P}(Y \ge \gamma), \quad \text{for all } \gamma \\ &\iff \alpha = \mathbb{P}(X \ge Q_X(1-\alpha)) \ge \mathbb{P}(Y \ge Q_X(1-\alpha)), \quad \text{for all } \alpha \\ &\iff 1 - \alpha = \mathbb{P}(X \le Q_X(1-\alpha)) \le \mathbb{P}(Y \le Q_X(1-\alpha)), \quad \text{for all } \alpha \\ &\iff Q_X(1-\alpha) \text{ is a valid } q_\alpha \text{for all } \alpha. \end{aligned}$$

This finishes the proof.

Joint formulation of change-constrained optimization problem (2.36)

In this section, we provide details on formulating the optimization problem (2.36) and the interval optimization problem (2.24) as a single chance-constrained optimization problem. By joining (2.36) and (2.24) as a single optimization problem, we can use a similar argument as in the proof of Theorem 2.2.4. Starting with problem (2.54):

$$\begin{aligned} \inf_{\mu}/\sup_{\mu} \quad & \mu \\ \text{st} \quad & \mu \in \mathbb{R} \\ & \inf_{\varphi(\boldsymbol{x})=\mu, \boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) - \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) \leq q_\alpha(\mu), \end{aligned} \tag{2.76}$$

and substituting $q_\alpha(\mu) = \sup_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} Q_{F_{\boldsymbol{x}}}(1-\alpha) = -\inf_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} -Q_{F_{\boldsymbol{x}}}(1-\alpha)$, we obtain:

$$\begin{aligned} \inf_{\mu}/\sup_{\mu} \quad & \mu \\ \text{st} \quad & \mu \in \mathbb{R} \\ & \inf_{\substack{\varphi(\boldsymbol{x}_1)=\mu, \boldsymbol{x}_1\in\mathcal{X}, \\ \varphi(\boldsymbol{x}_2)=\mu, \boldsymbol{x}_2\in\mathcal{X}}} \left[-2\ell_{\boldsymbol{x}_1}(\boldsymbol{y}) - Q_{F_{\boldsymbol{x}_2}}(1-\alpha)\right] \leq \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}), \end{aligned} \tag{2.77}$$

which can be transformed to parameter space as:

$$\begin{aligned} \inf_{\boldsymbol{x}_1,\boldsymbol{x}_2}/\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2} \quad & \varphi(\boldsymbol{x}_1) \\ \text{st} \quad & \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X} \\ & \varphi(\boldsymbol{x}_1) = \varphi(\boldsymbol{x}_2) \\ & -2\ell_{\boldsymbol{x}_1}(\boldsymbol{y}) - Q_{F_{\boldsymbol{x}_2}}(1-\alpha) \leq \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}). \end{aligned} \tag{2.78}$$

Further unpacking $Q_{F_{\boldsymbol{x}_2}}(1-\alpha)$ as in Lemma 2.3.1, we obtain the chance-constrained optimization problem:

$$\begin{aligned} \inf_{\boldsymbol{x}_1,\boldsymbol{x}_2,q}/\sup_{\boldsymbol{x}_1,\boldsymbol{x}_2,q} \quad & \varphi(\boldsymbol{x}_1) \\ \text{st} \quad & \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X} \\ & \varphi(\boldsymbol{x}_1) = \varphi(\boldsymbol{x}_2) \\ & -2\ell_{\boldsymbol{x}_1}(\boldsymbol{y}) \leq \inf_{\boldsymbol{x}\in\mathcal{X}} -2\ell_{\boldsymbol{x}}(\boldsymbol{y}) + q \\ & \mathbb{P}_{u\sim\mathcal{U}([0,1])}(\mathcal{F}(\boldsymbol{x}_2, u) \leq q) \leq 1-\alpha, \end{aligned} \tag{2.79}$$

where $\mathcal{F}(\boldsymbol{x}, u) = F_{\boldsymbol{x}}^{-1}(u)$.

Proof of Example 2.3.3

Similar to Example 2.2.6 in Section 2.7, this proof is divided into two cases.

Case of $\mu^* = 0$ Since the case when $\mu^* = 0$ is of particular interest, we show the result in this specific case and then generalize. Thus, when $\mu_0 = 0$, the log-likelihood ratio has the following distribution:

$$\ell_0 \sim \frac{1}{2} + \frac{1}{2}\chi_1^2. \tag{2.80}$$

Additionally, this distribution implies the following stochastic dominance:

$$\mathbb{P}_{\mu_0}(\ell_0 \leq c) = \frac{1}{2}\left(1 + \chi_1^2(c)\right) \geq \chi_1^2(c), \tag{2.81}$$

that is, the log-likelihood ratio CDF is stochastically dominated by the chi-squared with one degree of freedom distribution. This means that the type-I error of the test can be controlled at the α level.

When $\mu > 0$, the closed-form solution to the CDF of ℓ_0 becomes more complicated, as we can no longer use symmetry around the origin. From the result of Example 2.2.6, we have the following CDF:

$$\mathbb{P}_{\mu_0}(\ell_0 \leq c) = \chi_1^2(c) \cdot \mathbf{1}\{c < \mu_0^2\} + \left\{\Phi(\sqrt{c}) - \Phi\left(\frac{-\mu_0^2 - c}{2\mu_0}\right)\right\} \cdot \mathbf{1}\{c \geq \mu_0^2\}. \tag{2.82}$$

Note, a quick check of (2.82) when $\mu_0 = 0$ reveals agreement with (2.81) such that

$$\mathbb{P}_{\mu_0}(\ell_0 \leq c) = \Phi(\sqrt{c}) = \Phi(\sqrt{c}) - \frac{1}{2} + \frac{1}{2} = \frac{1}{2}\left(2\Phi(\sqrt{c}) - 1\right) + \frac{1}{2} = \frac{1}{2}\chi_1^2(c) + \frac{1}{2}. \tag{2.83}$$

This completes the case of $\mu^* = 0$.

Case of $\mu^* > 0$ We already demonstrated above the chi-squared with one degree of freedom dominates the log-likelihood ratio when $\mu_0 = 0$. We now show that the dominance holds when $\mu_0 > 0$. Clearly, when $c < \mu_0^2$, $\mathbb{P}_{\mu_0}(\ell_0 \leq 0) = \chi_1^2(c)$, making it in fact equal to the chi-squared with one degree of freedom. Suppose $c \geq \mu_0^2$. Define

$$h(c) := \Phi(\sqrt{c}) - \Phi\left(\frac{-\mu_0^2 - c}{2\mu_0}\right) - \chi_1^2(c).$$

The stochastic dominance occurs if and only if $h(c) \geq 0$ for all $c \geq \mu_0^2$.

Note first that $\chi_1^2(c) = \Phi(\sqrt{c}) - \Phi(-\sqrt{c})$ and therefore $h(c) = \Phi(-\sqrt{c}) - \Phi\left(\frac{-\mu_0^2 - c}{2\mu_0}\right)$. Since $\Phi(\cdot)$ is a monotonically increasing function, it is sufficient to show that $-\sqrt{c} - \frac{-\mu_0^2 - c}{2\mu_0} \geq 0$ for all $c \geq \mu_0^2$. We do so below.

Define a function f as follows:

$$f(c) = -\sqrt{c} - \frac{-\mu_0^2 - c}{2\mu_0}.$$

Observe that when $c = \mu_0^2$, $f(c) = 0$. Consider when $c > \mu_0^2$. We obtain the following first and second derivatives:

$$f'(c) = \frac{-\mu_0 + \sqrt{c}}{2\mu_0\sqrt{c}} \quad \text{and} \quad f''(c) = \frac{1}{4}c^{-3/2}.$$

By the constraint $c > \mu_0^2$, it follows that $-\mu_0 + \sqrt{c} > 0$, and therefore, $f'(c) > 0$ for all $c > \mu_0^2$. Additionally, $f''(c) > 0$ for all $c > \mu_0^2$, so f is convex. Hence, we conclude that f is a monotonically increasing function for $c > \mu_0^2$, which starts at 0 when $c = \mu_0^2$, and thus $f(c) \geq 0$ for all $c \geq \mu_0^2$. It therefore follows that

$$\Phi(-\sqrt{c}) \geq \Phi\left(\frac{-\mu_0^2 - c}{2\mu_0}\right),$$

and hence $h(c) \geq 0$ for all $c \geq \mu_0^2$. As such, we conclude that $\mathbb{P}_{\mu_0}(\ell_0 \leq c) \geq \chi_1^2(c)$ for all $c \geq 0$. In other words, that the sampling distribution for the log-likelihood ratio is stochastically dominated by a chi-squared distribution with one degree of freedom. This completes the case of $\mu^* > 0$.

2.9 Proofs in Section 2.4

Proof of Theorem 2.4.1

The proof follows by combining Lemmas 2.4.2 and 2.4.3.

Proof of Lemma 2.4.2

The proof follows by direct inspection and substitution of q_α and $-2\ell_{\boldsymbol{x}}(\boldsymbol{y}) = \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2$. The interval has the coverage if the q_α is valid by Theorem 2.2.4 and only if by Proposition 2.2.5.

Proof of Lemma 2.4.3

The proof follows by observing $z_{\alpha/2}^2 = Q_{\chi_1^2}(1-\alpha)$ and applying Lemma 2.3.2

Proof of Lemma 2.4.4

We argue by coupling. Note that $\frac{1}{2}(y_1 - y_2)^2 \sim \chi^2_1$, so that it suffices to show $\lambda \leq \frac{1}{2}(y_1 - y_2)^2$ for every y to constitute a valid coupling that proves stochastic dominance. This is clearly true when $y_1 + y_2 \geq 0$, since $\lambda - \frac{1}{2}(y_1 - y_2)^2$ is equal to non-positive terms only. When $y_1 + y_2 < 0$, if both are strictly negative, then $\lambda = 0 \leq \frac{1}{2}(y_1 - y_2)^2$. Then assume without loss of generality that y_1 is non-negative, then y_2 has to be negative. Then $\lambda = y_1^2$, but $y_1 \geq 0$, $y_2 < 0$ and $y_1 < -y_2$ imply that $|y_1 - y_2| = y_1 - y_2 \geq 2y_1 \geq \sqrt{2}y_1$, squaring both sides gives $\frac{1}{2}(y_1 - y_2)^2 < y_1^2 = \lambda$. This finishes the proof.

Proof of Lemma 2.4.5

Consider the LLR

$$\lambda(\mu^* = -1, \boldsymbol{y}) = \min_{\substack{x_1 + x_2 - x_3 = -1 \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2. \tag{2.84}$$

The goal of this proof is to show that χ^2_1 does not stochastically dominate (2.84) when $\boldsymbol{y} \sim \mathcal{N}(x^* = (0, 0, 1), \boldsymbol{I}_3)$. By Corollary 4.26 in (Roch, 2024), $X \succeq Y$ implies $\mathbb{E}[x] > \mathbb{E}[y]$, so it suffices to show that

$$\mathbb{E}[\lambda(\mu^* = -1, \boldsymbol{y})] > \mathbb{E}[\chi^2_1] = 1$$

to complete the proof.

Observe that

$$\mathbb{E}[\lambda(\mu^* = -1, \boldsymbol{y})] = \mathbb{E}\left[\min_{\substack{\boldsymbol{h}^\top \boldsymbol{x} = -1 \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right] - \mathbb{E}\left[\min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right]. \tag{2.85}$$

We begin by computing the second term. Since

$$\min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 = \sum_{i=1}^{3} (y_i - \max\{y_i, 0\})^2,$$

we have

$$\begin{aligned}\mathbb{E}\left[\min_{\boldsymbol{x} \geq \boldsymbol{0}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2\right] &= \sum_{i=1}^{3} \mathbb{E}\left[(y_i - \max\{y_i, 0\})^2\right] \\ &= 2\mathbb{E}_{z \sim \mathcal{N}(0,1)}\left[(z - \max\{z, 0\})^2\right] + \mathbb{E}_{z \sim \mathcal{N}(1,1)}\left[(z - \max\{z, 0\})^2\right].\end{aligned}$$

Let $g(z) := (z - \max\{z, 0\})^2$. Using in both cases, we obtain

$$\mathbb{E}[g(z)] = \underbrace{\mathbb{E}[g(z) \mid z \geq 0]}_{0} \cdot \mathbb{P}(z \geq 0) + \mathbb{E}[g(z) \mid z < 0] \cdot \mathbb{P}(z < 0) = \mathbb{E}[z^2 \mid z < 0] \cdot \mathbb{P}(z < 0).$$

Note that

$$\begin{aligned}\mathbb{E}[z^2 \mid z<0] &= (\mathbb{E}[z \mid z<0])^2 + \mathrm{Var}[z \mid z<0] \\ &= \begin{cases} \left(-\frac{\phi(0)}{\Phi(0)}\right)^2 + \left(1 - \left(\frac{\phi(0)}{\Phi(0)}\right)^2\right), & z \sim \mathcal{N}(0,1) \\ \left(1 - \frac{\phi(-1)}{\Phi(-1)}\right)^2 + 1 + \frac{\phi(-1)}{\Phi(-1)} - \left(\frac{\phi(-1)}{\Phi(-1)}\right)^2, & z \sim \mathcal{N}(1,1) \end{cases} \\ &= \begin{cases} 1, & z \sim \mathcal{N}(0,1) \\ 2 - \frac{\phi(-1)}{\Phi(-1)}, & z \sim \mathcal{N}(1,1), \end{cases}\end{aligned}$$

where we used the formulas for mean and variance of a truncated Gaussian. Finally,

$$\begin{aligned}\mathbb{E}\left[\min_{\boldsymbol{x}\geq\mathbf{0}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right] &= 2 \cdot 1/2 \cdot 1 + (2 - \phi(-1)/\Phi(-1)) \cdot (\Phi(-1)) \\ &= 1 + 2\Phi(-1) - \phi(-1) \\ &\approx 1.0753.\end{aligned}$$

It suffices to prove that

$$\mathbb{E}\left[\min_{\substack{\boldsymbol{h}^\top\boldsymbol{x}=-1 \\ \boldsymbol{x}\geq\mathbf{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right] > 2 + 2\Phi(-1) - \phi(-1) \approx 2.0753.$$

We will prove that

$$\mathbb{E}\left[\min_{\substack{\boldsymbol{h}^\top\boldsymbol{x}=-1 \\ \boldsymbol{x}\geq\mathbf{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right] = 13/6 \approx 2.166.$$

Note that the intersection of the plane $\boldsymbol{h}^\top\boldsymbol{x} = x_1 + x_2 - x_3 = -1$ and $\boldsymbol{x} \geq \mathbf{0}$ is the parametric surface $\mathcal{S} = \{(u,\ v,\ u+v+1), u \geq 0, v \geq 0\}$, so we can write

$$\min_{\boldsymbol{x}\in\mathcal{S}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = \min_{u\geq 0, v\geq 0} (y_1-u)^2 + (y_2-v)^2 + (y_3-u-v-1)^2. \tag{2.86}$$

It is convenient to define a new variable $z_3 = 1 - y_3 \sim \mathcal{N}(0,1)$, so that (y_1, y_2, z_3) is sampled from a standard three dimensional Gaussian. Abusing notation we will still write y_3 for z_3 and then $\boldsymbol{y} \sim \mathcal{N}((0,0,0), \boldsymbol{I})$. The optimization problem becomes:

$$\min_{u\geq 0, v\geq 0} (y_1-u)^2 + (y_2-v)^2 + (-y_3-u-v)^2. \tag{2.87}$$

This can be explicitly solved to yield

$$\min_{\substack{\boldsymbol{h}^\top \boldsymbol{x}=-1 \\ \boldsymbol{x}\geq \boldsymbol{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 = \begin{cases} y_1^2+y_2^2+y_3^2 & y_1-y_3\leq 0 \text{ and } y_2-y_3\leq 0 \\ \frac{1}{2}\left(y_1^2+2y_1y_3+2y_2^2+y_3^2\right) & y_1-y_3\geq 0 \text{ and } y_1-2y_2+y_3\geq 0 \\ \frac{1}{2}\left(2y_1^2+y_2^2+2y_2y_3+y_3^2\right) & y_2-y_3\geq 0 \text{ and } 2y_1-y_2-y_3\leq 0 \\ \frac{1}{3}\left(y_1+y_2+y_3\right)^2 & \begin{cases} 2y_1-y_3\geq y_2\geq \max\{y_1,y_3\} \\ 2y_2-y_3\geq y_1\geq \max\{y_2,y_3\} \end{cases} \end{cases}. \tag{2.88}$$

We split

$$\int_{\mathbb{R}^3} \min_{\substack{\boldsymbol{h}^\top \boldsymbol{x}=-1 \\ \boldsymbol{x}\geq \boldsymbol{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \, \phi(y_1)\phi(y_2)\phi(y_3)\,\mathrm{d}y$$

into the different domains given by (2.88), with the value of the expectation being equal to the sum of the different integrals, which we proceed to compute.

Region 1:

$$I_1 = \int_{y_3\geq y_1, y_3\geq y_2} \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{2\sqrt{2}\pi^{3/2}}(y_1^2+y_2^2+y_3^2)\,\mathrm{d}y.$$

Note that by symmetry of the variables in the integrand, we have

$$\begin{aligned} I_1 &= \int_{y_2\geq y_1, y_2\geq y_3} \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{2\sqrt{2}\pi^{3/2}}(y_1^2+y_2^2+y_3^2)\,\mathrm{d}y \\ &= \int_{y_1\geq y_3, y_1\geq y_2} \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{2\sqrt{2}\pi^{3/2}}(y_1^2+y_2^2+y_3^2)\,\mathrm{d}y. \end{aligned}$$

And since one of the y_i will always be the largest one, the sum of the domains is $\mathbb{R}^3$ (modulo measure zero intersections that do not affect integration) and we can write

$$I_1 = \frac{1}{3}\int_{\mathbb{R}^3} \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{2\sqrt{2}\pi^{3/2}}(y_1^2+y_2^2+y_3^2)\,\mathrm{d}y = \frac{1}{3}\cdot 3 = 1.$$

Here we used that the integral is the expected value of $y_1^2+y_2^2+y_3^2$, which is 3 since the y_i are centered with unit variance.

Region 2:

$$I_2 = \int_{y_1-y_3\geq 0, y_1-2y_2+y_3\geq 0} \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{4\sqrt{2}\pi^{3/2}}\left(y_1^2+2y_1y_3+2y_2^2+y_3^2\right)\,\mathrm{d}y. \tag{2.89}$$

Partition $\mathbb{R}^3$ in four spaces with measure zero intersection, and we aim to argue that the integral of the integrand in (2.89) has the same value when integrating over any of them:

$$\begin{aligned} A &:= \left\{ \boldsymbol{y} : y_1 \geq y_3, y_2 \geq \frac{y_1 + y_3}{2} \right\} \\ B &:= \left\{ \boldsymbol{y} : y_1 \geq y_3, y_2 \leq \frac{y_1 + y_3}{2} \right\} \\ C &:= \left\{ \boldsymbol{y} : y_1 \leq y_3, y_2 \geq \frac{y_1 + y_3}{2} \right\} \\ D &:= \left\{ \boldsymbol{y} : y_1 \leq y_3, y_2 \leq \frac{y_1 + y_3}{2} \right\}. \end{aligned}$$

Clearly $I_2 = I_B = \int_A \boldsymbol{h}(y_1, y_2, y_3)\,\mathrm{d}y$. Since $\boldsymbol{h}$ satisfies $\boldsymbol{h}(x_1, x_2, x_3) = \boldsymbol{h}(x_3, x_2, x_1)$, we can exchange y_1 and y_3 in the definitions of the sets, so $I_A = I_C$ and $I_B = I_D$. And since $\boldsymbol{h}$ is even with respect to x_2 and odd with respect to x_1, x_3 we can exchange y_i to $-y_i$ for $i = 1, 2, 3$ without the result changing. This flips both inequalities, proving $I_A = I_D$ and $I_B = I_C$. We therefore have

$$I_2 = \frac{1}{4} \int_{\mathbb{R}^3} \frac{e^{-\frac{1}{2}(y_1^2 + y_2^2 + y_3^2)}}{4\sqrt{2}\pi^{3/2}} \left(y_1^2 + 2y_1 y_3 + 2y_2^2 + y_3^2 \right) \mathrm{d}y = \frac{1}{4} \cdot 2 = \frac{1}{2}.$$

Here, in the integral, we factor out the sum and using that, the expected value of $y_i y_j$ is δ_{ij}.

Region 3:

$$I_3 = \int_{y_2 - y_3 \geq 0, y_2 - 2y_1 + y_3 \geq 0} \frac{e^{-\frac{1}{2}(y_1^2 + y_2^2 + y_3^2)}}{4\sqrt{2}\pi^{3/2}} \left(2y_1^2 + 2y_2 y_3 + y_2^2 + y_3^2 \right) \mathrm{d}y.$$

This is exactly the same integral as I_2 by switching y_2 with y_1, so $I_3 = \dfrac{1}{2}$.

Region 4:

$$\begin{aligned} I_4 = &\int_{2y_1 - y_3 \geq y_2 \geq \max(y_1, y_3)} \frac{e^{-\frac{1}{2}(y_1^2 + y_2^2 + y_3^2)}}{6\sqrt{2}\pi^{3/2}} (y_1 + y_2 + y_3)^2 \,\mathrm{d}y \\ &+ \int_{2y_2 - y_3 \geq y_1 \geq \max(y_2, y_3)} \frac{e^{-\frac{1}{2}(y_1^2 + y_2^2 + y_3^2)}}{6\sqrt{2}\pi^{3/2}} (y_1 + y_2 + y_3)^2 \,\mathrm{d}y. \end{aligned} \tag{2.90}$$

We partition $\mathbb{R}^3$ in 12 subspaces with measure 0 intersection and we aim to argue that the integral of the integrand in (2.90) (considering one of the integrals only) has the

same value when integrating over any of them. For σ a permutation of (y_1, y_2, y_3), we define the first six subsets as:

$$\{\boldsymbol{y} : 2y_{\sigma(1)} - y_{\sigma(2)} \geq y_{\sigma(3)} \geq \max\{y_{\sigma(1)}, y_{\sigma(2)}\}\},$$

and the last six subsets as:

$$\{\boldsymbol{y} : 2y_{\sigma(1)} - y_{\sigma(2)} \leq y_{\sigma(3)} \leq \min\{y_{\sigma(1)}, y_{\sigma(2)}\}\}.$$

We need to prove that the integral has the same value in any of the 12 subsets. Since that the integrand

$$\boldsymbol{h}(y_1, y_2, y_3) := \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{6\sqrt{2}\pi^{3/2}}(y_1 + y_2 + y_3)^2$$

satisfies $\boldsymbol{h}(y_1, y_2, y_3) = \boldsymbol{h}(y_{\sigma(1)}, y_{\sigma(2)}, y_{\sigma(3)})$ for all permutations σ, the value of the integral in between the first and second groups of 6 subsets is the same. For a fixed σ (say, the identity), since $\boldsymbol{h}(y_1, y_2, y_3) = \boldsymbol{h}(-y_1, -y_2, -y_3)$, the value over

$$\{\boldsymbol{y} : 2y_1 - y_2 \geq y_3 \geq \max\{y_1, y_2\}\}$$

is the same as the value over

$$\{\boldsymbol{y} : -2y_1 + y_2 \geq -y_3 \geq \max\{-y_1, -y_2\}\} = \{\boldsymbol{y} : 2y_1 - y_2 \leq y_3 \leq \min\{y_1, y_2\}\},$$

so the value over the 12 sets is complete.

It remains to be seen that for a generic $\boldsymbol{y} = (y_1, y_2, y_3)$, $y_1 \neq y_2 \neq y_3 \neq y_1$ (which can be assumed with probability 1 without affecting the integral), the point belongs to one and just one of the sets. Assume without loss of generality that y_1 is the greater of the three and y_3 is the smallest. Then since $y_1 > \max\{y_2, y_3\}$ and $y_3 < \min\{y_1, y_2\}$ the only subsets that $\boldsymbol{y}$ can belong to are:

$$\begin{aligned}
A &:= \{\boldsymbol{y} : 2y_2 - y_3 \geq y_1 \geq \max\{y_2, y_3\}\} \\
B &:= \{\boldsymbol{y} : 2y_3 - y_2 \geq y_1 \geq \max\{y_2, y_3\}\} \\
C &:= \{\boldsymbol{y} : 2y_1 - y_2 \leq y_3 \leq \min\{y_1, y_2\}\} \\
D &:= \{\boldsymbol{y} : 2y_2 - y_1 \leq y_3 \leq \min\{y_1, y_2\}\}.
\end{aligned}$$

But $\boldsymbol{y}$ is not in B because that would require $y_3 \geq \frac{y_1+y_2}{2}$ but $y_3 < y_1$ and $y_3 < y_2$, and it is also not in C because that would require $y_1 \leq \frac{y_2+y_3}{2}$ and $y_1 > y_2$ and $y_1 > y_3$. $\boldsymbol{y}$ will be in A if $y_2 > \frac{y_1+y_3}{2}$ and in D if, on the contrary, $y_2 < \frac{y_1+y_3}{2}$, both of which

are possible, but not at the same time. We conclude by identifying I_4 as the sum of two integrals over subsets that we have defined, and therefore

$$I_4 = \frac{2}{12}\int_{\mathbb{R}^3} \frac{e^{-\frac{1}{2}(y_1^2+y_2^2+y_3^2)}}{6\sqrt{2}\pi^{3/2}}(y_1+y_2+y_3)^2\,\mathrm{d}y = \frac{1}{6}\cdot 1 = \frac{1}{6}.$$

Here we expand the sum and use again that the expected value of $y_i y_j$ is δ_{ij}. The proof concludes by adding up

$$\mathbb{E}\left[\min_{\substack{\boldsymbol{h}^\top \boldsymbol{x}=-1 \\ \boldsymbol{x}\geq \boldsymbol{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right] = I_1 + I_2 + I_3 + I_4 = \frac{13}{6}.$$

Proof of Proposition 2.4.6

We construct a series of counterexamples, indexed by the dimension p, and prove that as $p \to \infty$, the expected value of the LLR diverges. Since stochastic dominance implies inequality of expectations (when expectations are finite), we conclude that the distribution can not be stochastically dominated. For all $p \in \mathbb{N}$, consider the example in $\mathbb{R}^p (= \mathbb{R}^m)$, $\boldsymbol{K} = \boldsymbol{I}_p$, $\boldsymbol{x}^* = (0,\ldots,0,1)$, $\boldsymbol{h} = (1,\ldots,1,-1)$ (such that $\mu^* = -1$). Let

$$\lambda_n(\mu^* = -1, \boldsymbol{y}) = \min_{\substack{\sum_{i=1}^{p-1} x_i - x_p = -1 \\ \boldsymbol{x}\geq \boldsymbol{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - \min_{\boldsymbol{x}\geq \boldsymbol{0}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2. \tag{2.91}$$

And compute

$$\mathbb{E}_{\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{x}^*,\boldsymbol{I}_n)}[\lambda_n(-1,\boldsymbol{y})] = \mathbb{E}_{\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{x}^*,\boldsymbol{I}_n)}\left[\min_{\substack{\sum_{i=1}^{p-1} x_i - x_p = -1 \\ \boldsymbol{x}\geq \boldsymbol{0}}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right] - \mathbb{E}_{\boldsymbol{y}\sim\mathcal{N}(\boldsymbol{x}^*,\boldsymbol{I}_n)}\left[\min_{\boldsymbol{x}\geq \boldsymbol{0}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right]. \tag{2.92}$$

For the second term, we have

$$\begin{aligned}
\mathbb{E}\left[\min_{\boldsymbol{x}\geq \boldsymbol{0}} \|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right] &= \sum_{i=1}^{p} \mathbb{E}\left[(y_i - \max\{y_i, 0\})^2\right] \\
&= (p-1)\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[(z-\max\{z,0\})^2\right] + \mathbb{E}_{z\sim\mathcal{N}(1,1)}\left[(z-\max\{z,0\})^2\right] \\
&= (p-1)\frac{1}{2} + (2 - \phi(-1)/\Phi(-1))\cdot(\Phi(-1)),
\end{aligned}$$

using similar arguments as the proof in Section 2.9. We will lower bound the first term using duality. For simplicity, define $\boldsymbol{z} = (y_1,\ldots,y_{p-1}, y_n - 1) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, and equivalently optimize

$$\min_{\substack{\sum_{i=1}^{p-1} x_i = x_p \\ \boldsymbol{x}\geq \boldsymbol{0}}} \|\boldsymbol{x}-\boldsymbol{z}\|_2^2, \tag{2.93}$$

where we defined the feasible $\widetilde{\boldsymbol{x}} = (x_1, \ldots, x_n - 1 = \sum_{i=1}^{p-1} x_i)$ and replaced $\widetilde{\boldsymbol{x}}$ by $\boldsymbol{x}$, abusing notation. Using Fenchel duality, we have that

$$\min_{\substack{\sum_{i=1}^{p-1} x_i = x_p \\ \boldsymbol{x} \geq \mathbf{0}}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 \geq \sup_{\boldsymbol{\xi} \in \mathbb{R}^p} (-f^*(\boldsymbol{\xi}) - g^*(-\boldsymbol{\xi})), \tag{2.94}$$

where we have noted by f^* the convex conjugate of $f(\boldsymbol{x}) := \|\boldsymbol{x} - \boldsymbol{z}\|_2^2$ and, letting S be the feasible set, we denoted by g^* the convex conjugate of

$$g(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \in S \\ \infty & \text{if } \boldsymbol{x} \notin S. \end{cases} \tag{2.95}$$

Note that with these definitions, $\min_{\substack{\sum_{i=1}^{p-1} x_i = x_p \\ \boldsymbol{x} \geq \mathbf{0}}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 = \inf_{\boldsymbol{x}}(f(x) + g(x))$ so the weak Fenchel duality applies. We compute $f^*(\boldsymbol{\xi}) = \frac{1}{4}\|\boldsymbol{\xi}\|_2^2 + \boldsymbol{z}^\top \boldsymbol{\xi} - \boldsymbol{z}^\top \boldsymbol{z}$, and

$$g^*(\boldsymbol{\xi}) = \begin{cases} 0 & \text{if } \xi_i + \xi_p \leq 0 \text{ for } i \in [p-1] \\ \infty & \text{otherwise,} \end{cases} \tag{2.96}$$

so that

$$\sup_{\boldsymbol{\xi} \in \mathbb{R}^p} (-f^*(\boldsymbol{\xi}) - g^*(-\boldsymbol{\xi})) = \sup_{\xi_i + \xi_n \geq 0, \text{ for all } i \in [p-1]} \left[-\frac{1}{4}\|\boldsymbol{\xi}\|_2^2 - \boldsymbol{z}^\top \boldsymbol{\xi} + \boldsymbol{z}^\top \boldsymbol{z} \right] \tag{2.97}$$

$$\geq \sup_{\xi_i + \xi_n \geq 0, \text{ for all } i \in [p-1]} \left[-\frac{1}{4}\|\boldsymbol{\xi}\|_2^2 - \boldsymbol{z}^\top \boldsymbol{\xi} \right]. \tag{2.98}$$

Since the supremum is lower bounded by any feasible point, we can further bound by picking a feasible $\boldsymbol{\xi}^*$ for each possible $\boldsymbol{z}$. We define the following:

$$\boldsymbol{\xi}^*(\boldsymbol{z}) = \begin{cases} -\boldsymbol{z} & \text{if } -\boldsymbol{z} \text{ is feasible } (-z_i \geq z_n \text{ for all } i) \\ (-z_1, \ldots, -z_{p-1}, \max_{i \in [p-1]} z_i) & \text{otherwise.} \end{cases} \tag{2.99}$$

Observe that

$$\begin{aligned} &\min_{\substack{\sum_{i=1}^{p-1} x_i = x_p \\ \boldsymbol{x} \geq \mathbf{0}}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2 \\ &\quad \geq -\frac{1}{4}\|\boldsymbol{\xi}^*(\boldsymbol{z})\|_2^2 - \boldsymbol{z}^\top \boldsymbol{\xi}^*(\boldsymbol{z}) \\ &\quad = \begin{cases} \frac{3}{4}\|\boldsymbol{z}\|^2 & \text{if } -\boldsymbol{z} \text{ is feasible } (-z_i \geq z_n \text{ for all } i) \\ \frac{3}{4}\sum_{i=1}^{p-1} z_i^2 + z_n \max_{i \in [p-1]} z_i - \frac{1}{4}\left(\max_{i \in [p-1]} z_i\right)^2 & \text{otherwise.} \end{cases} \end{aligned}$$

We note that $-\boldsymbol{z}$ is feasible with probability $1/p$, by symmetry. Taking expected value over the inequality and using the law of total expectation yields

$$\begin{aligned}
&\mathbb{E}\left[\min_{\substack{\sum_{i=1}^{p-1} x_i = x_p \\ \boldsymbol{x} \geq \boldsymbol{0}}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2\right] \\
&\geq \frac{1}{p} \times \frac{3}{4}\mathbb{E}\big[\|z\|_2^2\big] + \frac{p-1}{p}\left\{\mathbb{E}\left[\frac{3}{4}\sum_{i=1}^{p-1} z_i^2\right] + \mathbb{E}\left[z_n \max_{i \in [p-1])} z_i\right] - \frac{1}{4}\mathbb{E}\left[\Big(\max_{i \in [p-1]} z_i\Big)^2\right]\right\} \\
&= \frac{3}{4} + \frac{p-1}{p}\left\{\frac{3(p-1)}{4} + 0 - \frac{1}{4}\mathbb{E}\left[\Big(\max_{i \in [p-1]} z_i\Big)^2\right]\right\}.
\end{aligned}$$

To bound the last term, we use

$$\mathbb{E}\left[\Big(\max_{i \in [p-1]} z_i\Big)^2\right] = \mathbb{E}\left[\max_{i \in [p-1]} z_i\right] + \mathrm{Var}\left[\max_{i \in [p-1]} z_i\right] \leq \sqrt{2\log(p-1)} + 1,$$

where the moment bounds are standard results: the expectation bound can be found using Jensen's inequality on $\exp(\sqrt{2\log p}\max_i z_i)$ and then bounding $\max_i z_i \leq \sum_i z_i$, and the variance bound with Poincaré's inequality applied to a smooth maximum, even though it can be refined (Boucheron and M. Thomas, 2012). Putting everything together, we obtain

$$\mathbb{E}_{\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}^*, \boldsymbol{I}_n)}[\lambda_n(-1, \boldsymbol{y})] \geq \frac{p-1}{p}\left\{\frac{3(p-1)}{4} - \frac{1}{4}\sqrt{2\log(p-1)}\right\} - \frac{p}{2} + O(1),$$

which is $O(p)$ and therefore tends to ∞ as $p \to \infty$. This completes the proof.

2.10 Additional numerical illustrations in Section 2.5

Constrained Gaussian in three dimensions

We include the analog of Figure 2.9 with $1 - \alpha = 0.95$ in Figure 2.12.

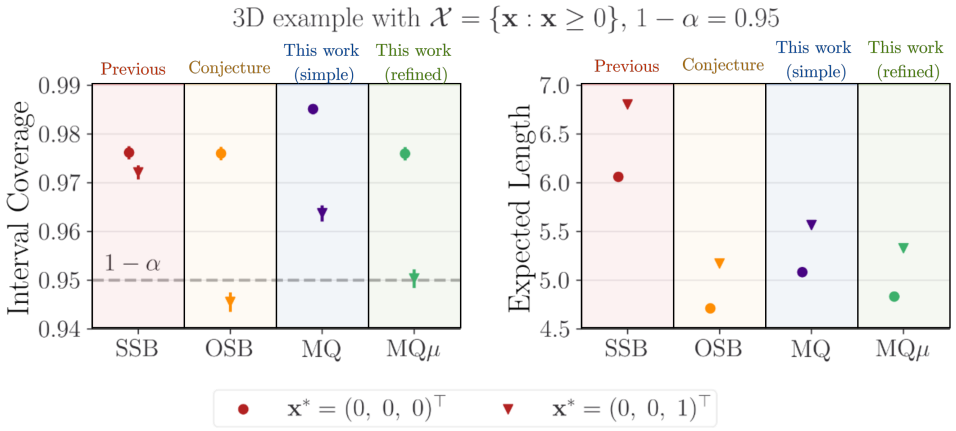


Figure 2.12: Estimated interval coverage **(left)** and expected lengths **(right)** for 95% intervals resulting from the SSB, OSB, MQ, and MQμ methods for the Gaussian linear model in (2.4) with $\boldsymbol{K} = \boldsymbol{I}_3$, $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x} = x_1 + x_2 - x_3$, and $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^3 : \boldsymbol{x} \geq 0\}$.

Chapter 3

CONFIDENCE INTERVALS FOR FUNCTIONALS IN CONSTRAINED INVERSE PROBLEMS VIA DATA-ADAPTIVE SAMPLING-BASED CALIBRATION

We address functional uncertainty quantification for ill-posed inverse problems where it is possible to evaluate a possibly rank-deficient forward model, the observation noise distribution is known, and there are known parameter constraints. We present four constraint-aware confidence interval constructions extending the theoretical test inversion framework in (Batlle, Stanley, et al., 2023) by making the intervals both computationally feasible and less conservative. Our approach first shrinks the potentially unbounded constraint set compact in a data-adaptive way, obtains samples of the relevant test statistic inside this set to estimate a quantile function, and then uses these computed quantities to produce the constraint-aware confidence intervals. Our approach to bounding the constraint set in a data-adaptive way is based on the approach by (Roger L. Berger and Boos, 1994), and involves defining a subset of the constraint set where the true parameter is guaranteed to exist with high probability. The probabilistic guarantee of this compact subset is then incorporated into the final coverage guarantee in the form of an uncertainty budget. We then propose custom sampling algorithms to efficiently sample from this subset, even when the parameter space is high-dimensional. Optimization-based interval methods formulate confidence interval computation as two endpoint optimizations, where the optimization constraints can be set to achieve different types of interval calibration while seamlessly incorporating parameter constraints. However, choosing optimization constraints to obtain coverage for a single functional has been elusive. We show that all four proposed intervals achieve nominal coverage for a particular functional both theoretically and in practice, with several numerical examples demonstrating superior performance of our intervals over the OSB interval in terms of both coverage and expected interval length. In particular, we show the superior performance of our intervals in a realistic unfolding simulation from high-energy physics that is severely ill-posed and involves a rank-deficient forward model.

3.1 Introduction

This paper proposes a novel uncertainty quantification (UQ) approach for ill-posed inverse problems characterized by a known parametric forward model[1] $f\colon \mathbb{R}^p \to \mathbb{R}^n$ mapping a parameter $\boldsymbol{x} \in \mathbb{R}^p$ to an observation $f(\boldsymbol{x}) \in \mathbb{R}^n$, additive noise with a known distribution $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and constraints on the forward model parameters denoted by $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ is primarily assumed to be of the form $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$. We consider scenarios where the parameter space dimension p may exceed the observation space dimension n, which often results in an ill-posed problem. Additionally, we do not require the forward model f to be injective, allowing for the possibility that f is many-to-one. We assume observations are generated by $\boldsymbol{y} = f(\boldsymbol{x}^*) + \boldsymbol{\varepsilon}$, where $\boldsymbol{x}^* \in \mathcal{X}$ is the true but unknown model parameter. Our UQ object of interest is a confidence interval[2], $C_\alpha(\boldsymbol{y})$, on a one-dimensional quantity of interest (QoI) derived from x^* by $\varphi : \mathcal{X} \to \mathbb{R}$. Our approach produces constraint-aware intervals with a finite-sample *frequentist coverage guarantee*, namely, for each $\alpha \in (0, 1)$

$$\mathbb{P}(\varphi(\boldsymbol{x}^*) \in C_\alpha(\boldsymbol{y})) \geq 1 - \alpha \text{ for all } \boldsymbol{x}^* \in \mathcal{X}. \tag{3.1}$$

The probability in the coverage guarantee (3.1) is taken over the distribution of the observation $\boldsymbol{y} = f(\boldsymbol{x}^*) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is the additive noise. The level $\alpha \in (0, 1)$ corresponds to the desired coverage probability, meaning that the interval $C_\alpha(\boldsymbol{y})$ will contain the true value $\varphi(\boldsymbol{x}^*)$ with probability at least $1 - \alpha$ for any realization of the noise $\boldsymbol{\varepsilon}$.

This paper builds on the confidence interval framework of (Batlle, Stanley, et al., 2023), which posits a more general data-generating process, $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$, and proposes a novel test-inversion framework for constructing constraint-aware confidence intervals for inverse problem QoI's. By proposing a testing framework to evaluate if $\boldsymbol{x}^* \in \Phi_\mu := \{\boldsymbol{x} \in \mathcal{X} : \varphi(\boldsymbol{x}) = \mu\}$ via the test statistic $\lambda(\mu, \boldsymbol{y})$, (Batlle, Stanley, et al., 2023) construct a confidence set as follows (see Equation (2.9)),

$$C_\alpha(\boldsymbol{y}) := \{\mu : \lambda(\mu, \boldsymbol{y}) \leq q_\alpha(\mu)\}. \tag{3.2}$$

As shown in Lemma 2.2 of (Batlle, Stanley, et al., 2023), there are two ways to set $q_\alpha(\mu)$ using a supremum over a particular quantile function, $Q_{\boldsymbol{x}}(1 - \alpha)$ (we show a particular realization of such a quantile function in the left-most panel of Figure 3.1),

[1] Even though we write the forward model as a known function f, we only require evaluation access to the map throughout the paper.

[2] We interchangeably use the terms "confidence set" and "confidence interval" since the latter may be obtained by the former by simply retaining the set endpoints.

such that $C_\alpha(\boldsymbol{y})$ achieves the desired coverage,

$$Q^{\max}_{\mu,1-\alpha} := \sup_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{X}} Q_{\boldsymbol{x}}(1-\alpha), \tag{3.3}$$

$$Q^{\max}_{1-\alpha} := \sup_{\boldsymbol{x}\in\varphi(\mathcal{X})} Q^{\max}_{\mu,1-\alpha} = \sup_{\boldsymbol{x}\in\mathcal{X}} Q_{\boldsymbol{x}}(1-\alpha). \tag{3.4}$$

In practice, not only are Equations (3.3) and (3.4) difficult to compute because $Q_{\boldsymbol{x}}(1-\alpha)$ is typically only accessible via sampling, but the optimizations can be expressed equivalently as chance-constrained programs (see Lemma 3.1 in (Batlle, Stanley, et al., 2023)), which are typically challenging to solve. Furthermore, although $Q^{\max}_{\mu,1-\alpha}$

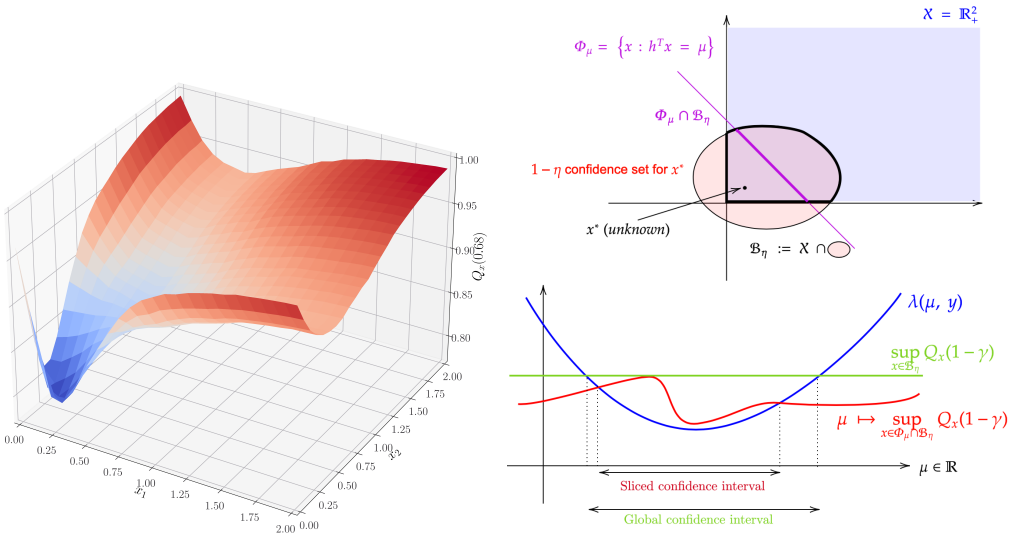


Figure 3.1: **(Left)** A particular quantile surface, $Q_{\boldsymbol{x}}(1-\alpha)$, where $\boldsymbol{x} \geq \mathbf{0}$ and $\alpha = 0.32$. This surface was obtained via Monte Carlo sampling the LLR test statistic over a grid of $\boldsymbol{x}$'s defined by the two-dimensional constrained-Gaussian scenario similar to that in Section 3.6, but with $\boldsymbol{h} := \begin{pmatrix}1 & 1\end{pmatrix}^\top$. **(Right-Top)** An illustration of the Berger–Boos set and other, a $1-\eta$ confidence set for $\boldsymbol{x}^*$, which prevents having to contend with an unbounded parameter space. Additionally, it can eliminate "worst-case" parameter settings in $\mathcal{X}$ that are far from the data, thus potentially making the resulting intervals less conservative. **(Right-Bottom)** An illustration of an LLR test statistic curve, $\lambda(\mu, \boldsymbol{y})$, over the QoI domain and the interval endpoints that result from using either Equation (3.5) for the "Sliced" interval or Equation (3.6) for the "Global" interval.

and $Q^{\max}_{1-\alpha}$ are optimal with respect to the setup in which they are analyzed, they can still produce overly conservative (larger than necessary) confidence sets if the true $\boldsymbol{x}^*$ is far away from the worst-case $\boldsymbol{x}$ against which the above maximum quantiles protect. This paper addresses these practical and statistical challenges by including a bounded data-informed subset of the parameter space, $\mathcal{B}_\eta$ with $\eta \in (0,1)$, such

that $\mathbb{P}\left(\boldsymbol{x}^* \in \mathcal{B}_\eta\right) \geq 1-\eta$ (see Section 3.3). We refer to $\mathcal{B}_\eta$ as the "Berger–Boos" set due to its inspiration from (Roger L. Berger and Boos, 1994). We show that we can reformulate Equations (3.3) and (3.4) to be constrained over $\mathcal{B}_\eta \subset \mathcal{X}$ if we also use a slightly larger quantile level, $1-\gamma$ for $\gamma \in (\alpha, 1)$, and such that $1-\alpha = (1-\eta)(1-\gamma)$, producing the following maximum quantiles:

$$\bar{q}_{\gamma,\eta}(\mu) := \sup_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{B}_\eta} Q_{\boldsymbol{x}}(1-\gamma), \tag{3.5}$$

$$\bar{q}_{\gamma,\eta} := \sup_{\boldsymbol{x} \in \mathcal{B}_\eta} Q_{\boldsymbol{x}}(1-\gamma). \tag{3.6}$$

Since $\bar{q}_{\gamma,\eta}(\mu)$ and $\bar{q}_{\gamma,\eta}$ are defined over bounded subsets of $\mathcal{X}$, we can propose a novel sampling procedure to obtain design points covering $\mathcal{B}_\eta$ and use quantile regression to estimate the desired quantile surface, $Q_{\boldsymbol{x}}(1-\gamma)$. The estimated quantile surface is then used to produce estimates of either $\bar{q}_{\gamma,\eta}(\mu)$ or $\bar{q}_{\gamma,\eta}$ and constructs a confidence set using a set definition similar to Equation (3.2),

$$C_\alpha(\boldsymbol{y}; \mathcal{B}_\eta) := \{\mu : \lambda(\mu, \boldsymbol{y}) \leq q\}, \tag{3.7}$$

where q is set to one of $\bar{q}_{\gamma,\eta}(\mu)$ or $\bar{q}_{\gamma,\eta}$ to obtain the desired interval. We present a variety of ways to perform the sampling and compare the different confidence set formulations and computations.

Context and related work

With advances in data collection and computational processing, high-dimensional ill-posed inverse problems have become more prevalent, especially in fields like remote sensing and data assimilation. This setting includes a wide array of physical science applications, spanning Earth science (Rodgers, 2000), atmospheric science and remote sensing (J. Liu, Bowman, Meemong, et al., 2016; Patil, Kuusela, and Hobbs, 2022), and high energy physics (Kuusela, 2016; Stanley, Patil, and Kuusela, 2022), among many others. Providing guarantees for UQ in parameter inference from indirect observations is essential for assessing the precision of scientific inferences made in these contexts. However, the inherent ill-posed nature of these problems often leads to inferences that are highly sensitive to noise, posing significant challenges for UQ. Namely, the ill-posedness leads to an identifiability issue in which statistical inference is impossible without providing some form of regularization, which usually takes either a deterministic (e.g., SVD truncation (Höcker and Kartvelishvili, 1996) and Tikhonov regularization (Schmitt, 2012)) or probabilistic (e.g., priors and Bayesian inference) form. Under some assumptions, these

two approaches are mathematically equivalent and are therefore subject to the same pitfalls, such as a disruption of the coverage guarantees of downstream intervals as a result of the incurred regularization bias as thoroughly discussed in (Kuusela, 2016). Our method's focus on one-dimensional QoI's and incorporation of parameter constraints allows for implicit regularization and therefore produces intervals with the promised coverage guarantee while avoiding the problem of regularization bias altogether.

Although including parameter constraints provides implicit regularization and enables handling non-trivial null spaces, it shifts complexity to statistical inference under constraints, a non-trivial problem in even elementary settings (Gouriéroux, Holly, and Monfort, 1982; Wolak, 1987; Robertson, F. T. Wright, and Dykstra, 1988; Alexander Shapiro, 1988; Wolak, 1989; Molenberghs and Verbeke, 2007). One elegant solution to this problem originates in what we refer to as *optimization-based UQ*, originating in the work of (W. R. Burrus, 1965) and (Burt W. Rust and Walter R. Burrus, 1972). (Philip B. Stark, 1992b) extended and generalized their approach, calling this method *strict bounds*, since it produced guaranteed simultaneous coverage interval estimators complying with known physical constraints on the model parameters. This collection of work defines endpoint optimization problems over the physically constrained parameter space to directly compute confidence intervals for one-dimensional QoI's. Not only does the optimization form of the confidence interval computation shift statistical inference complexity to numerical optimization, but it also allows known physical constraints to be directly included in the endpoint optimizations. This practical advantage is dulled by the difficulty of proving the coverage properties of the intervals resulting from the defined endpoint optimizations in the one-at-a-time setting. Even under the relatively strong assumptions in (Burt W. Rust and Walter R. Burrus, 1972) of a linear forward model, non-negativity parameter constraints and linear QoI, the authors were only able to conjecture the coverage of their interval (known as the Burrus Conjecture). The interval coverage was unsuccessfully proven in (O'Leary and Bert W. Rust, 1986) (the error pointed out in (Tenorio, Fleck, and Moses, 2007)) and finally generally refuted in (Batlle, Stanley, et al., 2023). These optimized-based intervals that have gained recent attention take the form

$$\mathcal{I}(\psi_\alpha^2, \boldsymbol{y}) := \left[\varphi^l(\psi_\alpha^2, \boldsymbol{y}), \varphi^u(\psi_\alpha^2, \boldsymbol{y})\right] = \left[\min_{\boldsymbol{x}\in D(\psi_\alpha^2,\boldsymbol{y})} \varphi(\boldsymbol{x}), \max_{\boldsymbol{x}\in D(\psi_\alpha^2,\boldsymbol{y})} \varphi(\boldsymbol{x})\right], \tag{3.8}$$

where

$$D(\psi_\alpha^2, \boldsymbol{y}) := \left\{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{y} - f(\boldsymbol{x})\|_2^2 \leq \psi_\alpha^2\right\}. \tag{3.9}$$

The statistical challenge for intervals of this form is the choosing of ψ_α^2 such that $\mathcal{I}(\psi_\alpha^2, \boldsymbol{y})$ has the desired coverage guarantee. It is important to note that the definition of $D(\psi_\alpha^2, \boldsymbol{y})$ assumes that the noise has been standardized; in other words, the covariance matrix of the noise has been transformed to the identity matrix.

The literature proposes two settings to guarantee coverage. One approach is to set ψ_α^2 such that $D(\psi_\alpha^2, \boldsymbol{y})$ is itself a confidence set for $\boldsymbol{x}^*$ in the parameter space. Since this choice would then automatically guarantee coverage for $\mathcal{I}(\psi_\alpha^2, \boldsymbol{y})$ regardless of the chosen QoI, this setting has been called "simultaneous" in (O'Leary and Bert W. Rust, 1986), and "Simultaneous Strict Bounds" (SSB) in (Stanley, Patil, and Kuusela, 2022; Batlle, Stanley, et al., 2023) since it also aligns with the setting of the strict bounds construction in (Philip B. Stark, 1992b). Under the Gaussian assumption, this can be achieved by setting $\psi_{SSB,\alpha}^2 := \chi_{n,\alpha}^2$, where $\chi_{n,\alpha}^2$ is the upper α-quantile for a chi-squared distribution with n degrees of freedom. Although this approach is relatively simple and obtains the desired coverage guarantee, it is conservative since the guarantee holds for all possible QoI choices simultaneously. To tailor the interval to one particular QoI, there is the "one-at-a-time" setting as described in (Burt W. Rust and Walter R. Burrus, 1972; O'Leary and Bert W. Rust, 1986), or "one-at-a-time strict bounds" (OSB) as called in (Stanley, Patil, and Kuusela, 2022; Batlle, Stanley, et al., 2023). Under the Gaussian assumption this proposed setting was $\psi_{OSB,\alpha}^2 = \chi_{1,\alpha}^2 + s(\boldsymbol{y})^2$, where $s(\boldsymbol{y})^2 := \min_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{y} - f(\boldsymbol{x})\|_2^2$. Unlike the simultaneous setting, $D(\psi_{OSB,\alpha}^2, \boldsymbol{y})$ is not a $1-\alpha$ confidence set for $\boldsymbol{x}^*$, which makes proving its coverage guarantee difficult. The validity of this claim was proposed by the Burrus conjecture (Burt W. Rust and Walter R. Burrus, 1972) and generally disproven by (Batlle, Stanley, et al., 2023).

To address the challenge of calibrating strict bounds and optimization-based confidence intervals, (Batlle, Stanley, et al., 2023) approached these intervals as an inverted likelihood ratio test as shown in Equation (3.2). The distribution of the log-likelihood ratio (LLR) statistic is non-standard due to the presence of the constraints, adding to the complexity of controlling the type-I error of the test. This framework development allowed for the general disproving of the long-standing Burrus conjecture, invalidating the general calibration of the intervals proposed by (Burt W. Rust and Walter R. Burrus, 1972). However, the framework revealed that the characterization of $Q_{\boldsymbol{x}}(1-\alpha)$ allowed for proper interval calibration via Equations (3.3) or (3.4) in much more general settings than those in which the Burrus conjecture applied (linear forward model, Gaussian noise, and linear functional).

Our approach and contributions

To address the computational challenges and statistical conservatism following from the use of quantiles (3.3) or (3.4) in confidence set (3.2), this paper develops analogous maximum quantiles over the bounded Berger–Boos set (quantiles (3.5) and (3.6)) for the altered confidence set (3.7). We refer to the confidence intervals produced by quantile (3.5) as "Sliced", since it considers the maximum quantile along slices of the Berger–Boos set as defined by level-sets of the QoI. We refer to the confidence intervals produced by quantile (3.6) as "Global", since it considers the maximum quantile over the entire Berger–Boos set. For both the sliced and global forms of the confidence interval, we present and develop both optimization and sampling approaches. In total, these confidence interval construction options produce four different interval varieties.

To validate the coverage guarantees of these interval varieties, we conduct a series of numerical experiments to demonstrate the effectiveness of the intervals in various scenarios. Since the OSB interval described in (Patil, Kuusela, and Hobbs, 2022) and (Stanley, Patil, and Kuusela, 2022) is the current standard approach in this setting, it is our primary point of comparison. First, we present low-dimensional examples. The first example is a two-dimensional constrained Gaussian noise model, a well-studied case in the literature (see (Tenorio, Fleck, and Moses, 2007), (Batlle, Stanley, et al., 2023)), illustrating how our intervals are competitive with the OSB interval in a scenario where the OSB interval is known to achieve nominal coverage. The second example is a three-dimensional constrained Gaussian noise model, illustrating the advantage of our method in achieving nominal coverage where the OSB interval fails. Second, we apply our method to a simulated version of particle unfolding from high-energy physics (see (Stanley, Patil, and Kuusela, 2022)), a binned deconvolution problem, in an 80-dimensional parameter space with a rank-deficient linear forward model. We use the previously studied realistic setting from (Stanley, Patil, and Kuusela, 2022) where the OSB interval is known to provide coverage to show that our Sliced intervals substantially outperform OSB in terms of expected length. Then, we use a parameter setting where the OSB interval does not achieve nominal coverage, but our intervals do, with the Sliced intervals still out-performing OSB in terms of expected length. This application highlights the practical significance of our approach, with our confidence intervals consistently achieving nominal coverage and often outperforming existing methods in terms of expected interval length.

The contributions of this work are both methodological and computational. Methodologically, we propose new confidence interval constructions for the setting described at the beginning of the section. Although the use of the Berger–Boos set is inspired by (Roger L. Berger and Boos, 1994), they originally applied it in a hypothesis testing setting for handling nuisance parameters. (Masserano, A. Shen, et al., 2024) applied a similar idea to nuisance parameters in a classification setting, but to the best of our knowledge, our paper is the first to apply the idea in the ill-posed inverse problem UQ setting. Inspired by the methods used in simulation-based inference (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Niccolo Dalmasso et al., 2024; Masserano, Dorigo, et al., 2023) we apply the approach of estimating the quantile function of a test statistic using simulated data. However, this work differs in the underlying model assumptions and composite nature of the null hypotheses. In addition, the prior work assumed bounded and relatively low-dimensional parameter spaces. The search for confidence interval endpoints is sometimes done via the Robbins-Monro (RM) procedure, which iteratively refines estimates to achieve desired coverage probabilities (Garthwaite and Buckland, 1992; Carpenter, 1999). Methods based on quantile regression (Roger Koenker, 2005) have been proposed to improve accuracy. For instance, (Fisher, Schweiger, and Rosset, 2020) introduced a technique that inverts estimated quantiles to determine the endpoints of the confidence intervals. This work is the first to combine test inversion, sampling, and quantile regression to estimate calibrated constraint-aware confidence intervals.

Computationally, we propose two custom sampling methods designed specifically to draw design points within the defined Berger–Boos set. This set can be particularly challenging to sample since it is the intersection between a pre-image based on the observed data and linear parameter constraints, which can produce sharp edges and corners. This set is also elongated due to the ill-posedness of the underlying inverse problem. While there are some approaches to handle similar scenarios arising in more traditional MCMC sampling (e.g., nested sampling (Skilling, 2004; Buchner, 2023) and Hamiltonian Monte Carlo sampling with constraints (Pakman and Paninski, 2014)), these algorithms can be intricate, so we develop the following alternatives. In the low-dimensional setting ($p < 10$) where the forward model is linear and the noise distribution is Gaussian, we design an accept-reject sampler based on sampling from p–balls as described in (Voelker, Gosmann, and Stewart, 2017). Although the approach in (Voelker, Gosmann, and Stewart, 2017) can rapidly sample from the pre-image ellipsoids as defined by the data-generating model assumptions, with even a simple non-negativity constraint on the parameters, the proportion of

rejected points quickly get impractically large for even moderate dimensions. For moderate and higher dimensional settings ($p \geq 10$), we leverage the Vaidya walk MCMC algorithm from (Yuansi Chen et al., 2018) to generate random walks about a polytope enclosing the Berger–Boos set. Using the Chebyshev ball center to define a notion of polytope centrality along with the extreme points of the Berger–Boos set with respect to the QoI, we create two lines along which starting positions are defined to generate a collection of parallel Markov chains. The design points resulting from these chains are then combined to create the complete set of sampled design points spanning the full range of the QoI values over the Berger–Boos set.

The rest of the paper proceeds as follows. In Section 3.2, we recapitulate the framework and theoretical components of (Batlle, Stanley, et al., 2023) upon which this paper's work is built. Section 3.3 then presents the theoretical Global and Sliced intervals under the Berger–Boos set formulation, along with the four interval constructions and theoretical justifications to guarantee that the constructed sampling-based intervals converge in probability to their theoretical counterparts. In Section 3.4, we give theoretical results regarding our proposed methods. Section 3.5 presents the custom sampling algorithms along with a brief description of how quantile regression works in our intervals. Section 3.6 presents four numerical experiments to demonstrate the coverage and length advantages of our intervals over the OSB interval. Finally, Section 3.7 provides some discussion and conclusions.

3.2 Background

A key part of the approach in (Batlle, Stanley, et al., 2023) was to view the optimization-based intervals shown by Equation (3.8) as inverted hypothesis tests, more specifically, a particular inverted likelihood ratio test. This change in perspective allowed for interval analysis through the lens of the properties of a particular log-likelihood ratio (LLR) test. We summarize this connection here, as it is critical to our method development carried out in Section 3.3. Although the numerical examples in Section 3.6 focus on the linear-Gaussian version of the data-generating process, we follow the more general exposition of (Batlle, Stanley, et al., 2023) to indicate that this framework is not limited to that scenario.

Suppose $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$, where $P_{\boldsymbol{x}^*}$ is a distribution parameterized by a fixed but unknown $\boldsymbol{x}^* \in \mathbb{R}^p$. Let $\ell_{\boldsymbol{x}}(\boldsymbol{y})$ denote the log-likelihood of $\boldsymbol{x}$ evaluated at $\boldsymbol{y}$. We furthermore suppose we know a set $\mathcal{X} \subseteq \mathbb{R}^p$ such that $\boldsymbol{x}^* \in \mathcal{X}$.

The duality between hypothesis tests and confidence sets is well known in statistics

(see, e.g., Chapter 7 of (George Casella and Roger L. Berger, 2002) or Chapter 5 of (Victor M. Panaretos, 2016)). We invert the following hypothesis test:

$$H_0 : \boldsymbol{x}^* \in \Phi_\mu \cap \mathcal{X} \quad \text{versus} \quad H_1 : \boldsymbol{x}^* \in \mathcal{X} - \Phi_\mu, \tag{3.10}$$

where $\Phi_\mu := \{\boldsymbol{x} : \varphi(\boldsymbol{x}) = \mu\}$. Since we are looking to obtain a confidence interval on the real line, it makes sense that each hypothesis test is defined by $\mu \in \mathbb{R}$. Notice that the composite structure of the null hypothesis includes all parameter settings within the μ-level set of the functional of interest.

We define the following test statistic to evaluate Test (3.10):

$$\lambda(\mu, \boldsymbol{y}) := \begin{cases} \inf_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} \left(-2\ell_{\boldsymbol{x}}(\boldsymbol{y})\right) \; - \; \inf_{\boldsymbol{x} \in \mathcal{X}} \left(-2\ell_{\boldsymbol{x}}(\boldsymbol{y})\right), & \text{if } \Phi_\mu \cap \mathcal{X} \neq \emptyset, \\ \infty, & \text{otherwise,} \end{cases} \tag{3.11}$$

where μ denotes the level set of the null hypothesis, $\boldsymbol{y}$ is the observed data, and $\mathcal{X}$ is the constraint set that is known to contain the parameter. The test is rejected if the test statistic is large, and hence we automatically reject the null if $\Phi_\mu \cap \mathcal{X} = \emptyset$ independently of the observed data. Alternatively, we can consider the test (3.10) only for $\mu \in \varphi(\mathcal{X})$ to perform test inversion. We control the behavior of this test statistic, and therefore the test, by bounding from above the probability of erroneously rejecting the null hypothesis (i.e., type-1 error). As such, we consider the distribution of $\lambda(\mu, \boldsymbol{y})$ under the null. For each $\boldsymbol{x} \in \mathcal{X}$, let $\mu = \varphi(\boldsymbol{x})$. Define $Q_{\boldsymbol{x}} : (0, 1) \to \mathbb{R}$ such that, for all $\alpha \in (0, 1)$,

$$\mathbb{P}\left(\lambda(\mu, \boldsymbol{y}) \leq Q_{\boldsymbol{x}}(1 - \alpha)\right) = 1 - \alpha, \tag{3.12}$$

where the probability is over $\boldsymbol{y} \sim P_{\boldsymbol{x}}$. We refer to $Q_{\boldsymbol{x}}$ as the *quantile function* of the LLR under the null hypothesis at $(\boldsymbol{x}, \varphi(\boldsymbol{x}) = \mu)$. Since the null hypothesis is composite, using this quantile function to define a cutoff is not enough to control type-1 error. Thus, we use the *sliced* maximum quantile function over the level-set under consideration:

$$Q^{\max}_{\mu, 1-\alpha} := \sup_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} Q_{\boldsymbol{x}}(1 - \alpha). \tag{3.13}$$

Note, we refer to this version as "sliced" since we are considering the maximum of the quantile function defined by the level-set of the functional which defines a slice through the constrained parameter space. $Q^{\max}_{\mu, 1-\alpha}$ as defined above controls the type-1 error for a specific value of μ. We can define a more conservative confidence set using the following *global* maximum quantile:

$$Q^{\max}_{1-\alpha} := \sup_{\mu \in \varphi(\mathcal{X})} Q^{\max}_{\mu, 1-\alpha} = \sup_{\boldsymbol{x} \in \mathcal{X}} Q_{\boldsymbol{x}}(1 - \alpha). \tag{3.14}$$

By Lemma 2.1 in (Batlle, Stanley, et al., 2023), the set

$$C_\alpha^\mu(\boldsymbol{y}; \mathcal{X}) := \{\mu : \lambda(\mu, \boldsymbol{y}) \leq Q_{\mu,1-\alpha}^{\max}\} \subset \mathbb{R} \tag{3.15}$$

defines a $1-\alpha$ confidence set for the true functional value, $\mu^* = \varphi(\boldsymbol{x}^*)$.The global quantile (3.14) is more conservative than the sliced quantile (3.13) since it holds for all null hypotheses, and therefore the set

$$C_\alpha(\boldsymbol{y}; \mathcal{X}) := \{\mu : \lambda(\mu, \boldsymbol{y}) \leq Q_{1-\alpha}^{\max}\} \subset \mathbb{R} \tag{3.16}$$

is also a $1-\alpha$ confidence set.

To connect this framework back to Interval (3.8), suppose the data generating process is of the form, $P_{\boldsymbol{x}} = \mathcal{N}(\boldsymbol{K}\boldsymbol{x}, \boldsymbol{I})$. The LLR test statistic is then

$$\lambda(\mu, \boldsymbol{y}) = \min_{\boldsymbol{x} \in \Phi_\mu \cap \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2. \tag{3.17}$$

(Batlle, Stanley, et al., 2023) (Theorem 2.4) proved that

$$\left[\inf_{\mu:\lambda(\mu,\boldsymbol{y}) \leq Q_{1-\alpha}^{\max}} \mu, \sup_{\mu:\lambda(\mu,\boldsymbol{y}) \leq Q_{1-\alpha}^{\max}} \mu\right] = \left[\inf_{\boldsymbol{x} \in D(Q_{1-\alpha}^{\max} + s(\boldsymbol{y})^2, \boldsymbol{y})} \varphi(\boldsymbol{x}), \sup_{\boldsymbol{x} \in D(Q_{1-\alpha}^{\max} + s(\boldsymbol{y})^2, \boldsymbol{y})} \varphi(\boldsymbol{x})\right], \tag{3.18}$$

where $D(\cdot, \boldsymbol{y})$ is defined by Equation (3.9). This equivalence asserts that by setting $\psi_\alpha^2 := Q_{1-\alpha}^{\max} + s(\boldsymbol{y})^2$, where $s(\boldsymbol{y})^2 := \min_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2$, we guarantee coverage for Interval (3.8). Furthermore, it asserts that the original OSB interval formulation is only valid if and only if $\chi_{1,\alpha}^2 \geq Q_{1-\alpha}^{\max}$ (see Lemma 2.2 and Corollary 2.3). (Batlle, Stanley, et al., 2023) showed that this inequality does not hold in general.

As explored in (Batlle, Stanley, et al., 2023) and reiterated above, Interval (3.8) can be calibrated by computing $Q_{\mu,1-\alpha}^{\max}$ or $Q_{1-\alpha}^{\max}$. However, pursuing calibration in this way is computationally challenging and statistically conservative. Both of these values require the ability to evaluate $Q_{\boldsymbol{x}}$. Without parameter constraints, this quantile function can be constant under Gaussian-linear assumptions (Batlle, Stanley, et al., 2023), i.e., the test statistic is pivotal. But even with a relatively simple two-dimensional Gaussian noise model with non-negativity constraints, this quantile function becomes non-trivial (e.g., see Figure 5.3 in (Batlle, Stanley, et al., 2023)). Beyond the practical difficulty of dealing with the underlying quantile function, both (3.13) and (3.14) can be expressed as chance-constrained optimization problems, which are known to be NP-hard in general and would need to be solved over an unbounded constraint set (Geng and Xie, 2019b; Pena-Ordieres, Luedtke, and

Wachter, 2020; Batlle, Stanley, et al., 2023). Statistically, since both (3.13) and (3.14) are the largest quantile function values subject to their respective constraints, they are by definition conservative, especially in scenarios where the true $\boldsymbol{x}^*$ is far from these most conservative points. Since we do not know $\boldsymbol{x}^*$, this conservatism is necessary under the framework of (Batlle, Stanley, et al., 2023).

As we see in Section 3.3, our method addresses both of these challenges. In scenarios where evaluating $Q_{\boldsymbol{x}}$ is difficult, we sample a collection of design points in a bounded subset of the constraint set, sample the LLR under the null at each design point and use quantile regression to estimate the quantile surface. Furthermore, we can potentially remove these most conservative points from consideration by only considering parameter values that are *not unlikely* given the observed data.

3.3 Interval constructions

In this section, we present four related interval constructions to build on the theory developed in (Batlle, Stanley, et al., 2023) by addressing the key aforementioned challenges. The first implementation challenge is to handle the potentially unbounded constraint set, for which we define and apply the "Berger–Boos" set to create a data-dependent subset of the original constraint set. This set leads to our four interval definitions, which follow a two-stage taxonomy: Global versus Sliced and Inverted versus Optimized. The second implementation challenge is computing these interval constructions in practice, which we achieve using a combination of novel sampling algorithms and quantile regression. We will cover this in a later section (Section 3.5).

We rewrite the data-generating process articulated in the introduction:

$$\boldsymbol{y} = f(\boldsymbol{x}^*) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{x}^* \in \mathcal{X}, \tag{3.19}$$

where $f : \mathbb{R}^p \to \mathbb{R}^n$ is a known forward model and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{I}$, without loss of generality. Let $f^{-1}(A) := \{\boldsymbol{x} : f(\boldsymbol{x}) \in A\}$ be the pre-image of a set $A \subset \mathbb{R}^n$ under the forward map f.

Although the assumed data-generating process given by Equation (3.19) is less general than that of $\boldsymbol{y} \sim P_{\boldsymbol{x}^*}$ assumed in Section 3.2, it is still sufficiently general to contain the application areas mentioned in Section 3.1. Slightly generalizing the form of the additive noise distribution would make the construction of the Berger–Boos set more complicated, but is possible in principle.

Global and sliced confidence sets using the Berger–Boos set

Both $Q^{\max}_{\mu,1-\alpha}$ and $Q^{\max}_{1-\alpha}$ in Section 3.2 suffer from the same theoretical and practical concerns. Theoretically, they are both *conservative* in the sense that they must control type-1 error probability under the worst case for the truth (i.e., the parameter setting with the largest quantile). Practically, not only can $Q_{\boldsymbol{x}}$ be difficult to compute, but if the constraint set $\mathcal{X}$ is unbounded, computing the aforementioned quantiles becomes even more difficult. These challenges both stem from the composite nature of the null hypothesis. (Roger L. Berger and Boos, 1994) introduce a compelling solution in the context of hypothesis testing with nuisance parameters (a special case of hypothesis testing on $\varphi(\boldsymbol{x})$), which is to control type-1 error only over a data-informed region of the parameter space. Following their construction, we build a confidence interval that, instead of using the maximum $1-\alpha$ quantile over $\mathcal{X}$ in (3.14) (or $\mathcal{X} \cap \Phi_\mu$ in (3.13)), uses a larger quantile ($1-\gamma$ with $\gamma < \alpha$) but that is maximized over a smaller set. The construction follows a three-step process:

(1) Choose $\eta \in [0, \alpha]$ and build a $1-\eta$ confidence interval for x^*, $\mathcal{B}_\eta$. Under the additive Gaussian noise assumption in Equation (3.19), letting $\Gamma_\eta(\boldsymbol{y}) := \{\boldsymbol{y}' \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{y}'\|_2^2 \leq \chi^2_{n,\eta}\}$ we have $\mathbb{P}\left(f(\boldsymbol{x}^*) \in \Gamma_\eta(\boldsymbol{y})\right) = 1-\eta$, and hence

$$\mathcal{B}_\eta := f^{-1}(\Gamma_\eta(\boldsymbol{y})) \cap \mathcal{X} = \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{y} - f(\boldsymbol{x})\|_2^2 \leq \chi^2_{n,\eta}\} \tag{3.20}$$

is a $1-\eta$ confidence set for x^*[3]. We refer to this pre-image confidence set, $\mathcal{B}_\eta$, as the "Berger–Boos" set.

(2) Optimize the $1-\gamma$ quantile of the test statistic only over $\mathcal{B}_\eta$ (or $\mathcal{B}_\eta \cap \Phi_\mu$), instead of $\mathcal{X}$ (or $\mathcal{X} \cap \Phi_\mu$) . Here, $\gamma < \alpha$ is chosen to ensure calibration. As proved in lemma 3.3.1, $\gamma = \alpha - \eta$ is a valid choice.

(3) Use the obtained quantiles, which we define as

$$\bar{q}_{\gamma,\eta}(\mu) := \sup_{\boldsymbol{x} \in \mathcal{B}_\eta \cap \Phi_\mu} Q_{\boldsymbol{x}}(1-\gamma), \tag{3.21}$$

$$\bar{q}_{\gamma,\eta} := \sup_{\boldsymbol{x} \in \mathcal{B}_\eta} Q_{\boldsymbol{x}}(1-\gamma), \tag{3.22}$$

that are used to construct the following sliced (sl) and global (gl) confidence sets:

$$C^{\mathrm{sl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta) := \{\mu : \lambda(\mu, \boldsymbol{y}) \leq \bar{q}_{\gamma,\eta}(\mu)\}, \tag{3.23}$$

$$C^{\mathrm{gl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta) := \{\mu : \lambda(\mu, \boldsymbol{y}) \leq \bar{q}_{\gamma,\eta}\}. \tag{3.24}$$

[3]Intersecting the pre-image $f^{-1}(\Gamma_\eta(\boldsymbol{y}))$ with the constraint set $\mathcal{X}$ does not change the coverage probability since we know $\boldsymbol{x}^* \in \mathcal{X}$.

Analogous to the presentation in (Batlle, Stanley, et al., 2023), we define *sliced* and *global* max-quantiles that include the Berger–Boos set to control the type-1 error probability of Test (3.10). Both of these max-quantiles maximize a larger γ-quantile over the set of interest instead of the α-quantile as shown in (3.13) and (3.14).

The following Lemma (analogous to Lemma 2.1 in (Batlle, Stanley, et al., 2023)) gives a sufficient condition to select the values γ and η to ensure a $1-\alpha$ coverage.

Lemma 3.3.1 (Setting η and γ to guarantee $1-\alpha$ coverage)**.** *Let $\alpha \in (0,1)$ and define $\bar{q}_{\gamma,\eta}(\mu)$ according* (3.21) *and $C^{\mathrm{sl}}_{\alpha}(\boldsymbol{y}; \mathcal{B}_\eta)$ according to* (3.23)*. For $\eta \in (0, \alpha)$, $C^{\mathrm{sl}}_{\alpha}(\boldsymbol{y}; \mathcal{B}_\eta)$ is a $1-\alpha$ confidence set for $\mu^* = \varphi(\boldsymbol{x}^*)$ if $\gamma \leq \alpha - \eta$. The length of the obtained interval is a non-increasing function of γ so the tightest interval will be obtained when equality is satisfied.*

Proof. See Section 3.8. □

The coverage guarantee implied by Lemma 3.3.1 also implies coverage for $C^{\mathrm{gl}}_{\alpha}(\boldsymbol{y}; \mathcal{B}_\eta)$, as shown in the following corollary.

Corollary 3.3.2. *In the setting of Lemma 3.3.1. $C^{\mathrm{gl}}_{\alpha}(\boldsymbol{y}; \mathcal{B}_\eta)$ is a $1-\alpha$ confidence set for $\mu^* = \varphi(\boldsymbol{x}^*)$ if $\gamma \leq \alpha - \eta$.*

Proof. See Section 3.8. □

Remark 6. Following the generality of the original construction, Lemma 3.3.1 generalizes to any $1-\eta$ confidence set of x^* and test statistic calibrated with the $1-\gamma$ quantile. In our particular case, since both the confidence set $\mathcal{B}_\eta$ and the test statistic $\lambda(\mu, y)$ depend on the Gaussian log-likelihood $\|\boldsymbol{y} - f(\boldsymbol{x})\|_2^2$, there is a stronger interpretation of the Berger-Boos construction. Whenever $\mathcal{B}_\eta \cap \Phi_\mu \neq \emptyset$, the Berger-Boos construction is equivalent to a data-dependant reduction of the constraint set, replacing $\mathcal{X}$ for $\mathcal{B}_\eta$ both in the test statistic and the quantile optimization problems. Since this is done after seeing the data, the optimized quantile needs to increase to $1-\alpha+\eta$ to maintain $1-\alpha$ coverage.

Navigating the Berger–Boos parameter choices and trade-offs. Setting $\eta = 0$ in $C^{\mathrm{gl}}_{\alpha}(\boldsymbol{y}; \mathcal{B}_\eta)$ we maximize $Q_{\boldsymbol{x}}(1-\gamma) = Q_{\boldsymbol{x}}(1-\alpha)$ over $\mathcal{X}$, which returns $Q^{\max}_{1-\alpha}$ from Equation (3.14). For $\eta > 0$, the Berger–Boos construction restricts the quantile maximization to a smaller subset of the parameter space while maximizing a larger $1-\gamma$ quantile to maintain the desired $1-\alpha$ confidence level of the final confidence

set. Fundamentally, the choice of η reflects a trade-off between these two opposing effects: as we increase η from 0, the set over which the quantile function is optimized shrinks, but the quantile level is increased. In Section 3.6, we perform a numerical experiment comparing average interval length for different values of η, showing that a small $\eta > 0$ can be beneficial not only numerically (since in certain cases it allows sampling over a bounded set) but also in terms of the average length of the resulting interval.

Interval constructions

There are possibly many ways to compute $C_\alpha^{\mathrm{gl}}(\boldsymbol{y}; \mathcal{B}_\eta)$ and $C_\alpha^{\mathrm{sl}}(\boldsymbol{y}; \mathcal{B}_\eta)$ in practice. Obtaining either set comes down to the computation of $\bar{q}_{\gamma,\eta}(\mu)$ and $\bar{q}_{\gamma,\eta}$. If the quantile function $Q_{\boldsymbol{x}}$ and its gradient $\nabla_{\boldsymbol{x}} Q_{\boldsymbol{x}}$ could be evaluated, computing these quantities could potentially be achieved using a first-order numerical optimizer. However, we emphasize that since the sliced max-quantile, $\bar{q}_{\gamma,\eta}(\mu)$, is a function of the level-set parameter, μ, such an optimization would have to be done for each possible functional value. Since such easy function and gradient evaluations rarely exist (see (Batlle, Stanley, et al., 2023) for some examples where such evaluations are possible), this paper develops a *sampling*-based approach to estimate these quantities to construct Confidence Sets (3.23) and (3.24). As we demonstrate below, once we estimate $\bar{q}_{\gamma,\eta}(\mu)$ and $\bar{q}_{\gamma,\eta}$, we can either use the output from the sampling algorithm to compute the Global or Sliced interval via classical test inversion, or we can use the estimated max-quantiles in optimizations similar to those in Interval (3.18). As such, we introduce four interval constructions: Global Inverted, Global Optimized, Sliced Inverted, and Sliced Optimized. These four options are summarized in Section 3.3.

We leverage our ability to sample $\lambda(\mu, \boldsymbol{y})$ by sampling $\boldsymbol{y} \sim \mathcal{N}(f(\boldsymbol{x}), \boldsymbol{I})$ to estimate the desired max-quantiles. We present two algorithms, both of which first generate a random set of design points from the Berger–Boos set. If we can directly *efficiently* compute the desired quantile at each design point, this capability is leveraged in Algorithm 2. If we cannot afford such a computation, Algorithm 3 presents an alternative, which first samples one realization of the LLR test statistic at each sampled design point, and then performs quantile regression with the generated pairs of design points and LLR values to estimate the underlying quantile surface. The sampled design points and either the exact or estimated quantiles at the design points are then used to produce the final intervals.

Both algorithms start by sampling $\mathcal{B}_\eta$ uniformly at random to generate a collection

of M design points across the Berger–Boos set, i.e., $\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \ldots, \bar{\boldsymbol{x}}_M \sim \mathcal{U}\left(\mathcal{B}_\eta\right)$. Even under the additive Gaussian noise assumption where the pre-image $f^{-1}(\Gamma_\eta(\boldsymbol{y}))$ is an ellipsoid, this step is non-trivial since the ill-posedness can produce an extremely narrow or unbounded ellipsoid, resulting in sharp boundaries when intersected with the known constraints in addition to a large portion of the pre-image lying outside of $\mathcal{X}$, making rejection-sampling challenging. We address these sampling challenges in Section 3.5. If computing $\lambda(\mu, \boldsymbol{y})$ for a given μ and $\boldsymbol{y}$ is inexpensive, Algorithm 2 directly estimates the quantile function at each design point. This computation is most likely inexpensive when there are closed-form solutions to the LLR's subordinate optimizations. As shown in Algorithm 2, an easy way to estimate each design point's quantile is to sample its test statistic N times and take the appropriate percentile. More often, computing $\lambda(\mu, \boldsymbol{y})$ is expensive since it involves two constrained optimizations which are possibly non-convex due to either the constraints or the forward model, or just numerically challenging due to the ill-posedness of the problem. In this scenario, one can use Algorithm 3 to sample one realization of the test statistic at each design point and estimate the quantile function over the Berger–Boos set using quantile regression. While Algorithm 2 relies upon computational strength to compute the LLR $N \times M$ times, Algorithm 3 shifts complexity to the quantile regression and relieves the computational burden by assuming that there is information to be shared about the quantile surface between design points (namely that the quantile surface is smooth). The emphasis on the quantile regression further necessitates that the quantile regression be performed well. We discuss some considerations to this end in Section 3.5. We note that although Algorithm 2 and Algorithm 3 make use of the additive Gaussian noise assumption, they are not necessarily limited to this assumption given the proper adjustments to the Berger–Boos set, assuming one still has the ability to sample from the data-generating process. Since these approaches are sampling-based, it is necessary to show convergence as the number of samples gets large.

The approach of Algorithm 3 is inspired by recent uncertainty quantification approaches in likelihood-free inference where one can sample from a likelihood but cannot easily compute it. Specifically, we draw inspiration from the use of quantile regression in (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Niccolo Dalmasso et al., 2024; Masserano, Dorigo, et al., 2023; Masserano, A. Shen, et al., 2024). Although these approaches differ in detail and implementation from our approach (e.g., they typically focus on settings with low-dimensional $\boldsymbol{x}$), they overlap in the sampling and quantile regression perspectives, which effectively allow machine learning

to supplement for computationally intensive or intractable forward models. Namely, rather than assuming a purely stochastic forward model and therefore only being able to sample from the likelihood, we assume f is a deterministic function involved in $\lambda(\mu, \boldsymbol{y})$, which is a random quantity due to the additive noise. Since Algorithm 3 involves training a quantile regressor, it includes separate training and testing sets of design points over the Berger–Boos set and samples from their respective test statistics.

Algorithm 2 Direct estimation of quantiles

Input: $\alpha, \gamma, \eta \in (0, 1)$ such that $\gamma = \alpha - \eta$, $M, N \in \mathbb{N}$.

1: **Construct Berger–Boos confidence set**: Create $\Gamma_\eta(\boldsymbol{y}) \subseteq \mathbb{R}^n$ such that $\mathbb{P}\left(f(\boldsymbol{x}^*) \in \Gamma_\eta(\boldsymbol{y})\right) \geq 1 - \eta$. $f^{-1}\left(\Gamma(\boldsymbol{y})\right)$ is also a $1 - \eta$ confidence set for $\boldsymbol{x}^*$, as is $\mathcal{B}_\eta = f^{-1}(\Gamma_\eta(\boldsymbol{y})) \cap \mathcal{X}$.
2: **Sample from the Berger–Boos confidence set,** $\mathcal{B}_\eta$: Sample $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_M \sim \mathcal{U}\left(\mathcal{B}_\eta\right)$.
3: **for** $k = 1, 2, \ldots, M$ **do**
4: **Sample noise realizations**: Sample N noise realizations: $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_N \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
5: **Sample from the LLR distribution**: Create an ensemble of LLR draws under $\bar{\boldsymbol{x}}_k$: $\{\lambda_i\}_{i=1}^N$, where $\lambda_i := \lambda\left(\varphi(\bar{\boldsymbol{x}}_k), f(\bar{\boldsymbol{x}}_k) + \boldsymbol{\varepsilon}_i; \mathcal{X}\right)$.
6: **Compute percentile estimate of the γ-quantile**: Compute the $(1-\gamma)\times 100$ percentile of the LLR samples for the data generating process under $\bar{\boldsymbol{x}}_k$, i.e., $\widehat{q}_\gamma^k := \lambda_{(\{(1-\gamma)N\})}$, where $\{\cdot\}$ denotes the nearest whole number and $\lambda_{(i)}$ denotes the i-th order statistic.
7: **end for**

Output: Pairs of sampled design points and their respective γ-quantiles, i.e., $\{(\bar{\boldsymbol{x}}_k, \widehat{q}_\gamma^k)\}_{k=1}^M$.

With the generated pairs from Algorithms (2) and (3), we present two strategies to estimate each of $C_\alpha^{\mathrm{gl}}(\boldsymbol{y}; \mathcal{B}_\eta)$ and $C_\alpha^{\mathrm{sl}}(\boldsymbol{y}; \mathcal{B}_\eta)$. To streamline notation, let $\left(\boldsymbol{x}_k, q_\gamma^k\right)$ denote the k-th pair from either algorithm. This is notationally helpful since Algorithm 2 only generates one set of parameter samples, whereas Algorithm 3 generates two. That is, using Algorithm 2, $\boldsymbol{x}_k := \bar{\boldsymbol{x}}_k$ and $q_\gamma^k := \widehat{q}_\gamma^k$ and using Algorithm 3, $\boldsymbol{x}_k := \widetilde{\boldsymbol{x}}_k$ and $q_\gamma^k := \widehat{q}_\gamma(\widetilde{\boldsymbol{x}}_k)$.

To estimate the global confidence set, we estimate $\bar{q}_{\gamma,\eta}$ using the empirical maximum, $\widehat{q} := \max_k q_\gamma^k$. This results in the following two interval constructions:

$$C_{\mathrm{opt}}^{\mathrm{gl}}(\boldsymbol{y}) := \min/\max\left\{\varphi(\boldsymbol{x}) : \boldsymbol{x} \in D(\widehat{q} + s(\boldsymbol{y})^2, \boldsymbol{y})\right\} \tag{3.25}$$

$$C_{\mathrm{inv}}^{\mathrm{gl}}(\boldsymbol{y}) := \min/\max\left\{\varphi(\boldsymbol{x}_k) : k = 1, \ldots, M \text{ and } \lambda(\varphi(\boldsymbol{x}_k), \boldsymbol{y}; \mathcal{X}) \leq \widehat{q}\right\}. \tag{3.26}$$

Algorithm 3 Quantile regression estimate of quantile surface

Input: $\alpha, \gamma, \eta \in (0,1)$ such that $\gamma = \alpha - \eta$; $M_{\text{tr}}, M \in \mathbb{N}$.

1: **Construct Berger–Boos confidence set**: Create $\Gamma_\eta(\boldsymbol{y}) \subseteq \mathbb{R}^n$ such that $\mathbb{P}\left(f(\boldsymbol{x}^*) \in \Gamma_\eta(\boldsymbol{y})\right) \geq 1 - \eta$. $f^{-1}\left(\Gamma(\boldsymbol{y})\right)$ is also a $1-\eta$ confidence set for $\boldsymbol{x}^*$, as is $\mathcal{B}_\eta = f^{-1}(\Gamma_\eta(\boldsymbol{y})) \cap \mathcal{X}$.
2: **Sample from the Berger–Boos confidence set** $\mathcal{B}_\eta$: Sample $\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_{M_{\text{tr}}} \sim \mathcal{U}\left(\mathcal{B}_\eta\right)$ design points to train the quantile regressor and $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_M \sim \mathcal{U}\left(\mathcal{B}_\eta\right)$ test points to invert the interval, generating $M_{\text{tr}} + M$ total samples. Since the test points are used for the interval inversion, they are used as out-of-sample points for the quantile regressor.
3: **for** $k = 1, 2, \ldots, M_{\text{tr}}$ **do**
4: **Sample a noise realization:** Sample a noise realization: $\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
5: **Sample from the LLR distribution:** Compute the LLR under $\bar{\boldsymbol{x}}_k$ with sampled noise $\boldsymbol{\varepsilon}_k$: $\lambda_k := \lambda\left(\varphi(\bar{\boldsymbol{x}}_k), f(\bar{\boldsymbol{x}}_k) + \boldsymbol{\varepsilon}_k; \mathcal{X}\right)$.
6: **end for**
7: **Estimate the quantile function using quantile regression:** Using the generated pairs $\{(\bar{\boldsymbol{x}}_k, \lambda_k)\}_{k=1}^{M_{\text{tr}}}$, estimate the upper γ-conditional quantile function, $\widehat{q}_\gamma(\boldsymbol{x})$, using quantile regression.

Output: Generate γ-quantile predictions at out-of-sample test points, $\left\{(\widetilde{\boldsymbol{x}}_k, \widehat{q}_\gamma(\widetilde{\boldsymbol{x}}_k))\right\}_{k=1}^{M}$.

We refer to Interval (3.26) as the Global Inverted interval construction since it is endpoints defined by only those sampled parameter values that comport with the maximum estimated quantile LLR cutoff. We refer to Interval (3.25) as the Global Optimized interval since its endpoints are defined by the extreme functional values of a feasible region defined by $\widehat{q}$. Although, $C_{\text{inv}}^{\text{gl}}(\boldsymbol{y}) \neq C_{\text{opt}}^{\text{gl}}(\boldsymbol{y})$ due to finite sample, they are asymptotically equal and in practice show similar performance in terms of coverage and expected length, as shown in Section 3.6. We prove the consistency of the interval constructed via inversion and via optimization in Theorem 3.4.1. Since the endpoints of both the inverted and optimized intervals converge in probability to the endpoints of $C_\alpha^{\text{gl}}(\boldsymbol{y}; \mathcal{B}_\eta)$, the two interval constructions are asymptotically equivalent.

To estimate the sliced set, we estimate $\bar{q}_{\gamma,\eta}(\mu)$ using a rolling maximum of the sampled q_γ^k values, as ordered by the sampled functional values, and directly accept or reject functional values based on their estimated quantile. These approaches result in two interval constructions:

$$C_{\text{opt}}^{\text{sl}}(\boldsymbol{y}) := \min/\max\left\{\mu \in \mathbb{R} : \lambda(\varphi(\mu), \boldsymbol{y}; \mathcal{X}) \leq \widehat{m}_\gamma(\mu)\right\}, \tag{3.27}$$

$$C_{\text{inv}}^{\text{sl}}(\boldsymbol{y}) := \min/\max\left\{\varphi(\boldsymbol{x}_k) : k = 1, \ldots, M \text{ and } \lambda(\varphi(\boldsymbol{x}_k, \boldsymbol{y}; \mathcal{X}) \leq q_\gamma^k)\right\}, \tag{3.28}$$

where $\widehat{m}_\gamma(\mu)$ denotes rolling estimate of $\bar{q}_{\gamma,\eta}(\mu)$ defined as follows. We refer to $I^{\text{sl}}(\boldsymbol{y})$ as the "sliced" index set, i.e., those indices for which the LLR at a particular functional value is less than the estimated quantile at a design point generating that functional value.

The rolling maximum quantile is defined using estimated quantiles ordered by the sampled functional values. Choose "rolling" parameter, $T \in \mathbb{N}$, and let $\sigma(1), \sigma(2), \ldots, \sigma(M)$ define an ordering such that $\mu_{\sigma(k)} \leq \mu_{\sigma(k+1)}$ for all $k = 1, \ldots, M-1$. Define $Q_k := \{q_\gamma^{\sigma(k)}, q_\gamma^{\sigma(k-1)}, \ldots, q_\gamma^{\sigma(k-T)}\}$ Then, for a given $\mu \in \left[\min_{\boldsymbol{x}\in\mathcal{B}_\eta} \varphi(\boldsymbol{x}), \max_{\boldsymbol{x}\in\mathcal{B}_\eta} \varphi(\boldsymbol{x})\right]$, define $k^*(\mu) := \text{argmin}_k |\mu - \mu_{\sigma(k)}|$ and define $\widehat{m}_\gamma(\mu) := \max\{q \in Q_{k^*(\mu)}\}$. Using the rolling maximum quantile is one possible way to estimate $\bar{q}_{\gamma,\eta}(\mu)$. One could also bin the quantiles by functional value, compute the maximum predicted quantile in each bin, and then fit a nonparametric regression to fit the maximum binned quantiles to the functional values.

This estimator choice for $\widehat{m}_\gamma(\mu)$ affects how one computes $C^{\text{sl}}_{\text{opt}}(\boldsymbol{y})$. To see why, note that we can re-express $C^{\text{sl}}_{\text{opt}}(\boldsymbol{y})$ as follows:

$$C^{\text{sl}}_{\text{opt}}(\boldsymbol{y}) := \min/\max \left\{\varphi(\boldsymbol{x}) : \boldsymbol{x} \in D(\widehat{m}_\gamma(\varphi(\boldsymbol{x})) + s(\boldsymbol{y})^2, \boldsymbol{y})\right\}. \tag{3.29}$$

As such, one could substitute the estimated $\widehat{m}_\gamma(\cdot)$ into each endpoint optimization. However, this estimated curve is likely not convex in $\boldsymbol{x}$ and therefore complicates the optimizations. One could also pursue a root-finding approach to find the set of μ such that $\lambda(\mu, \boldsymbol{y}) = \widehat{m}_\gamma(\mu)$ since these intersection points define a set of accepted functional values. This approach can also be complex in proportion to the complexity of the estimated curve. In our view, the most pragmatic approach is to simply determine all μ_k such that $\lambda(\mu_k, \boldsymbol{y}; \mathcal{X}) \leq \widehat{m}_\gamma(\mu_k)$, and then define the endpoints of $C^{\text{sl}}_{\text{opt}}(\boldsymbol{y})$ to be the minimum and maximum of those accepted sampled values. The consistency of the computed endpoints for both the inverted and optimized sliced intervals is proven in Theorem 3.4.1. Similar to the global interval constructions, the consistency of both constructions to $C^{\text{sl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta)$ also establishes the asymptotic equivalent of the inverted and optimized approaches.

Although both Algorithms 2 and 3 can produce a quantile estimate, it is worth noting a few key differences between the two. First, Algorithm 2 involves a nested sampling loop. Thus, any statistical guarantee regarding the validity of its output has the two moving parts of the accuracy of $\widehat{q}^k_\gamma$ as N gets large and the proximity of the approximated quantile to the true maximum quantile function value over $\mathcal{B}_\eta$ as M gets large. By contrast, Algorithm 3 only has one sampling loop and estimates the

Table 3.1: Summary of methods based on global and sliced max quantile, and whether they are optimization-based or inversion-based.

Category	Global max quantile	Sliced max quantile
Inversion-based methods	Interval (3.26)	Interval (3.28)
Optimization-based methods	Interval (3.25)	Interval (3.27)

full quantile function surface over $\mathcal{B}_\eta$, producing an estimate of the maximum as a consequence, but is reliant on smoothness to do an accurate regression with finitely many data points.

3.4 Theoretical justification

In this section, we provide convergence results for the interval constructions of the previous section. We prove that, under certain assumptions, whenever Algorithm 2 or Algorithm 3 are used to estimate the maximum quantiles, the global interval constructions (3.25) and (3.26) converge in probability to the true global interval $C_\alpha^{\mathrm{gl}}(\boldsymbol{y}; \mathcal{B}_\eta)$ in (3.24), and the sliced interval constructions (3.27) and (3.28) converge in probability to the true sliced interval $C_\alpha^{\mathrm{sl}}(\boldsymbol{y}; \mathcal{B}_\eta)$ in (3.23). Although our approach involving quantile regression is inspired by the approaches in (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Niccolo Dalmasso et al., 2024; Masserano, Dorigo, et al., 2023) and therefore requires similar theoretical results to connect the consistency of the quantile estimation with the interval validity, our approach is sufficiently different and requires novel theoretical insights. First, the inversion of a hypothesis test with a composite null adds a layer of complexity to the proofs. (Niccolo Dalmasso et al., 2024) does address composite null hypotheses, but in the context of nuisance parameters where they use a profile likelihood approach. Second and more fundamentally, (Niccolo Dalmasso et al., 2024; Masserano, Dorigo, et al., 2023) construct confidence sets which are shown to achieve the desired coverage level asymptotically in the number of samples used to train the quantile regressor. By contrast, we show that our computed intervals converge in probability to the theoretical Intervals (3.23) and (3.24) which achieve the correct coverage by definition. Third and finally, since we use the sampled points to invert the interval, it is insufficient to show that the cutoff is consistent as the results in (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Niccolo Dalmasso et al., 2024; Masserano, Dorigo, et al., 2023) show. Our inverted intervals require proof that the samplers can get arbitrarily close to the true endpoint boundaries.

Throughout the following results, we assume that for fixed $\gamma \in (0,1)$, $Q_{\boldsymbol{x}}(1-\gamma)$ is a continuous function of $\boldsymbol{x}$. We refer the reader to (Kibzun and Kan, 1997) for a full analysis of the properties of parametrized quantile functions. We furthermore assume that $\mathcal{B}_\eta$ is a compact set without isolated points, so that the sampling methods eventually sample close to every point.

Theorem 3.4.1. *Assume that $Q_{\boldsymbol{x}}(1-\gamma)$ is a continuous function of $\boldsymbol{x}$, and let $\mathcal{B}_\eta$ be compact and without isolated points. Let the quantile regression in Algorithm 2 be consistent for all $\boldsymbol{x}$, i.e, such that $\mathbb{P}(|\widehat{q}_\gamma(\boldsymbol{x}) - Q_{\boldsymbol{x}}(1-\gamma)| > \varepsilon) \to 0$ as $M_{tr} \to \infty$ is satisfied $\forall \varepsilon > 0$. We will write $\xrightarrow{\text{p}}$ for convergence in probability, understood as $N, M \to \infty$ if Algorithm 1 is used and as $M_{tr}, M \to \infty$ if Algorithm 2 is used. For either algorithm, we have, for a given observation $\boldsymbol{y}$:*

1. $C^{\text{gl}}_{\text{inv}}(\boldsymbol{y}) \xrightarrow{\text{p}} C^{\text{gl}}_{\alpha}(\boldsymbol{y})$,
2. $C^{\text{sl}}_{\text{inv}}(\boldsymbol{y}) \xrightarrow{\text{p}} C^{\text{sl}}_{\alpha}(\boldsymbol{y})$.

Further assume that there exists a point $\bar{\mu} \in \varphi(\mathcal{B}_\eta)$ satisfying $\lambda(\bar{\mu}, \boldsymbol{y}; \mathcal{X}) < \bar{q}_{\gamma,\eta}$, and then

3. $C^{\text{gl}}_{\text{opt}}(\boldsymbol{y}) \xrightarrow{\text{p}} C^{\text{gl}}_{\alpha}(\boldsymbol{y})$.

Finally, further assuming technical conditions discussed in Section 3.9, and then

4. $C^{\text{sl}}_{\text{opt}}(\boldsymbol{y}) \xrightarrow{\text{p}} C^{\text{sl}}_{\alpha}(\boldsymbol{y})$.

Proof. See Section 3.9. □

Note that if $\lambda(\mu, \boldsymbol{y})$ is convex in μ, as it is for linear forward models and quantities of interest (Batlle, Stanley, et al., 2023, Proposition 2.5), the condition of the existence of $\bar{\mu}$ satisfying $\lambda(\bar{\mu}, \boldsymbol{y}; \mathcal{X}) < \bar{q}_{\gamma,\eta}$ is equivalent to the interval not being empty. For nonlinear forward models, it is a slightly stronger condition. For the convergence of the sliced optimized version, one needs to show that $\inf_{\mu:\lambda(\mu,\boldsymbol{y})\le\widehat{m}_\gamma(\mu)} \mu$ converges to $\inf_{\mu:\lambda(\mu,\boldsymbol{y})\le m_\gamma(\mu)} \mu$ as $\widehat{m}_\gamma(\mu)$ converges to $m_\gamma(\mu)$. In order to do so, we study the convergence of optimization problems of the form $\inf_{\mu:\widehat{f}(\mu)\ge 0} \mu$ to $\inf_{\mu:f(\mu)\ge 0} \mu$ as $\widehat{f}$ converges to f. Although the result is not true in general, we provide sufficient technical conditions about f and the uniformness of convergence of $\widehat{f}$ to f for the result to hold. We discuss the details in Section 3.9.

3.5 Implementation methodology

Sampling the pre-image Berger–Boos set

The viability of this method directly relies upon our ability to sample from the Berger–Boos set, $\mathcal{B}_\eta = f^{-1}(\Gamma_\eta(\boldsymbol{y})) \cap \mathcal{X}$. Under the assumed data generating process in Equation (3.19), the Berger–Boos set is defined as follows:

$$\mathcal{B}_\eta = f^{-1}(\Gamma_\eta(\boldsymbol{y})) \cap \mathcal{X} = \left\{\boldsymbol{x} \in \mathcal{X} : (\boldsymbol{y} - f(\boldsymbol{x}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - f(\boldsymbol{x})) \leq \chi^2_{n,\eta}\right\}, \quad (3.30)$$

where we have generalized to a non-identity covariance matrix, $\boldsymbol{\Sigma}$, to make the following exposition more general. This set is equivalent to the set over which the strict bounds intervals are optimized in (Philip B. Stark, 1992a), also called "SSB" intervals in (Stanley, Patil, and Kuusela, 2022; Batlle, Stanley, et al., 2023). Note, however, that one would use $\chi^2_{n,\alpha}$ instead of $\chi^2_{n,\eta}$ for $1-\alpha$ interval computation in that scenario.

We discuss sampling $\mathcal{B}_\eta$ in the linear-Gaussian case, where $f(\boldsymbol{x}) = \boldsymbol{K}\boldsymbol{x}$, as this scenario aligns with the numerical experiments presented in Section 3.6. We present two sampling approaches: the "Voelker-Gossman-Stewart" (VGS) algorithm based on (Voelker, Gosmann, and Stewart, 2017) for efficiently sampling ellipsoids uniformly at random in low-dimensional scenarios and an MCMC-based algorithm we call the Polytope sampler based on the Vaidya walk presented in (Yuansi Chen et al., 2018). When f is linear and has full column rank, $\mathcal{B}_\eta$ is an ellipsoid intersected with the constraint set, making the VGS algorithm an effective option in low-dimensional settings (see Section 3.5). In all other scenarios, especially higher dimensional ones, the Polytope sampler is a better option, as naive accept-reject algorithms become intractable (see Section 3.5).

VGS Sampler for low-dimensional and full column rank settings

Under the linear-Gaussian assumptions, when $\boldsymbol{K}$ has full column rank, it can be shown that

$$\mathcal{B}_\eta = \left\{\boldsymbol{x} \in \mathcal{X} : (\boldsymbol{x} - \widehat{\boldsymbol{x}})^\top \boldsymbol{K}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{K} (\boldsymbol{x} - \widehat{\boldsymbol{x}}) \leq \chi^2_{n,\eta}\right\}, \quad (3.31)$$

where $\widehat{\boldsymbol{x}}$ is the Generalized Least-Squares (GLS) estimator. Equation (3.31) describes an ellipsoid in $\mathbb{R}^p$ intersected with $\mathcal{X}$ with axis directions and lengths determined by $\chi^2_{n,\eta}$ and the eigenvectors and eigenvalues of $\boldsymbol{K}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{K}$, respectively (see (Burt W. Rust and Walter R. Burrus, 1972) for more details). An ellipsoid is nothing but a deformed ball. As such, we can sample uniformly at random from the ellipsoid in $\mathbb{R}^p$ using an algorithm sampling uniformly at random from a p–ball, followed

by an appropriate linear transformation and translation. We can then include an additional accept-reject step to account for the constraint set $\mathcal{X}$. Note, all subsequent discussions of spheres and balls assume unit radii and centering at the origin. Also note that because the ellipsoid in (3.31) is centered around the GLS estimator, we can sample points first from within an ellipsoid of the same shape centered at the origin, and then translate those points by $\widehat{\boldsymbol{x}}$.

(Voelker, Gosmann, and Stewart, 2017) propose a particularly efficient and clever algorithm to sample uniformly at random from the p–ball by proving a connection between uniform sampling on the $(p+1)$-sphere (i.e., the surface of the ball in $\mathbb{R}^{p+2}$) and uniform sampling within the p–ball (i.e., *within* the unit ball in $\mathbb{R}^p$). Namely, one can sample a Gaussian $\widetilde{\boldsymbol{z}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p+2})$ and normalize $\boldsymbol{z} := \widetilde{\boldsymbol{z}}/\|\widetilde{\boldsymbol{z}}\|_2$ to sample a point uniformly at random from the $(p+1)$-sphere. Next, one simply drops the last two elements of $\boldsymbol{z}$ to obtain a point that is uniformly sampled from within a p–ball. The validity of this approach is substantiated by Lemma 1 and Theorem 1 in (Voelker, Gosmann, and Stewart, 2017) and relies upon a distribution result about the ratio of chi-squared distributions and preservation of distribution under an orthogonal transformation. To denote sampling a point following this procedure, we use the notation $\boldsymbol{x} \sim \mathrm{VGS}(p)$.

To sample from our desired ellipsoid in (3.31), first consider the eigendecomposition of $\boldsymbol{K}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{K} = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$, where $\boldsymbol{\Omega} = \mathrm{diag}\left(\omega_1^2, \omega_2^2, \ldots, \omega_p^2\right)$, ω_i is the i-th eigenvalue of $\boldsymbol{K}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{K}$ and $\boldsymbol{P}$ is an orthonormal matrix where the columns vectors are the eigenvectors of $\boldsymbol{K}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{K}$. Denote $\boldsymbol{\Omega}^{1/2} = \mathrm{diag}\left(\omega_1, \omega_2, \ldots, \omega_p\right)$ and by extension, $\boldsymbol{\Omega}^{-1/2} = \mathrm{diag}\left(\omega_1^{-1}, \omega_2^{-1}, \ldots, \omega_p^{-1}\right)$ when $\omega_i > 0$ for all i. These decompositions imply the correct transformation to apply to points sampled from the p–ball via $\boldsymbol{x} \sim \mathrm{VGS}(p)$. Namely, define $\boldsymbol{w} := \sqrt{\chi^2_{n,\eta}}\boldsymbol{P}\boldsymbol{\Omega}^{-1/2}\boldsymbol{x}$. To know that $\boldsymbol{w}$ is sampled from the correct ellipsoid, it should be the case that $\boldsymbol{w}^\top\boldsymbol{K}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{K}\boldsymbol{w} \leq \chi^2_{n,\eta}$, as then we can simply make the update $\boldsymbol{w} \leftarrow \boldsymbol{w} + \widehat{\boldsymbol{x}}$ to ensure that we have a point sampled from (3.31). This guarantee is verified as follows:

$$\begin{aligned}
\boldsymbol{w}^\top\boldsymbol{K}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{K}\boldsymbol{w} &= \boldsymbol{w}^\top\boldsymbol{P}\boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{P}^\top\boldsymbol{w} \\
&= \chi^2_{n,\eta} \cdot \boldsymbol{x}^\top\boldsymbol{\Omega}^{-1/2}\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{P}^\top\boldsymbol{P}\boldsymbol{\Omega}^{-1/2}\boldsymbol{x} \\
&= \chi^2_{n,\eta} \cdot \boldsymbol{x}^\top\boldsymbol{x} \leq \chi^2_{n,\eta},
\end{aligned} \tag{3.32}$$

where $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}$ by definition and the last line follows since we know $\boldsymbol{x}$ is sampled from the p–ball and therefore $\boldsymbol{x}^\top\boldsymbol{x} \leq 1$. Equation (3.32) justifies that the $\boldsymbol{w}$ samples lie within the desired ellipsoid, and their uniform distribution follows from the

uniform distribution of $\boldsymbol{x}$ in the p–ball since the distribution is preserved under a linear transformation. Finally, we accept $\boldsymbol{w}$ if $\boldsymbol{w} \in \mathcal{X}$ and reject if $\boldsymbol{w} \notin \mathcal{X}$. This procedure for sampling at random from $\mathcal{B}_\eta$ is summarized in Algorithm 4.

Algorithm 4 VGS Sampler

Input: $p \in \mathbb{N}$, $M \in \mathbb{N}$, $\boldsymbol{P}$, $\boldsymbol{\Omega}^{-1/2}$, $\widehat{\boldsymbol{x}}$, $\chi^2_{n,\eta}$.

1: Define $\mathcal{S} := \{\}$ to be the initialized set in which the sampled points are to be placed.
2: **for** $k = 1, 2, \ldots, M$ **do**
3: **Sample from the p–ball using VGS:** $\boldsymbol{x}_k := z_{1:p}/\|z\|_2$, where $z \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{p+2})$ and $z_{1:p}$ denotes taking the first through p-th indices (inclusive).
4: **Transform VGS output:** $\boldsymbol{w}_k := \sqrt{\chi^2_{n,\eta}}\boldsymbol{P}\boldsymbol{\Omega}^{-1/2}\boldsymbol{x}_k$.
5: **Translate $\boldsymbol{w}_k$ by the GLS estimator:** $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k + \widehat{\boldsymbol{x}}$.
6: **Accept-Reject to incorporate constraints**: If $\boldsymbol{w}_k \in \mathcal{X}$, then add $\mathcal{S} \leftarrow \mathcal{S} \cup \{\boldsymbol{w}_k\}$, else start loop iteration k again.
7: **end for**

Output: $\mathcal{S}$ containing uniformly sampled points over $\mathcal{B}_\eta$ (as defined in (3.31)).

Generating samples using the VGS Sampler is efficient, but its feasibility diminishes in high dimensions because of the accept-reject step. To illustrate this point, consider the simple scenario where $\mathcal{X} = \mathbb{R}^p_+$, i.e., the non-negative orthant of $\mathbb{R}^p$ and suppose we sample from the p–ball intersected with $\mathbb{R}^p_+$ by sampling $\boldsymbol{x} \sim \text{VGS}(p)$ where $\boldsymbol{P} = \boldsymbol{\Omega} = \boldsymbol{I}$, $\widehat{\boldsymbol{x}} = \boldsymbol{0}$ and we use 1 instead of $\chi^2_{n,\eta}$. Then, $\mathbb{P}(\boldsymbol{x} \in \mathbb{R}^p_+) = 2^{-p}$ and the acceptance probability of each sample goes to zero exponentially in p. To make this point slightly more general, we generate data $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}^*_p, \boldsymbol{I}_p)$ where $\boldsymbol{x}^*_p \in \mathbb{R}^p_+$ and is a vector of ones. For a collection of dimensions $p \in [2, 30]$, we sample $\boldsymbol{x} \sim \text{VGS}(p)$ with $\boldsymbol{P} = \boldsymbol{\Omega} = \boldsymbol{I}$ and $\widehat{\boldsymbol{x}} = \boldsymbol{y}$ to estimate the probability that $\boldsymbol{x}$ is in $\mathbb{R}^p_+$ and plot the results in the left panel of Figure 3.2. Since the acceptance probability decays exponentially (under 10^{-4} for 30 dimensions), it becomes clear that this algorithm is inefficient in dimensions larger than 10, since the acceptance probability quickly becomes prohibitively small in regimes where a larger sample size is even more important.

Dimension is only one of two primary complicating factors. Not only is there less of a (potentially) shifted p–ball's volume in the non-negative orthant as p grows but if our data are generated via $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{K}\boldsymbol{x}^*, \boldsymbol{I}_p)$ with $\boldsymbol{x}^* \in \mathbb{R}^p_+$ where the condition number of $\boldsymbol{K}$ is large, it is possible that even less of the ellipsoid from which the VGS Sampler draws points intersects with the parameter constraint. To illustrate this point, we generate a single observation from the aforementioned

model using a $\boldsymbol{K} \in \mathbb{R}^{40\times 40}$ as described in Section 3.6. The $\boldsymbol{x}^* \in \mathbb{R}^{40}_+$ is created in the same way as that of Section 3.6. The right panel of Figure 3.2 shows a computed probability mass function for the number of coordinates in a VGS Sampler draw (with $\boldsymbol{P}$ and $\boldsymbol{\Omega}$ defined such that $\boldsymbol{K}^\top \boldsymbol{K} = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$) complying with the non–negativity parameter constraints. Equivalently, the right panel of Figure 3.2 shows the computed probability mass function for the number of coordinates lying within the Berger–Boos set. For this particular setup, we critically note that out of 5×10^4 draws from the VGS Sampler, none of the draws had all coordinates comply with the non-negativity constraint. By contrast, for the aforementioned noise model where we more simply sample from the p–ball intersected with the non-negative orthant, the computed acceptance probability is approximately 3.35×10^{-6}, in both cases emphasizing the VGS Sampler's poor performance in high-dimensional and ill-conditioned forward model regimes.

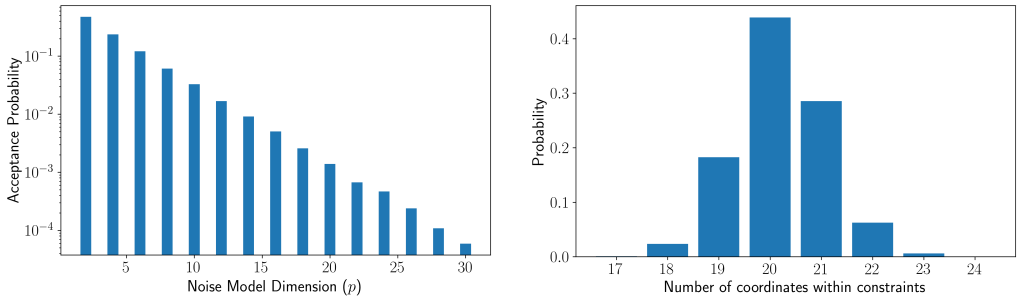


Figure 3.2: Numerical illustrations of the VGS Sampler's infeasibility in high dimensional regimes. The **left** panel shows the computed acceptance probability of a point drawn by the VGS Sampler with data generated from a non-negatively constrained Gaussian noise model. Crucially, at only 30 dimensions, the acceptance probability is already less than 10^{-4} for this particular setup. The **right** panel shows the computed probability mass function for the number of non-negative constraint complying coordinates of a VGS sample with data generated from a non-negatively constrained linear Gaussian model in 40 dimensions with a non-identity forward model. Since this is an example using a forward model with a large condition number ($\approx 1.6 \times 10^4$), we critically note that there is empirically zero probability of generating a sample within the non-negativity constraints.

Polytope sampler for general settings

In settings where the forward model is not linear and full column rank, Algorithm 4 fails. The ineffectiveness of this algorithm expands if the condition number of the linear forward model is large such that most of the pre-image ellipsoid defining the Berger–Boos set lies outside of the constraint set. Although these scenarios

induce particular geometric challenges, in the linear-Gaussian case with convex $\mathcal{X}$ we are still fundamentally sampling a convex set for which there is a vast literature. For example, there is a vast sampling literature for computing Bayesian posteriors in high dimensions via nested sampling (Skilling, 2004; Ashton et al., 2022; Buchner, 2023). In particular, nested sampling has been successfully applied in high-dimensional cosmology settings using sophisticated approaches to strategically sample the parameter space by restricting prior sampling in various ways (Buchner, 2023; Montel, Alvey, and Weniger, 2023). Although these approaches provide tools addressing a sampling setting similar to ours (i.e., the Berger–Boos set can be viewed as a portion of the parameter space defined by a cutoff on the likelihood) they are ultimately aimed at sampling from a particular distribution (i.e., the posterior), which is a stronger criterion than required here. For sampling general convex sets, simple algorithms like Hit-and-Run are available (Smith, 1984; Lovasz, 1999; Lovasz and Vempala, 2006). However, in the particular case considered here, more sophisticated and efficient algorithms can be devised. In particular, there exists a deep literature on random walks over polytopes such that the asymptotic stationary distribution of the walk is a uniform distribution over the polytope of interest (Kannan and Narayanan, 2012; Narayanan, 2016; Yuansi Chen et al., 2018). As such, we propose to first construct a bounding polytope, $\mathcal{P}^d$ composed of d half-spaces around $\mathcal{B}_\eta$, sample C random walks within the Berger–Boos set using the Vaidya walk as described in (Yuansi Chen et al., 2018) each starting at a point from a collection of strategically chosen locations, and then combine the parallel chains to create the final sample set. The detailed algorithm can be seen in Algorithm 5.

Although MCMC algorithms are typically evaluated using trace plots on individual dimensions of the parameter space, given the high-dimensionality of the problem and the final step combining several MCMC chains in the parameter space started from different positions, it is more meaningful to evaluate this algorithm's ability to sample fully from the functional space. Namely, we can solve for the largest and smallest values the functional can take within the Berger–Boos set as follows:

$$I_{\mathsf{BB}}(\boldsymbol{y}) := \left[\mu^l_{\mathsf{BB}}, \mu^u_{\mathsf{BB}}\right] = \left[\min_{\boldsymbol{x}\in\mathcal{B}_\eta} \varphi(\boldsymbol{x}), \max_{\boldsymbol{x}\in\mathcal{B}_\eta} \varphi(\boldsymbol{x})\right]. \tag{3.33}$$

When $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x}$ for some $\boldsymbol{h} \in \mathbb{R}^p$, $I_{\mathsf{BB}}(\boldsymbol{y})$ corresponds to the SSB interval in (Stanley, Patil, and Kuusela, 2022). Computing the sets $C^{\mathrm{sl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta)$ and $C^{\mathrm{gl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta)$ well is then contingent upon sampling functional values within $I_{\mathsf{BB}}(\boldsymbol{y})$ well since a functional value can only be included in the inverted set if the sampler has a non-zero probability of sampling arbitrarily close to it. By "well", we informally

mean that the sampled functional values range at least between the endpoints of $I_{\mathsf{BB}}(\boldsymbol{y})$ and that if we partition $I_{\mathsf{BB}}(\boldsymbol{y})$ into n sub–intervals, that most if not all of the sub–intervals contain at least one sample. In practice, it will often be the case that the sampled functional values can lie outside the endpoints of $I_{\mathsf{BB}}(\boldsymbol{y})$ since the MCMC chains are sampling a bounding polytope of the Berger–Boos set.

Choosing the bounding polytope. Any bounded polytope of $\mathcal{B}_\eta$ is the intersection of a finite number of half-spaces defined in $\mathbb{R}^p$. The task of constructing a bounding polytope is then equivalent to choosing a collection of d hyperplanes in $\mathbb{R}^p$ to construct a set $\mathcal{P}^d := \{\boldsymbol{x} \in \mathbb{R}^p : \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}\}$ such that $\mathcal{B}_\eta \subseteq \mathcal{P}^d$, where $\boldsymbol{A} \in \mathbb{R}^{d\times p}$ and $\boldsymbol{b} \in \mathbb{R}^d$. Let $\boldsymbol{a}_i^\top \in \mathbb{R}^p$ denote the i-th row vector of $\boldsymbol{A}$. We compute b_i (the i-th element of $\boldsymbol{b}$) as

$$b_i = \max_{\boldsymbol{x}\in\mathcal{B}_\eta} \boldsymbol{a}_i^\top \boldsymbol{x}. \tag{3.34}$$

This construction ensures the necessary inclusion. We consider three approaches to pick the vectors $\boldsymbol{a}_i$: (i) using the constraints $\mathcal{X}$, (ii) using the known eigenvectors defining the bounded directions of the pre-image ellipsoid, and (iii) randomly. In practice, we combine these approaches to ensure that we consider only parameter settings in agreement with our physical constraints, and to tighten the bounding Berger–Boos set polytope as much as possible. There is a tradeoff with respect to the latter consideration since the mixing time and computational cost for the Vaidya walk increase with the number of hyperplanes (Yuansi Chen et al., 2018).

To incorporate the known parameter constraints, consider the non-negativity constraint used in Section 3.6, i.e., $\boldsymbol{x} \in \mathbb{R}^p_+$. To enforce non-negativity, we set $\boldsymbol{a}_i = -\boldsymbol{e}_i$ for $i = 1, \ldots, p$, where $\boldsymbol{e}_i$ is defined by its i-th element set to one and the rest of its elements set to zero and $b_i = 0$. These choices produce p rows in $\boldsymbol{A}$ corresponding to the desired lower bounds (i.e., $x_i \geq 0$ for all i), but we can compute an additional p constraints using (3.34) with $\boldsymbol{a}_i := \boldsymbol{e}_i$ for $i = p+1, \ldots 2p$. These $2p$ constraints define a hyperrectangle enclosing the Berger–Boos set in the parameter space. To incorporate polytope constraints based upon the forward model, we use the ellipsoidal definition of the pre-image as shown in (3.31) and eigendecomposition of $\boldsymbol{K}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{K} = \boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^\top$ shown in (3.32). Note that this ellipsoid form is valid under the linear-Gaussian noise assumption and for all $\boldsymbol{K}$. In the event that $\boldsymbol{K}$ is not full column rank, the ellipsoid defined by Equation (3.31) is still defined via the pseudo–inverse of of $\boldsymbol{K}^\top\boldsymbol{\Sigma}\boldsymbol{K}$ but where the ellipsoid is unbounded in some directions. The column vectors of $\boldsymbol{P}$ corresponding to the non-zero entries on the diagonal of $\boldsymbol{\Omega}$ are the

eigenvectors corresponding to the bounded principal axes. As such, both $\boldsymbol{p}_i$ and $-\boldsymbol{p}_i$ for $i = 1, \ldots, p$ (the column vectors of $\boldsymbol{P}$) can be used as rows of $\boldsymbol{A}$ with their corresponding bounds defined by (3.34). Finally, to further tighten the polytope around the Berger–Boos set, we sample a multivariate Gaussian, i.e., $\boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)$ to include random hyperplanes. Note, it is possible that the unbounded directions of the ellipsoid defined by Equation (3.31) are not bounded when intersecting with the parameter constraint set and hence there is no bounding polytope. In this case, our interval construction should be unbounded since there is not enough information to produce a bounded confidence set for the quantity of interest.

Once the components $(\boldsymbol{A}, \boldsymbol{b})$ have been defined using some or all of the above hyperplane generation strategies, the Vaidya walk can immediately be employed to perform a random walk around the polytope. The primary intuitive requirement we wish to satisfy with any sampling scheme in this context is that every region of the Berger–Boos set has a non-zero probability of being sampled. Asymptotically, the Vaidya walk samples the desired polytope uniformly at random, which satisfies a stronger requirement. In practice, the need to sample non-uniformly can arise if there are particularly meaningful parameter settings in regions that are difficult for the random walk to reach. We explore such a case in Section 3.6. Additionally, although the asymptotic distribution of the Vaidya sampler is theoretically sufficient, in practice it often has some difficulty reaching the corners of the generated polytope. Although MCMC chain mixing is typically evaluated by looking at trace plots, this diagnostic is insufficient here because the dimension is high and the defined polytope can make different dimensions difficult to compare. Instead, we consider how well the sampler samples the functional values $\varphi(\boldsymbol{x})$. This motivates taking a collection of starting points between the SSB endpoints and the Chebyshev center of the polytope, as described in the following section.

Constructing the parallel chain starting points. Although the random walks defined in (Yuansi Chen et al., 2018) asymptotically sample uniformly over $\mathcal{P}^d$, given the long and thin shape of the pre-image, running the Vaidya walk from even a "good" starting position does not consistently sample the functional space well. Instead, we use the following heuristic to construct several parallel chains that constitute a complete sample when combined. We define an even number of starting points, $C \in 2\mathbb{N}$, to roughly span the Berger–Boos set, run the Vaidya walk for M_p steps from each starting point to collect parameter settings $\{\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_{M_p}\}$, and combine the samples from each walk, resulting in $C \times M_p$ total samples. In practice,

C and M_p are chosen to yield a total of M samples as desired for Algorithm 2 or Algorithm 3. Since we have designed this process to avoid the neccessity of any individual chain reaching its asymptotic distribution, we do not need a burn–in period for any chain as long as we are confident that the union of the sampled points sufficiently covers Interval (3.33) so that we may accept or reject any quantity of interest value in that interval. We define the collection of starting points using the line segments defined by the parameters generating the endpoints of $I_{\mathsf{BB}}(\boldsymbol{y})$ and the Chebyshev center of $\mathcal{P}^d$ as defined in (Boyd and Vandenberghe, 2004), and denoted by $\boldsymbol{x}_c$. This point is the center of the largest ball contained within $\mathcal{P}^d$ and therefore acts as a way to characterize the center of $\mathcal{P}^d$. We denote the parameter settings generating the endpoints of $I_{\mathsf{BB}}(\boldsymbol{y})$ by $\widehat{\boldsymbol{x}}^l$ for the lower endpoint and $\widehat{\boldsymbol{x}}^u$ for the upper endpoint. We then define a uniform grid of values $\{\tau_l\}_{l=1}^{C/2}$ such that $\tau_l \in (0,1)$ and $\tau_l < \tau_{l+1}$ for all l. For each $k \in [C]$, define $k' := \lceil k/2 \rceil$ and set $\boldsymbol{x}_k^{start} := \tau_{k'}\widehat{\boldsymbol{x}}^l + (1-\tau_{k'})\boldsymbol{x}_c$ if k is odd and $\boldsymbol{x}_k^{start} := \tau_{k'}\widehat{\boldsymbol{x}}^u + (1-\tau_{k'})\boldsymbol{x}_c$ if k is even. Creating starting positions along the lines connecting these endpoints and the Chebyshev center accommodates the chosen polytope while helping ensure that samples are chosen spanning the range of possible functional values over the Berger–Boos set.

The empirical performance of the Polytope sampler can be seen in Figure 3.3, showing (left) a histogram of the functional values sampled and (right) a trace plot for the sampled Vaidya walks starting from points constructed as described above. These plots are generated for one observation from the 80-dimensional ill-posed inverse problem in the coverage study performed in Section 3.6. Critically, the histogram shows that the Polytope sampler samples the functional space well and the trace plot shows that our starting point construction heuristic performs well in practice.

Sampling from the Berger–Boos set. With $\mathcal{P}^d$ and the starting positions defined, we construct a sample $\mathcal{S} := \{\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_{C\times M_p}\}$. We leave the details of the Vaidya walk to the original paper Yuansi Chen et al. (2018), but note an essential radius tuning parameter of the algorithm that must be chosen. This radius impacts the spread of a Gaussian proposal distribution for the walk and thus affects a new proposed point's acceptance probability. Choosing a too large radius results in a low acceptance rate of proposed steps, thus creating walk that does not mix well. In contrast, choosing a too small radius results in a high acceptance rate with relatively small step sizes. In practice, we find a radius setting of 0.5 works well as it produces

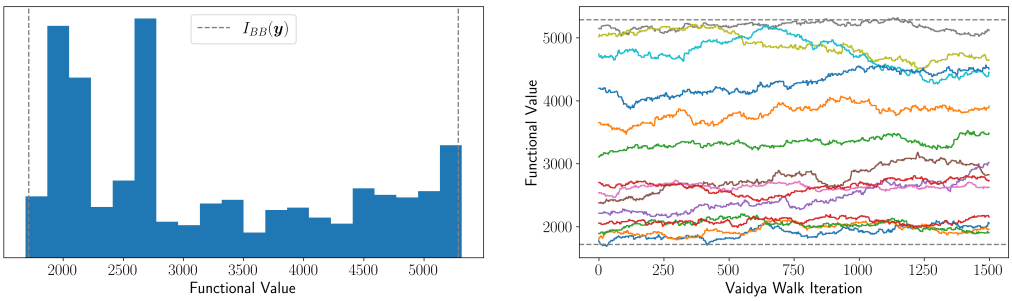


Figure 3.3: Polytope sampler output for a realization of the 80-dimensional ill-posed inverse problem studied in Section 3.6. The **left** panel contains a histogram of sampled functional values which both span and cover well the range of $I_{\mathsf{BB}}(\boldsymbol{y})$ (shown by the dashed gray lines in both plots). The **right** panel contains trace plots of the 14 Vaidya walks (each indicated by a different color) which together constitute the full sample. Our heuristic for choosing starting points along the lines connecting the parameter settings generating the endpoints of $I_{\mathsf{BB}}(\boldsymbol{y})$ and the Chebyshev center of the polytope provides a good initial spread of starting functional values.

an acceptance probability of $\approx$ 33.3%, producing a reasonable trade-off between taking meaningful steps and not rejecting too many steps.

Quantile regression

Algorithm 3 explained in Section 3.3 involves using quantile regression to learn a quantile surface from a collection of pairs of design points and samples from the LLR test statistic. As previously mentioned, similar approaches have been taken in (Niccolò Dalmasso, Izbicki, and A. B. Lee, 2020; Niccolo Dalmasso et al., 2024; Masserano, Dorigo, et al., 2023; Masserano, A. Shen, et al., 2024), and since quantile regression is a technique facilitating our interval constructions, we will only give a brief overview of quantile regression and some different ways to implement it. Fundamentally, given a one-dimension random variable $z \sim P_{\boldsymbol{x}}$ that depends on the parameter $\boldsymbol{x} \in \mathbb{R}^p$, we are interested in the upper γ-quantile at every parameter setting, i.e.,

$$\mathbb{P}_{\boldsymbol{x}}(z > Q_{\boldsymbol{x}}(1-\gamma)) = \gamma. \tag{3.36}$$

We note that the quantile surface itself is not random, so we can use draws from the distribution $P_{\boldsymbol{x}}$ to estimate $Q_{\boldsymbol{x}}$ at a given parameter setting $\boldsymbol{x}$. However, as noted in Section 3.3, performing such an estimate is not always computationally feasible, and intuitively, we might expect quantiles to vary smoothly over the parameter space, which would imply that information about a quantile at $\boldsymbol{x}_1$ should be related to a quantile at $\boldsymbol{x}_2$ if these points are close. In statistics literature, estimating $Q_{\boldsymbol{x}}$ is

Algorithm 5 Polytope sampler

Input: $M_p \in \mathbb{N}$ and $C \in 2\mathbb{N}$. $\boldsymbol{A} \in \mathbb{R}^{d\times p}$, $\boldsymbol{b} \in \mathbb{R}^d$. $\{\tau_l\}_{l=1}^{C/2}$, where $\tau_l \in (0,1)$ and $\tau_l < \tau_{l+1}$ for all l.

1: Let $\mathcal{S} := \{\}$ be the set in which we store all sampled parameter settings.
2: **Construct Chebyshev center:** Solve for $\boldsymbol{x}_c$ using the following optimization.

$$\begin{aligned} \underset{\boldsymbol{x}_c, r}{\text{maximize}} \quad & r \\ \text{subject to} \quad & \boldsymbol{a}_i^\top \boldsymbol{x}_c + r\|\boldsymbol{a}_i\|_2 \leq b_i, \quad i = 1, \ldots, d. \end{aligned}$$

3: **Compute φ extremes of Berger–Boos set:** Extreme points are computed with respect to the functional of interest:

$$\begin{aligned} \widehat{\boldsymbol{x}}^l &:= \text{argmin} \;\; \varphi(\boldsymbol{x}) \quad \text{subject to} \;\; \boldsymbol{x} \in \mathcal{B}_\eta, \\ \widehat{\boldsymbol{x}}^u &:= \text{argmax} \;\; \varphi(\boldsymbol{x}) \quad \text{subject to} \;\; \boldsymbol{x} \in \mathcal{B}_\eta. \end{aligned} \tag{3.35}$$

4: **for** $k = 1, 2, \ldots, C$ **do**
5: **Construct starting point:** Define $k' := \lceil k/2 \rceil$. Define $\boldsymbol{x}_k^{start} = \tau_{k'}\widehat{\boldsymbol{x}}^l + (1 - \tau_{k'})\boldsymbol{x}_c$ if k is odd and $\boldsymbol{x}_k^{start} = \tau_{k'}\widehat{\boldsymbol{x}}^u + (1 - \tau_{k'})\boldsymbol{x}_c$ if k is even.
6: **Run the Vaidya walk for M_p steps:** Collect samples $\{\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_{M_p}\}$ and add them to $\mathcal{S}$.
7: **end for**

Output: $\mathcal{S}$ containing sampled points over $\mathcal{B}_\eta$.

framed as estimating a quantile function conditional on known covariates and is often thought of as a generalization of estimating the conditional median (Roger Koenker and Hallock, 2001). Just as conditional mean and conditional median estimation can be accomplished by using an appropriate loss function (sum of squares and absolute differences, respectively), estimating conditional quantiles can be accomplished by minimizing the pinball loss defined as follows:

$$L_\gamma(z, q) := \begin{cases} (1-\gamma)(q - z), & z < q, \\ -\gamma(z - q), & z \geq q, \end{cases} \tag{3.37}$$

(Roger Koenker, 2005; Steinwart and Christmann, 2011). As such, estimating the quantile surface can be framed as a risk-minimization problem, leaving only standard modeling choices to fill in for q in (3.37). Although initial efforts were focused on linear parametric quantile regressors (R. Koenker and Bassett Jr, 1978), in recent years, modeling efforts have focused on nonparametric varieties. (Meinshausen,

2006) adapted random forests to quantile regression. (Takeuchi et al., 2006) leveraged Reproducing Kernel Hilbert Spaces to construct smooth quantile regressors. Closer to our application, (Masserano, Dorigo, et al., 2023) used neural networks to optimize the pinball loss to learn the quantile surface for their application.

Using the design points in the Berger–Boos set as sampled via the VGS or Polytope samplers, Algorithm 3 shows how we sample from the test statistic distribution defined at each design point to define a data set to fit a quantile regressor. We use the Gradient Boosting Regressor implemented in scikit-learn (Pedregosa et al., 2011) with the "quantile" loss function (i.e., pinball loss defined in eq. (3.37)) to fit the quantile surface for our numerical examples in Section 3.6. This algorithm involves a collection of hyperparameters (i.e., the minimum number of samples required to split an internal node, the minimum number of samples required to be a leaf node, the maximum depth of any individual estimator, the learning rate, and the number of estimators) which we determine using 10-fold cross validation in a pilot study ahead of our simulation experiments in Section 3.6. Although one may use any quantile regression approach to estimate the quantile surface, we emphasize the importance of choosing an approach that can accommodate a nonlinear surface in the parameters (such as the Gradient Boosting Regressor) as the parameter constraints are known to produce nonlinear quantile surfaces in even simple examples as seen in (Batlle, Stanley, et al., 2023).

3.6 Numerical experiments

For scenarios within the linear-Gaussian case of data-generating process (3.19), the OSB interval can be regarded as the previous the state-of-the-art option for computing constraint-aware confidence intervals. Although these intervals have empirically achieved nominal coverage in applications (Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022), they generally do not guarantee coverage (Batlle, Stanley, et al., 2023). As such, we use the OSB interval in the following numerical experiments as a main comparison point to the intervals defined in this paper. In scenarios where the OSB interval achieves at least nominal coverage, we show that our intervals are either competitive or better in terms of expected interval length. In scenarios where the OSB interval does not achieve nominal coverage, our intervals do achieve nominal coverage and can have shorter expected length. We provide four numerical experiments to make these points. The first set of two uses a constrained Gaussian noise model setup in two or three dimensions. These two experiments illustrate the aforementioned points in addition to constituent parts of

the interval computation process due to the relatively low dimensions. The second set of two considers a wide-bin deconvolution setup inspired by particle unfolding in high-energy physics (Stanley, Patil, and Kuusela, 2022). This setup features an 80-dimensional parameter space with a rank-deficient forward model and is thus a substantially more complicated computational scenario compared to the two or three-dimensional Gaussian noise models. These examples demonstrate the superior performance of our intervals over OSB in terms of both coverage and expected length.

In the following experiments, we form 68% confidence intervals, set $\eta = 0.01$ and compute γ according to Lemma 3.3.1. We draw 10^3 observations from each data-generating process to estimate both coverage and expected interval length for our four interval constructions and the OSB interval. We additionally provide 95% confidence intervals in the form of orange line segments to characterize statistical error for both coverage and expected length estimates. The coverage confidence intervals are Clopper-Pearson intervals for the success probability parameter of a binomial distribution, while the expected length confidence intervals are the average length plus/minus the appropriately scaled standard error of the mean.

Constrained Gaussian in two dimensions

The two-dimensional Gaussian noise model is defined as follows:

$$\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_2), \quad \boldsymbol{x}^* \in \mathbb{R}^2_+, \tag{3.38}$$

where $\varphi(\boldsymbol{x}) = x_1 - x_2$ and $\boldsymbol{x}^* = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}^\top$. The LLR is then given as follows:

$$\lambda(\mu, \boldsymbol{y}) = \min_{\substack{x_1 - x_2 = \mu \\ \boldsymbol{x} \in \mathbb{R}^2_+}} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \in \mathbb{R}^2_+} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{3.39}$$

This example first appeared in (Tenorio, Fleck, and Moses, 2007) as a case where the OSB interval allegedly fails to achieve nominal coverage when the true parameter $\boldsymbol{x}^*$ is such that $\varphi(\boldsymbol{x}^*) = 0$. However, (Batlle, Stanley, et al., 2023) overturned this result by proving OSB validity in this case. As such, this example is important to include because of its historical context and OSB interval validity. The proof that the OSB interval covers in this particular example relies upon showing that $Q^{\max}_{1-\alpha} = \chi^2_{1,\alpha}$ for all $\alpha \in (0, 1)$, where $\chi^2_{1,\alpha}$ is the upper α-quantile of a chi-squared distribution with one degree of freedom. Alternatively stated, it holds that $\lambda(\varphi(\boldsymbol{x}^*), \boldsymbol{y}; \mathbb{R}^2_+)$ is stochastically dominated by χ^2_1. This result is shown in Lemma 4.4 of (Batlle, Stanley, et al., 2023).

Estimated coverage and length results are shown in Figure 3.4. We note that all four of our interval constructions are competitive with OSB in terms of coverage, while

all of our interval constructions have higher estimated expected length, apart from the Sliced constructions which are within statistical error of OSB. Since the OSB interval is defined using $Q_{1-\alpha}^{\max} = \chi^2_{1,\alpha}$ and the α-quantile surface of the LLR rapidly approaches this global max-quantile as one moves away from the origin (see Figure 5.3 in (Batlle, Stanley, et al., 2023)), the OSB interval lengths are difficult to beat in practice with intervals based on the Berger–Boos sets since these sets likely contain parameter settings with quantiles near $\chi^2_{1,\alpha}$. The left panel of Figure 3.5 shows four realizations of the data-generating process with the observations shown as red points. For each observation, the blue points show uniformly distributed draws within its Berger–Boos set, sampled using the VGS sampler. Cross-referencing the spread of the Berger–Boos set samples in Figure 3.5 with Figure 5.3 in (Batlle, Stanley, et al., 2023), it is clear that there are always samples in the parameter space where the quantile surface is nearly the same as the $\chi^2_{1,\alpha}$ quantile. Further, when including the Berger–Boos set in the interval construction, we instead construct our intervals using the γ-quantile, where $\gamma < \alpha$, as the LLR cutoff, resulting in a more relaxed constraint. This fact can be clearly observed in the central panel of Figure 3.5, showing the sampled γ-quantiles within the Berger–Boos set of one observation from the data generating process. Since a non-trivial portion of this distribution is above $\chi^2_{1,\alpha}$, the longer average length of the Global intervals is explained. In the right panel of Figure 3.5, for the same observation, we show the estimated sliced max-quantile function, $\widehat{m}_\gamma(\mu)$, in orange alongside $\chi^2_{1,\alpha}$. Since this estimated function is above $\chi^2_{1,\alpha}$ at their intersection points with the underlying LLR function shown in the solid blue line, it further makes sense that the Sliced interval constructions provide no additional length improvement compared to the OSB interval in this particular setting.

Constrained Gaussian in three dimensions

As seen in the previous example, in a case where the OSB interval is known to achieve nominal coverage, its expected length can be difficult to beat. However, OSB coverage guarantee can be difficult to prove or disprove, since it amounts to proving stochastic dominance on the non-trivial LLR statistic. In such situations, our intervals immediately provide a clear theoretical advantage. One such case involving a three-dimensional constrained Gaussian case was explored in (Batlle, Stanley, et al., 2023), to which we now apply our four interval constructions. The three-dimensional Gaussian noise model is defined as follows:

$$\boldsymbol{y} = \boldsymbol{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_3), \quad \boldsymbol{x}^* \in \mathbb{R}^3_+, \tag{3.40}$$

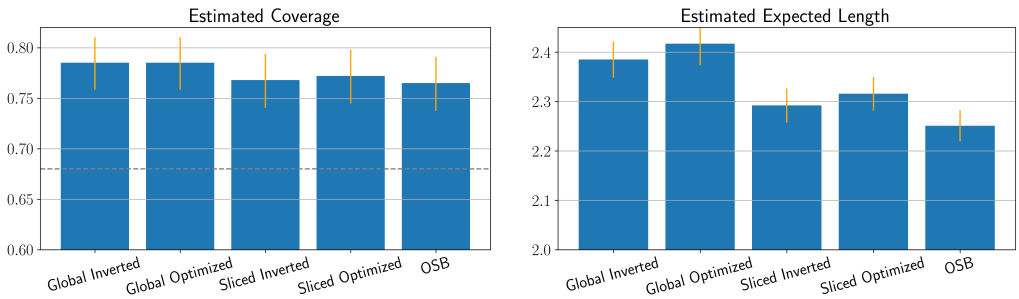


Figure 3.4: Estimated coverages and expected lengths across all four interval constructions and OSB for comparison at the 68% level for the two-dimensional constrained Gaussian setting. All four of our interval constructions are comparable to OSB with respect to coverage, but OSB shows better-expected length performance, aside from our two sliced interval constructions. Although the OSB intervals are defined using the global max-quantile ($Q^{\max}_{1-\alpha}$) and therefore can potentially be improved upon by limiting the considered parameter space via the Berger–Boos set, due to the rapidity with which the α-quantile surface meets the $\chi^2_{1,\alpha}$ quantile (see Figure 5.3 in (Batlle, Stanley, et al., 2023)), the OSB interval lengths are difficult to beat in practice.

where $\varphi(\boldsymbol{x}) = x_1 + x_2 - x_3$ and $\boldsymbol{x}^* = \begin{pmatrix} 0.03 & 0.03 & 1 \end{pmatrix}^\top$. The LLR is then defined as follows:

$$\lambda(\mu, \boldsymbol{y}) = \min_{\substack{x_1+x_2-x_3=\mu \\ \boldsymbol{x}\in\mathbb{R}^3_+}} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x}\in\mathbb{R}^3_+} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{3.41}$$

Figure 3.6 shows estimated coverage and expected length across all four interval constructions and the OSB interval. While the OSB interval fails to attain nominal coverage in this example, all four of our interval constructions do, with the Sliced constructions providing the best calibration. While the Global constructions pay a fairly steep price for coverage in expected interval length, the Sliced constructions navigate the trade-off well, paying for coverage with only slightly longer intervals compared to the OSB interval.

The setting of $\boldsymbol{x}^*$ used here is slightly different than that of (Batlle, Stanley, et al., 2023), where $\boldsymbol{x}^* = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^\top$ was used. In (Batlle, Stanley, et al., 2023), this setting was used as a counter-example for OSB coverage, since $Q_{\boldsymbol{x}^*}(1-\alpha) > \chi^2_{1,\alpha}$ for at least some $\alpha \in (0, 1)$. However, as shown in Figure 5.5 of (Batlle, Stanley, et al., 2023), when $\alpha = 0.05$, $Q_{\boldsymbol{x}^*}(1-\alpha) \leq \chi^2_{1,\alpha}$ for $\boldsymbol{x}^* = \begin{pmatrix} t & t & 1 \end{pmatrix}^\top$ when t is approximately greater than $e^{-2} \approx 0.135$, which indicates that parameter settings violating stochastic dominance by χ^2_1 exist close to the parameter constraint boundary. The location of these key parameter settings presents a challenge for the

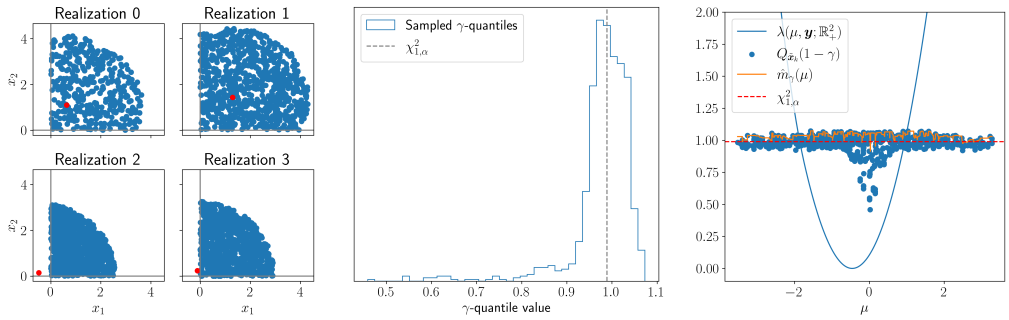


Figure 3.5: (**Left**) Four realizations of the data-generating process where the observations are shown in red. For each realization, the blue points are uniformly distributed samples from its Berger–Boos set, sampled using the VGS sampler. (**Center**) For a realization of the data-generating process, we plot the distribution of γ-quantiles for the points sampled by the VGS sampler. Notably, a non-trivial percent of these are above $\chi^2_{1,\alpha}$ defining the OSB interval. (**Right**) For the same realization, we plot the estimated sliced max-quantile function, $\widehat{m}_\gamma(\mu)$ in orange alongside $\chi^2_{1,\alpha}$ in red. The blue points correspond to sampled parameter values, each of which has a functional and quantile value, while the solid blue line shows the LLR over the functional varies. All intervals can be read immediately from this image by inspecting where the blue LLR curve intersects the sampled points.

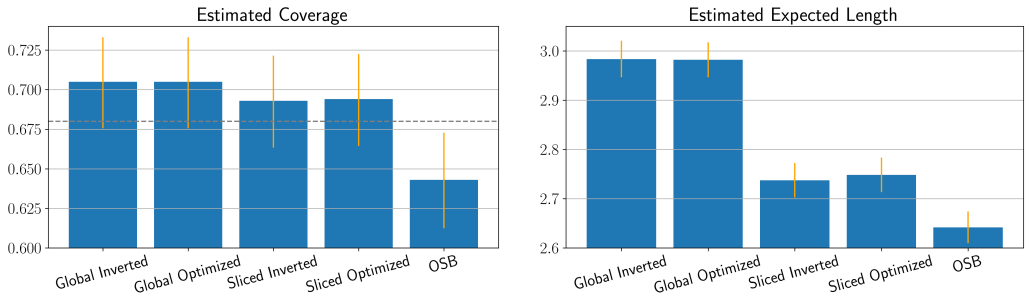


Figure 3.6: Estimated coverage and expected length across all four interval constructions and OSB for comparison at the 68% level for the three-dimensional constrained Gaussian example. All four of our interval constructions achieve nominal coverage while the OSB interval does not. While the Global interval constructions pay a steep price in expected length compared to OSB, the Sliced constructions are only slightly longer than OSB.

Polytope sampler described by Algorithm 5. In Section 3.10, Algorithm 6 presents a modified version of Algorithm 5 that better handles sampling in this example.

Berger–Boos set experiment. For this model, we investigate the effect of changing the parameter η that controls the Berger–Boos construction in the global interval. We compute intervals for different $\boldsymbol{y}$ and fixed $\boldsymbol{x}^* = (2, 2, 0), \boldsymbol{x}^* = (3, 3, 0)$ and

$\boldsymbol{x}^* = (5, 5, 0)$ as η ranges between 0 and $\alpha = 0.32$. The lengths of such intervals, averaged over the data $\boldsymbol{y}$, are shown in Figure 3.7. In this example the maximum quantile is achieved at $\boldsymbol{x} = (0, 0, t)$ for large t, and, as expected, the benefit of using a small $\eta > 0$ becomes more pronounced as the true $\boldsymbol{x}^*$ becomes farther from the point that achieves the maximum quantile. However, as η grows too close to α the downside of optimizing the $1 - \alpha + \eta$ quantile instead of $1 - \alpha$ outweighs the benefit of optimizing it over a smaller set, resulting in larger intervals. This suggests that a small $\eta > 0$ is a reasonable default, as suggested originally by (Roger L. Berger and Boos, 1994).

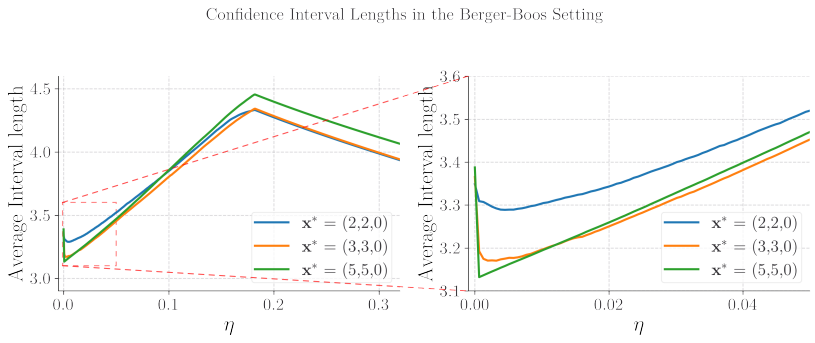


Figure 3.7: Confidence interval lengths in the Berger–Boos setting, averaged over values of $\boldsymbol{y}$, for varying η and $\boldsymbol{x}^*$. The minimum average length occurs at a small $\eta > 0$, showing that the construction is beneficial if η is tuned correctly. This occurs because even for moderately small η, the Berger–Boos set, which is a three-dimensional sphere intersected with the non-negative orthant, avoids the point with the highest $1 - \alpha$ quantile.

Wide-bin deconvolution

While the numerical experiments in Sections (3.6) and (3.6) show that our interval constructions are competitive with the OSB interval in scenarios where it is known to provide coverage and superior to the OSB interval by achieving nominal coverage when OSB does not, this section shows the superior performance of our interval constructions relative to OSB in a more complex high-dimensional setting. We consider the problem of computing a confidence interval for the sum of adjacent bins of a deconvolved histogram as described in (Stanley, Patil, and Kuusela, 2022). This problem is a core statistical problem of particle unfolding in high-energy physics. For more detailed information, we refer the reader to (Kuusela and Victor M. Panaretos, 2015b; Kuusela, 2016; Kuusela and Philip B. Stark, 2017b; CMS Collaboration, 2016; CMS Collaboration, 2019).

The data-generating process is linear with Gaussian noise,

$$\boldsymbol{y} = \boldsymbol{K}\boldsymbol{x}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad \boldsymbol{x}^* \geq \boldsymbol{0}, \tag{3.42}$$

where $\boldsymbol{K} \in \mathbb{R}^{40\times 80}$ and $\varphi(\boldsymbol{x}) = \boldsymbol{h}^\top \boldsymbol{x}$. The vector $\boldsymbol{h}$ defines the bin-adjacent aggregation. In particle unfolding, the vectors $\boldsymbol{x}^*$ and $\boldsymbol{y}$ represent particle counts within discretized bins. A collection of $\boldsymbol{h}$ vectors can act to sum the contents of adjacent bins to effectively lower the resolution of the inference problem. The LLR is then defined as follows:

$$\lambda(\mu, \boldsymbol{y}) = \min_{\substack{\boldsymbol{h}^\top \boldsymbol{x} = \mu \\ \boldsymbol{x} \in \mathbb{R}^{80}_+}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2 - \min_{\boldsymbol{x} \in \mathbb{R}^{80}_+} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{x}\|_2^2. \tag{3.43}$$

We emphasize two features of this setup that complicate the task of computing confidence intervals for $\varphi(\boldsymbol{x})$. First, the forward model, $\boldsymbol{K}$, has a non-trivial null space and a large condition number, making any inverse problem point estimation and UQ markedly challenging. Typically, this sort of ill-posedness is handled with regularization of some kind, but as is well-known in the inverse problem literature and specifically shown in (Kuusela, 2016), including such regularization induces a bias, which can undercut desired statistical guarantees (e.g., coverage) of the inference object of interest. Including constraints and focusing on a particular functional of the parameter vector *implicitly* regularizes the problem (Patil, Kuusela, and Hobbs, 2022; Stanley, Patil, and Kuusela, 2022), but shifts the problem difficulty to inference with constraints. Although (Batlle, Stanley, et al., 2023) and this paper proposes a theoretical framework to perform inference with constraints, the second challenge is in the practical implementation due to the high-dimensional parameter space in scenarios like this example. As we show in the following sections, the Polytope sampler described by Algorithm 5 and quantile regression do an adequate job producing samples and fitting quantile surfaces in this high-dimensional space to ensure the desired coverage of the interval constructions.

As extensively discussed in (Stanley, Patil, and Kuusela, 2022), the OSB interval (i.e., using $\chi^2_{1,0.05}$ in the optimization-based interval construction) produces empirically valid confidence intervals in all tested scenarios, albeit typically with over-coverage. In the scenarios considered in (Stanley, Patil, and Kuusela, 2022), the underlying function generating the true histogram means ($\boldsymbol{x}^* \in \mathbb{R}^{80}_+$) was relatively smooth, likely contributing to the over-coverage. As such, we present two true parameter settings for $\boldsymbol{x}^*$ in (3.42) to highlight two advantages of our interval constructions over the OSB interval. First, we use the original smooth parameter setting from (Stanley,

Patil, and Kuusela, 2022) to show how our intervals improve over-coverage relative to the OSB interval by reducing the expected interval length. Second, we present an "adversarial" setting where our interval constructions achieve nominal coverage while the OSB interval does not. Figure 3.8 shows the smooth and adversarial settings for $\boldsymbol{x}^*$. We constructed the adversarial setting by first computing our interval constructions on the smooth setting and then looking at the maximum out-of-sample predicted quantile for a generated observation with a large predicted quantile. For each observation drawn within both settings, we draw 2.1×10^4 samples using the Polytope sampler as described by Algorithm 5.

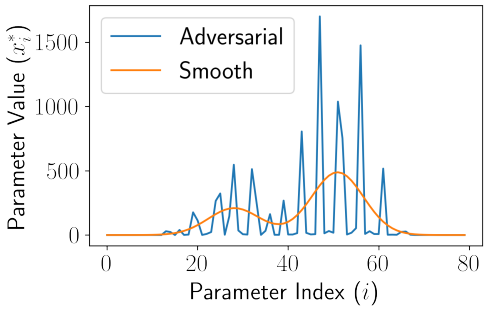


Figure 3.8: Parameter values for the smooth and adversarial settings for $\boldsymbol{x}^*$ used to illustrate our interval construction versus the OSB interval. The adversarial setting is made more difficult by the sharp jumps in parameter values.

Smooth setting

Using the smooth $\boldsymbol{x}^*$ shown in Figure 3.8, (Stanley, Patil, and Kuusela, 2022) showed that the OSB interval over-covers at the 95% level. Furthermore, it was shown that the OSB interval was the shortest across a range of other interval options, including SSB, prior optimized, and minimax. As such, for this setting, we show that our interval constructions not only achieve nominal coverage, but the sliced constructions dramatically reduce over-coverage compared to OSB by producing substantially shorter intervals on average. Estimated coverage and expected interval lengths are shown in Figure 3.9.

Both Global interval constructions and OSB dramatically over-cover which highlights the conservatism of the Global constructions. These estimated coverage values

indicate that within each observation's Berger–Boos set, there is a parameter setting against which the method has to protect that is substantially more difficult to cover than the true realistic parameter setting. Interestingly, both Global constructions produce markedly longer intervals on average compared with the OSB interval. Aligning with the intuition from the Global and Sliced construction definitions, the Sliced intervals are less conservative as seen by their lower over-coverage and significantly smaller average lengths. Importantly, both Sliced constructions are shorter on average compared to the OSB interval, with the Sliced Inverted showing an 18.7% reduction in average length and the Sliced Optimized showing a 11.1% reduction in average length.

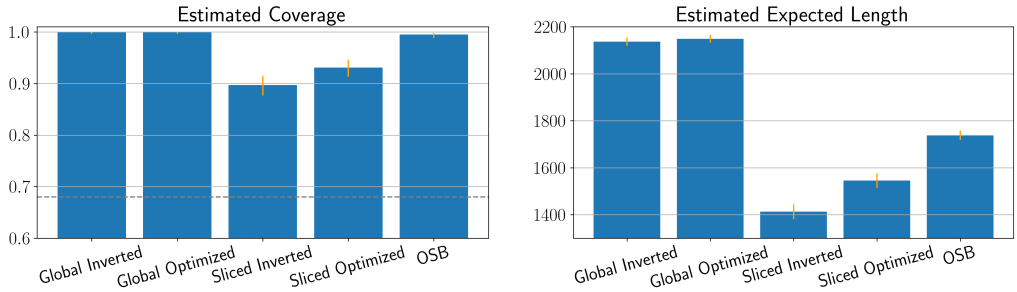


Figure 3.9: Estimated coverage and expected length across all four interval constructions and OSB at the 68% level for the **smooth** wide-bin deconvolution experiment. While the Global interval constructions over-cover like the OSB interval, the Sliced interval constructions reduce both over-coverage and expected interval length.

Adversarial setting

The key result in Section 3.6 is that the Sliced interval constructions both reduce over-coverage and expected interval length compared with the OSB interval. In this section, we show that for the adversarial parameter setting the OSB interval does not achieve nominal coverage, whereas all four of our interval constructions do achieve nominal coverage while still reducing the expected interval length in the case of the Sliced interval constructions compared to the OSB interval. The corresponding estimated coverage and expected length results are shown in Figure 3.10.

Both Global interval constructions and the Sliced Optimized interval over-cover, with the Sliced Optimized over-covering to a lesser extent than the Global intervals. The Sliced Inverted interval achieves nominal coverage within the statistical uncertainty. The estimated expected lengths tell a story similar to that of the smooth example, with the Global intervals showing the longest average interval lengths, the

Sliced intervals showing the shortest, and the OSB interval being between the two. Importantly, the Sliced intervals are again significantly shorter than the OSB interval, even though the OSB interval does not achieve nominal coverage. The Sliced Inverted interval shows a 18.9% average interval length reduction over OSB while the Sliced Optimized interval shows an 11.4% average interval length reduction.

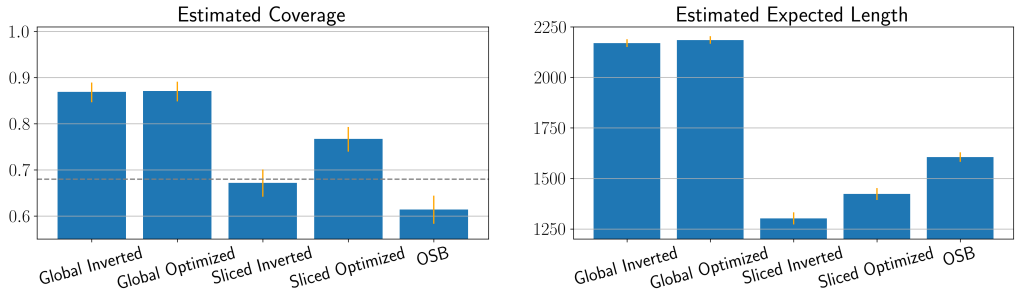


Figure 3.10: Estimated coverage and expected length across all four interval constructions and OSB at the 68% level for the **adversarial** wide-bin deconvolution experiment. While the OSB interval fails to achieve nominal coverage, all four of our interval constructions do. Interestingly, the Sliced interval constructions are meaningfully shorter than the OSB interval while also providing coverage.

For both the smooth and adversarial experiments, we note the differences in coverage and expected length between the two Sliced interval constructions. Although both approaches theoretically compute the same interval, significant differences can arise from their construction. For instance, the Sliced Inverted intervals are constructed by accepting individual functional values, which makes each point's acceptance dependent upon the quality of the quantile regressor at that point. By contrast, since the Sliced Optimized intervals essentially smooth over the max quantiles as a function of the functional space, the intervals are less sensitive to the quantile regressor's performance at any individual point. In the above simulation studies, there were realizations of the data for which the Sliced Inverted construction only accepted a single functional value sample, thereby making the confidence interval a single point. We found the Sliced Optimized construction to be more robust in these settings as the max quantile values were shared in a sliding window over the functional space.

3.7 Conclusion

This paper proposes several confidence interval constructions for functionals in constrained ill-posed inverse problems. Our approach is based on two key ideas: data-adaptive constraints using a Berger–Boos construction and sampling-based

inversion. Two independent decisions when constructing intervals provide four different valid intervals: Global versus Sliced, using the quantile function either over the entire or along level-set slices of the Berger–Boos set and Inverted versus Optimized, constructing the interval either by individually accepted functional values via the estimated quantile function or using the estimated quantile function in endpoint optimizations. All of the constructions are built upon the preliminary constraint by the data-informed Berger–Boos set, followed by a sampling procedure to estimate a quantile function that can be used to invert or optimize the interval endpoints. We have validated the method (including all four aforementioned interval constructions) through several numerical examples, demonstrating its ability to provide correct coverage, better calibration, and comparable or shorter interval length compared to the OSB interval baseline. Overall, our approach offers a flexible framework that can incorporate constraints directly and can be tailored to various types of inverse problems. The main takeaway is that data–adaptive constraining helps improve the length of the resulting confidence intervals, and enables sampling which makes it feasible to carry out the test inversion needed to construct confidence intervals with a desired nominal coverage.

There are several promising directions for future work. One direction is to find ways to extend our method to even higher-dimensional problems, which are more challenging. This would involve developing improved techniques to handle the curse of dimensionality and exploring the trade-off between accuracy and computational complexity. For the approaches in this paper in particular, this extension would require a more tailored sampling approach. Another direction is to leverage more sophisticated machine learning algorithms (deep learning models or ensemble methods) to improve the estimate of the quantile function and thus improve the accuracy and efficiency of our confidence intervals. Additionally, applying our approach to other applications involving ill-posed inverse problems, such as medical imaging or geophysics, would provide further validation of the effectiveness of our approach. Finally, it is of interest to conduct a further theoretical analysis of our approach under different constraints and noise conditions to better understand its limitations and strengths. This would involve studying the statistical properties of our confidence intervals and investigating the impact of various assumptions on their performance. Of particular interest are relaxations of the linear forward model and Gaussian noise assumptions to extend our method's application domain. Although the original theoretical foundation developed in (Batlle, Stanley, et al., 2023) does not make these assumptions, our implementation relies upon them for the tractability of the

sampling algorithms. Overall, these future research directions have the potential to demonstrate the applicability and robustness of our approach in a wide range of domains.

3.8 Proofs in Section 3.3

Proof of Lemma 3.3.1

We reproduce the original argument in (Roger L. Berger and Boos, 1994), originally stated in terms of p-values, translated into our quantile setting. Fix any $\boldsymbol{x}^* \in \mathcal{X}$ and consider the sets:

- $A_1 = \{\boldsymbol{y} : B_\eta(\boldsymbol{y}) \ni \boldsymbol{x}^*\}$
- $A_2 = \{\boldsymbol{y} : \lambda(\mu^*, \boldsymbol{y}) \leq Q_{\boldsymbol{x}}(1-\gamma)\}$
- $A_3 = \{\boldsymbol{y} : \lambda(\mu^*, \boldsymbol{y}) \leq \bar{q}_{\gamma,\eta}(\mu^*)\} = \{\boldsymbol{y} : \lambda(\mu^*, \boldsymbol{y}) \leq \sup_{\boldsymbol{x} \in \mathcal{B}_\eta \cap \Phi_{\mu^*}} Q_{\boldsymbol{x}}(1-\gamma)\}$

We now have that $\mathbb{P}(\boldsymbol{y} \in A_1) = 1-\eta, \mathbb{P}(\boldsymbol{y} \in A_2) = 1-\gamma$, and that $(A_1 \cap A_2) \subset (A_1 \cap A_3)$. Therefore,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{y} \notin A_3) &= \mathbb{P}(\boldsymbol{y} \notin A_3, \boldsymbol{y} \in A_1) + \mathbb{P}(\boldsymbol{y} \notin A_3, \boldsymbol{y} \notin A_1) && (3.44)\\
&\leq \mathbb{P}(\boldsymbol{y} \notin A_2, \boldsymbol{y} \in A_1) + \mathbb{P}(\boldsymbol{y} \notin A_1) && (3.45)\\
&\leq \mathbb{P}(\boldsymbol{y} \notin A_2) + \mathbb{P}(\boldsymbol{y} \notin A_1) && (3.46)\\
&= \gamma + \eta && (3.47)
\end{aligned}$$

so that $\mathbb{P}(\boldsymbol{y} \in A_3) \geq 1-\gamma-\eta$. Imposing $1-\gamma-\eta \geq 1-\alpha$ gives the desired result.

Proof of Corollary 3.3.2

By definition, $\bar{q}^{\mu}_{\gamma,\eta} \leq \bar{q}_{\gamma,\eta}$ for all $\mu \in \mathbb{R}$. Therefore, $C^{\mathrm{sl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta) \subseteq C^{\mathrm{gl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta)$, and thus

$$\mathbb{P}_{\boldsymbol{x}^*}\big(\mu^* \in C^{\mathrm{gl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta)\big) \geq \mathbb{P}_{\boldsymbol{x}^*}\big(\mu^* \in C^{\mathrm{sl}}_\alpha(\boldsymbol{y}; \mathcal{B}_\eta)\big) \geq 1-\alpha. \tag{3.48}$$

3.9 Proof of Theorem 3.4.1

Throughout the proof, we make use of the following lemma:

Lemma 3.9.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$ such that $\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ is achieved, and at least one of the minimizers $\boldsymbol{x}^*$ satisfies:*

1. *f is continuous at $\boldsymbol{x}^*$,*

2. $\boldsymbol{x}^*$ *is not an isolated point of* $\mathcal{X}$ *(i.e.,* $\forall \delta > 0, B_\delta(\boldsymbol{x}^*) \cap \mathcal{X} \neq \emptyset$*).*

Let μ *be a measure on* $\mathcal{X}$ *such that* $\mu(B) > 0$ *for all* $B \subseteq \mathcal{X}$ *such that* $\lambda_{Leb}(B) > 0$ *(where* $\lambda_{Leb}(B)$ *refers here to the Lebesgue measure of the set* B*). Let* $Y_m = \min_{i=1,\dots,m} f(\boldsymbol{x}_i)$*, where* $\boldsymbol{x}_i$ *are i.i.d. samples from* μ*. Then* $Y_m \xrightarrow{p} f(\boldsymbol{x}^*)$*.*

Proof. Fix $\varepsilon > 0$ and let us show that $\mathbb{P}(|Y_m - f(\boldsymbol{x}^*)| > \varepsilon) \to 0$. Since f is continuous at $\boldsymbol{x}^*$ there exists a $\delta > 0$ such that $f(B_\delta(\boldsymbol{x}^*)) \subset B_\varepsilon(f(\boldsymbol{x}^*))$ so that

$$\mathbb{P}(|Y_m - f(\boldsymbol{x}^*)| \geq \varepsilon) \leq \mathbb{P}(\boldsymbol{x}_i \notin B_\delta(\boldsymbol{x}^*),\ \forall i = 1, \dots, m) = (\mathbb{P}_{\boldsymbol{x}\sim\mu}(\boldsymbol{x} \notin B_\delta(\boldsymbol{x}^*)))^m. \tag{3.49}$$

Since $\boldsymbol{x}^*$ is not an isolated point, we have $\lambda_{\text{Leb}}(B_\delta(\boldsymbol{x}^*) \cap \mathcal{X}) > 0$ and therefore $\mathbb{P}_{\boldsymbol{x}\sim\mu}(\boldsymbol{x} \notin B_\delta(\boldsymbol{x}^*)) < 1$ and $\mathbb{P}(|Y_m - f(\boldsymbol{x}^*)| \geq \varepsilon) \to 0$. □

We begin by proving that the empirical maximums of the quantiles obtained both by Algorithm 1 and Algorithm 2 (assuming the quantile regressor is consistent) converge to the true max quantile.

Algorithm 1 Let $\bar{q}^{\text{de}}_{\gamma,\eta} := \max_{i=1,\dots,M} \widehat{q}^i_\gamma(N)$, where we explicitly write the dependence with the number of samples and the index i refers to the quantile estimated at the i-th sampled point $\boldsymbol{x}_i$. We aim to show that $\bar{q}^{\text{de}}_{\gamma,\eta} \xrightarrow{\text{p}} \bar{q}_{\gamma,\eta}$. We know that $\widehat{q}^i_\gamma(N)$ converges in probability to $Q_{P_{\boldsymbol{x}_i}}(1-\gamma)$ as $N \to \infty$. We have

$$\left|\bar{q}^{\text{de}}_{\gamma,\eta} - \bar{q}_{\gamma,\eta}\right| \leq \left|\bar{q}^{\text{de}}_{\gamma,\eta} - \max_{i=1,\dots,M} Q_{P_{\boldsymbol{x}_i}}(1-\gamma)\right| + \left|\max_{i=1,\dots,M} Q_{P_{\boldsymbol{x}_i}}(1-\gamma) - \bar{q}_{\gamma,\eta}\right|. \tag{3.50}$$

The first term can be made smaller than $\varepsilon/2$ as $N \to \infty$ by convergence of the estimator, and the second term can be made smaller than $\varepsilon/2$ as $M \to \infty$ by application of Lemma 3.9.1 to the quantile function (maximizing instead of minimizing).

Algortihm 2 The proof is identical to that of Algorithm 1, with the only difference of replacing the quantiles estimated via Monte Carlo sampling to those estimated by the quantile regression, and N to M_{tr}, the number of samples needed to train the quantile regression. Since the quantile regression is assumed to be consistent, the first term can be made arbitrarily small as M_{tr} grows, and the result follows.

Since identical convergence results apply for both algorithms, henceforth we will not explicitly distinguish: the proof is written in terms of N, which can be replaced by M_{tr}.

Proof of Statement 1 (Global Inverted)

Recall,

$$C_{\text{inv}}^{\text{gl}}(\boldsymbol{y}) = \left[\min_{k\in\{1,\dots,M\}:\lambda(\varphi(\boldsymbol{x}_k),\boldsymbol{y})\leq\widehat{q}(N)} \varphi(\boldsymbol{x}_k), \max_{k:\lambda(\varphi(\boldsymbol{x}_k),\boldsymbol{y})\leq\widehat{q}(N)} \varphi(\boldsymbol{x}_k) \right] \tag{3.51}$$

$$= \left[\min_{k\in\{1,\dots,M\}:\lambda(\mu_k,\boldsymbol{y})\leq\widehat{q}(N)} \mu_k, \max_{k:\lambda(\mu_k,\boldsymbol{y})\leq\widehat{q}(N)} \mu_k \right], \tag{3.52}$$

where $\widehat{q}(N) := \max_{i=1,\dots,M} \widehat{q}_\gamma^i(N)$, the estimated quantiles of the sampled $\boldsymbol{x}_i \in \mathcal{B}_\eta$ and $\mu_i := \varphi(\boldsymbol{x}_i)$. Note that μ_i are samples in $\varphi(\mathcal{B}_\eta) \subset \mathbb{R}$. Also note that we have more explicitly written out the interval definition (i.e., Equation (3.26)) to emphasize clarity rather than presentation.

Consider the left extreme of the interval, a similar argument follows from the right extreme. Consider three quantities:

$$\mu_1(N,M) := \min \mu_k \quad \text{s.t.} \quad k = 1,\dots,M \text{ and } \lambda(\mu_k,\boldsymbol{y}) \leq \widehat{q}(N) \tag{3.53}$$

$$\mu_2(M) := \min \mu_k \quad \text{s.t.} \quad k = 1,\dots,M \text{ and } \lambda(\mu_k,\boldsymbol{y}) \leq \bar{q}_{\gamma,\eta} \tag{3.54}$$

$$\mu_3 := \min \mu \quad \text{s.t.} \quad \mu \in \varphi\left(\mathcal{B}_\eta\right) \text{ and } \lambda(\mu,\boldsymbol{y}) \leq \bar{q}_{\gamma,\eta}, \tag{3.55}$$

Our goal is to show that as $N, M \to \infty$, $\mu_1 \xrightarrow{\text{p}} \mu_3$. Lemma 3.9.1 shows that $\mu_2 \xrightarrow{\text{p}} \mu_3$. Indeed, the minimization over the indices k such that the condition is satisfied can be seen as a rejection sampling strategy in which all accepted samples are samples of the feasible region of the optimization in μ_3. As M grows, since the sampler eventually samples all areas of $\varphi(\mathcal{B}_\eta)$, some samples are guaranteed to be close to the optimum with high probability. Finally, for fixed M and N going to infinity, $\mu_1 \xrightarrow{\text{p}} \mu_2$. This follows from the continuity of the optimization problem with respect to the right-hand side of the constraint, and the fact that $\max_{i=1,\dots,M} q(\boldsymbol{x}_i) \xrightarrow{\text{p}} \bar{q}_{\gamma,\eta}$. It follows that as $N, M \to \infty$, $\mu_1 \xrightarrow{\text{p}} \mu_3$.

Proof of Statement 2 (Sliced Inverted)

The proof technique is similar to the one of Statement 1, replacing $\widehat{q}(N) := \max_{i=1,\dots,M} \widehat{q}_\gamma^i(N)$ for $\widehat{q}_\gamma^k(N)$. Defined then, similarly as in the previous proof:

$$\mu_1(N,M) := \min \mu_k \quad \text{s.t.} \quad k = 1,\dots,M \text{ and } \lambda(\mu_k,\boldsymbol{y}) \leq \widehat{q}_\gamma^k(N) \tag{3.56}$$

$$\mu_2(M) := \min \mu_k \quad \text{s.t.} \quad k = 1,\dots,M \text{ and } \lambda(\mu_k,\boldsymbol{y}) \leq \bar{q}_{\gamma,\eta}(\mu_k) \tag{3.57}$$

$$\mu_3 := \min \mu \quad \text{s.t.} \quad \mu \in \varphi\left(\mathcal{B}_\eta\right) \text{ and } \lambda(\mu,\boldsymbol{y}) \leq \bar{q}_{\gamma,\eta}(\mu) \tag{3.58}$$

where $\bar{q}_{\gamma,\eta}(\mu) = \max_{\boldsymbol{x}\in\Phi_\mu\cap\mathcal{B}_\eta} Q_{\boldsymbol{x}}(1-\gamma)$. As $N \to \infty$, $\widehat{q}^k_\gamma(N) \to Q_{\boldsymbol{x}_k}(1-\gamma)$. Lemma 3.9.1 can be used to show $\mu_2 \xrightarrow{\text{p}} \mu_3$, because the sampler strategy used that first samples $\boldsymbol{x}_k$ and then accepts $\mu_k = \varphi(\boldsymbol{x}_k)$ as a sample if $\lambda(\varphi(\boldsymbol{x}_k), \boldsymbol{y}) < q^k_\gamma$, it can be shown that for every feasible point μ, there is eventually a sample close to it.

Finally, μ_1 will become arbitrarily close to μ_2 as M grows large, since both are taking the minimum over samples that are sampled densely from the feasible set, meaning accepted μ_k will eventually be close to μ_3 for both the case of μ_1 and the case of μ_2

Proof of Statement 3 (Global Optimized)

We prove the continuity of the optimization problem,

$$\max_{\mu\in\varphi(\mathcal{B}_\eta)} \mu \text{ s.t. } \lambda(\mu, \boldsymbol{y}) \leq q, \tag{3.59}$$

as a function of q in the positive measure interval $(\lambda(\bar{\mu}, \boldsymbol{y}), \bar{q}_{\gamma,\eta}]$. Therefore, convergence in probability follows as we have convergence in probability to $\bar{q}_{\gamma,\eta}$ as $N, M \to \infty$, which in particular implies that the maximum quantile estimate is in the interval $(\lambda(\bar{\mu}, \boldsymbol{y}), \bar{q}_{\gamma,\eta}]$ almost surely. We do so by appealing to the maximum theorem (Ok, 2007, S E.3), which, in general, guarantees continuity of functions of the form $f^*(\theta) = \sup\{f(x,\theta) : x \in C(\theta)\}$ as long as f is continuous, C is a continuous compact-valued correspondence and $C(\theta)$ is non-empty for all $\theta \in \Theta$. The continuity of C comes from the continuity of the LLR, f is equal to the identity and the strict feasibility condition ensures C is non-empty in $\Theta := (\lambda(\bar{\mu}, \boldsymbol{y}), \bar{q}_{\gamma,\eta}]$.

Proof of Statement 3 (Sliced Optimized)

We will prove sufficient conditions for convergence of $\inf_{\mu:\widehat{f_k}(\mu)\geq 0} \mu$ to $\inf_{\mu:f(\mu)\geq 0} \mu$ as $\widehat{f_k}$ converges to f, and the result will follow by taking $\widehat{f_k}(\mu) = \widehat{m}_\gamma(\mu) - \lambda(\mu, \boldsymbol{y})$ and $f(\mu) = m_\gamma(\mu) - \lambda(\mu, \boldsymbol{y})$. A similar argument can be repeated for the supremum. Use use the notation $\widehat{f_k}$ to indicate that k sampled points are used to estimate this function via the definition of $\widehat{m}_\gamma(\mu)$ (see Section 3.3).

Lemma 3.9.2. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function, and let f_k be a sequence of functions $f_k : \mathbb{R} \to \mathbb{R}$. Let $\mu^* = \inf_{f(\mu)\geq 0} \mu$ and $\mu_k = \inf_{f_k(\mu)\geq 0} \mu$. Let the sequence of functions $\{f_k\}$ be such that for all $\delta > 0$,*

$$\mathbb{P}\left(\sup_\mu |f_k(\mu) - f(\mu)| > \delta\right) \to 0 \text{ as } k \to \infty, \tag{3.60}$$

namely, f_k converges in probability uniformly to f. Furthermore, let f be such that for all $\varepsilon > 0$, there exists $\delta' > 0$ such that $|f(\mu)| > \delta'$ if and only if $|\mu - \mu^*| < \varepsilon$. Then, we have $\mu_k \xrightarrow{\text{p}} \mu^*$.

We then have $\mu_k \xrightarrow{p} \mu^*$.

Proof. Assume for the sake of contradiction that there exists $\varepsilon > 0$ such that $\mathbb{P}(|\mu_k - \mu^*| \geq \varepsilon)$ does not go to 0. Then, by the condition on f, it must follow that $\mathbb{P}(|f(\mu_k)| > \delta)$ does not go to 0. But,

$$\mathbb{P}(|f(\mu_k)| > \delta) \leq \mathbb{P}(|f(\mu_k) - f_k(\mu_k)| > \delta) + \mathbb{P}(|f_k(\mu_k)| > \delta), \tag{3.61}$$

and the right-hand side goes to 0 by uniform convergence in probability (first term) and by feasibility of μ_k (second term). □

3.10 Additional details and illustrations in Section 3.6

Importance-like sampler for the three-dimensional example in Section 3.6

Since the parameter settings with quantiles meaningfully larger than $\chi^2_{1,\alpha}$ are located close to the constraint boundary, a sampling challenge is presented. Using the samplers as described in Section 3.5 results in under-sampling of this large-quantile region since both samplers provide uniform random samples over the Berger–Boos set. Algorithm 6 presents a modified version of Algorithm 2, an importance-like sampler to increase the probability mass of samples close to the constraint boundary. Note, we say "importance-like" because we do not provide any theoretical guarantee regarding this sampler's ability to produce draws from a particular target distribution. We tailored Algorithm 6 to settings with a non–negativity constraint and hand-tuned the length scale parameter to the particular three-dimensional example in Section 3.6.

The key to Algorithm 6 is the additional accept/reject step where the k-th sample is accepted with probability p_k. Additionally, by setting $q \in (0, 1)$, we prioritize retaining samples closer to the non-negativity boundary. This effect can be seen in Figure 3.11, where the vast majority of samples are found closer to the constraint boundary. The ability of Algorithm 6 to better sample the high-quantile regions of the Berger–Boos set can be seen in Figure 3.12. In particular, for the functional values $\mu \in (-4, -2)$, the importance-like sampler is substantially more effective than the Polytope sampler at finding parameter setting with γ-quantiles greater than 1.1.

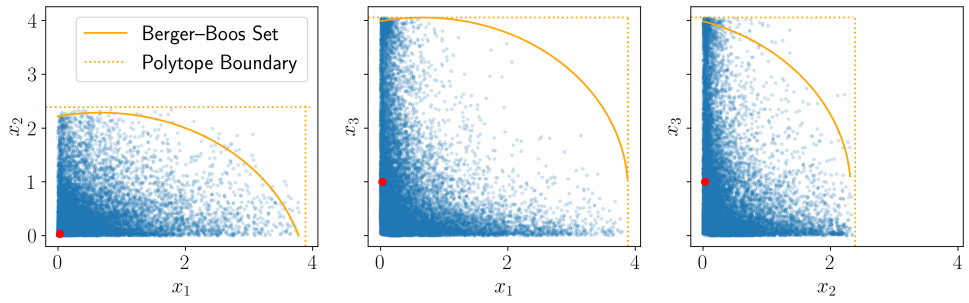


Figure 3.11: When sampling points using Algorithm 6, the vast majority of samples are found closer to the non-negativity constraint boundary. The true parameter setting is shown by the red point, while the parameter settings sampled by Algorithm 6 are shown by the blue points. This sampling prioritization helps adequately sample the regions of the Berger–Boos set where the quantile surface is larger than $\chi^2_{1,\alpha}$. Furthermore, the vast majority of sampled points lie within the Berger–Boos set with some lying outside within the bounding polytope.

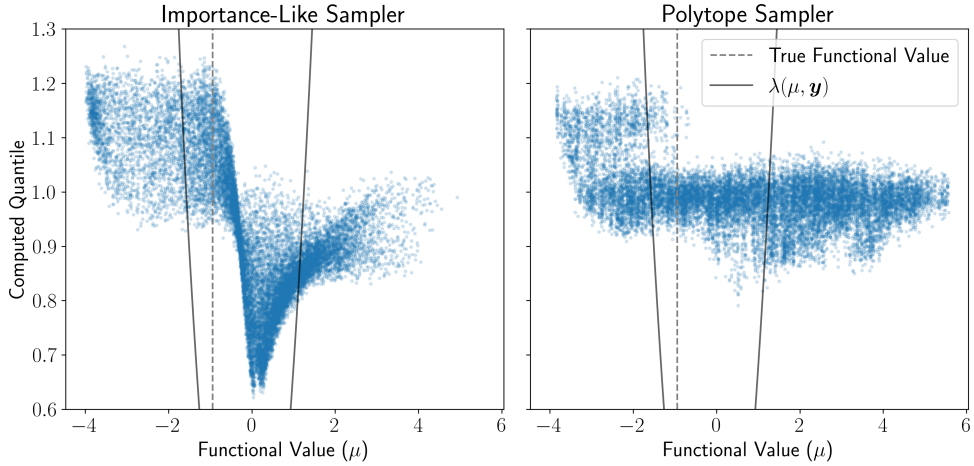


Figure 3.12: The importance-like sampler described by Algorithm 6 is more effective than the Polytope sampler described by Algorithm 5 at sampling parameter settings with γ-quantile greater than 1.1. Each parameter setting sampled by Algorithm 6 is shown by a blue point. This improved ability helps ensure the coverage guarantee shown in the left panel of Figure 3.6.

Algorithm 6 Importance-like sampler for three-dimensional constrained Gaussian

Input: Number of samples: $M \in \mathbb{N}$, inverse length scale: γ_p, and order of norm: $q \in (0, 1)$.

Instantiate a list $\mathcal{S}$ of length M to store sampled points.
while $|\mathcal{S}| < M$ **do**
  Draw $M - |\mathcal{S}|$ realizations from Algorithm 5: $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_{M-|\mathcal{S}|}$
  for $k = 1, \ldots, M - |\mathcal{S}|$: **do**
    Compute the probability of accepting the k-th draw: $p_k := \exp\left(-\gamma_p \|\widetilde{\boldsymbol{x}}_k\|_q\right)$.
    Draw $z_k \sim \text{Bernoulli}(p_k)$.
    if $z_k = 1$ **then**
      $\mathcal{S}[k] \leftarrow \widetilde{\boldsymbol{x}}_k$
    end if
  end for
end while

Output: Sampled parameters in Berger–Boos set $\mathcal{S}$.

Chapter 4

UNCERTAINTY QUANTIFICATION OF THE 4TH KIND; OPTIMAL POSTERIOR ACCURACY-UNCERTAINTY TRADEOFF WITH THE MINIMUM ENCLOSING BALL

Uncertainty quantification (UQ) is, broadly, the task of determining appropriate uncertainties to model predictions. There are essentially three kinds of approaches to Uncertainty Quantification: (A) robust optimization (min and max), (B) Bayesian (conditional average), and (C) decision theory (minmax). Although (A) is robust, it is unfavorable with respect to accuracy and data assimilation. (B) requires a prior, it is generally non-robust (brittle) with respect to the choice of that prior, and posterior estimations can be slow. Although (C) leads to the identification of an optimal prior, its approximation suffers from the curse of dimensionality and the notion of loss/risk used to identify the prior is one that is averaged with respect to the distribution of the data. We introduce a fourth kind which is a hybrid between (A), (B), (C), and hypothesis testing. It can be summarized as, after observing a sample x, (1) defining a likelihood region through the relative likelihood and (2) playing a minmax game in that region to define optimal estimators and their risk. The resulting method has several desirable properties: (a) an optimal prior is identified after measuring the data and the notion of loss/risk is a posterior one, (b) the determination of the optimal estimate and its risk can be reduced to computing the minimum enclosing ball of the image of the likelihood region under the quantity of interest map (such computations are fast and do not suffer from the curse of dimensionality). The method is characterized by a parameter in $[0, 1]$ acting as an assumed lower bound on the rarity of the observed data (the relative likelihood). When that parameter is near 1, the method produces a posterior distribution concentrated around a maximum likelihood estimate (MLE) with tight but low confidence UQ estimates. When that parameter is near 0, the method produces a maximal risk posterior distribution with high confidence UQ estimates. In addition to navigating the accuracy-uncertainty tradeoff, the proposed method addresses the brittleness of Bayesian inference by navigating the robustness-accuracy tradeoff associated with data assimilation.

4.1 Introduction

The past century has seen a steady increase in the need of estimating and predicting complex systems and making (possibly critical) decisions with limited information (H. Owhadi and C. Scovel, 2017c). These decisions are currently being formed based on increasingly complex models with imperfectly known parameters estimated based on available (limited) data whose distribution depends on the unknown/imperfectly known parameters of the model (if the model is well specified, i.e., if the distribution of the data belongs to the parametric family of distributions represented by the model). Making decisions and assessing the risk of these decisions requires identifying methods for data assimilation (estimating the parameters of the model based on data) and quantifying the risk/uncertainties of these decisions/parametric models. Such UQ methods are not unique, and they essentially differ through assumptions made on the generation of the true parameter of the model. In all inference/UQ methods, there is a tradeoff between robustness and accuracy (H. Owhadi and C. Scovel, 2017b), and these assumptions lead to the accuracy of the underlying method when they hold true but also to their lack of robustness when they do not hold true. In this paper, we introduce a new and rigorous UQ method that navigates (in a Pareto optimal manner) this tradeoff between accuracy and robustness in data assimilation and UQ for parametric models.

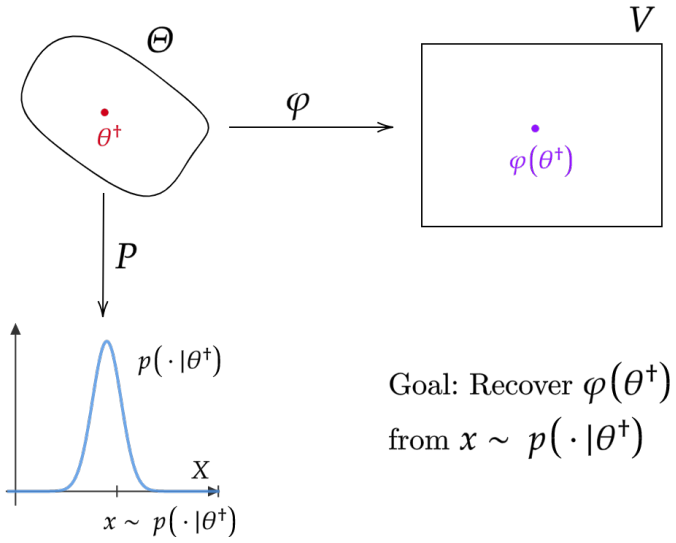


Figure 4.1: The Uncertainty Quantification (UQ) problem. Here, Θ is the space of parameters, $\theta^\dagger$ is the true unknown parameter, φ a quantity of interest, and P is the physical model determining the distribution p from which the data x is observed

The problem

To describe this method we formalize the underlying UQ problem as follows (see Fig. 4.1). Given a parameter space Θ and a quantity of interest $\varphi : \Theta \to V$ we seek to estimate $\varphi(\theta^\dagger)$, where $\theta^\dagger \in \Theta$ is an unknown parameter, based on the observation of some data $x \in X$ sampled from a probability distribution $p(\cdot|\theta^\dagger)$ (given by our model P) depending on the unknown parameter $\theta^\dagger$. Note that if our goal is to recover $\theta^\dagger$ itself, then we can let φ be the identity function. A simple example (detailed in Sec. 4.1) is to recover the probability $\theta^\dagger$ that a coin lands on heads, given the observation $x = (x_1, \ldots, x_n) \in \{H, T\}^n$ of n tosses of that coin. Note that this general setup combines parametric uncertainty ($\theta^\dagger$ is unknown) with aleatoric uncertainty (the x data is a sample from a random variable whose distribution depends on $\theta^\dagger$), and they need be merged to estimate $\varphi(\theta^\dagger)$ and quantify the uncertainty/risk of the estimation.

The three main approaches to UQ

There are currently three main approaches (detailed in Sec. 4.2) to addressing this UQ problem. The **worst case** (robust optimization) approach is (if φ is real-valued) to compute, the minimum and maximum possible value of $\varphi(\theta)$ over all possible values the parameter $\theta \in \Theta$. Although the data may be incorporated through empirical distribution inequalities (H. Owhadi and C. Scovel, 2017a), the worst-case approach is conservative and, due to its lack of assumptions on the generation of $\theta^\dagger$, it is at the robust end of the tradeoff between accuracy and robustness. Indeed this approach is simply based on the observation that

$$\varphi(\theta^\dagger) \in \left[\min_{\theta\in\Theta} \varphi(\theta), \max_{\theta\in\Theta} \varphi(\theta) \right]. \tag{4.1}$$

The **Bayesian approach** is to assume that $\theta^\dagger$ is a sample from a prior distribution π on Θ, then estimate $\varphi(\theta^\dagger)$ and quantify the uncertainty of that estimation by computing the posterior distribution of $\theta^\dagger$ given the data x. Writing $d(x)$ for the estimation of $\varphi(\theta^\dagger)$ ($d : X \to V$), the Bayesian decision theoretic variant of the Bayesian approach is to introduce a loss/cost

$$\mathcal{L}(\theta, d) = \mathbb{E}_{x\sim P(\cdot|\theta)} \mathbb{E}\left[\|\varphi(\theta) - d(x)\|^2 \right] \tag{4.2}$$

for the choice of the estimator d if the true value of unknown parameter is θ, assume that $\theta^\dagger$ is sampled from a known prior distribution π and identify an optimal estimator d_π as a minimizer

$$d_\pi = \operatorname{argmin}_d \mathbb{E}_{\theta\sim\pi}\left[\mathcal{L}(\theta, d) \right], \tag{4.3}$$

of the π-averaged loss $\mathbb{E}_{\theta\sim\pi}\big[\mathcal{L}(\theta,d)\big]$, whose value at the minimum defines the risk of that estimator. Due to the strength of the assumption that $\theta^\dagger$ is sampled from a known prior distribution, the Bayesian approach is at the accurate end of the tradeoff between accuracy and robustness: in particular, it is brittle to the choice of prior (H. Owhadi and C. Scovel, 2017b; H. Owhadi, C. Scovel, and T. Sullivan, 2015b; H. Owhadi, C. Scovel, and T. Sullivan, 2015a; H. Owhadi and C. Scovel, 2016). The **game/decision theoretic** approach formulates the underlying UQ problem as a zero-sum game in which θ is chosen by an adversarial player (Player I) seeking to maximize the loss $\mathcal{L}(\theta,d)$ and d is chosen by Player II seeking to minimize that loss. As in classical game theory (Neumann, 1928), identifying a Nash equilibrium requires lifting this game by letting Player I randomize the selection of θ according to some mixed strategy/prior distribution π on Θ and considering the average loss,

$$\mathcal{L}(\pi,d)=\mathbb{E}_{\theta\sim\pi,x\sim P(\cdot|\theta)}\mathbb{E}\big[\|\varphi(\theta)-d(x)\|^2\big]\,,\quad \pi\in\mathcal{P}(\Theta)\,,d\,:\,X\to V\,. \tag{4.4}$$

A saddle point (π^*,d_{π^*}) for (4.4) is then identified by letting d_π be the best Bayesian response (4.3) to π and π^* be a maximizer of the average-loss $\mathbb{E}_{\theta\sim\pi}\big[\mathcal{L}(\theta,d_\pi)\big]$, i.e.,

$$\pi^*\in\operatorname{argmax}_{\theta\in\Theta}\mathbb{E}_{\theta\sim\pi}\big[\mathcal{L}(\theta,d_\pi)\big]\,. \tag{4.5}$$

Although this approach achieves a balance in the accuracy/robustness tradeoff by relaxing the assumption that θ is sampled from a known distribution, it does not explicitly enable a navigation of that tradeoff. Furthermore, (1) the numerical approximation of an optimal mixed strategy for Player II suffers from the curse of dimensionality, and (2) d_π is the best response to a data-averaged notion of risk rather than a data-given notion of risk.

Our new approach to UQ

In this paper, we present a **new approach** that does not suffer from weaknesses present in previous UQ methods such as brittleness and curse of dimensionality and that explicitly navigates the tradeoff between accuracy and robustness in the estimation of the quantity of interest. Motivated by the fact that the main cause of brittleness in inference is the possible rarity of the observed data (H. Owhadi and C. Scovel, 2017b; H. Owhadi, C. Scovel, and T. Sullivan, 2015b; H. Owhadi, C. Scovel, and T. Sullivan, 2015a; H. Owhadi and C. Scovel, 2016), the first step of this approach is to make the hypothesis that the parameter θ that has generated the data is such that the data is not rare and bound the probability that this hypothesis is false. To describe this, given the observation x, for $\alpha\in[0,1]$ let $\Theta_x(\alpha)$ be the set

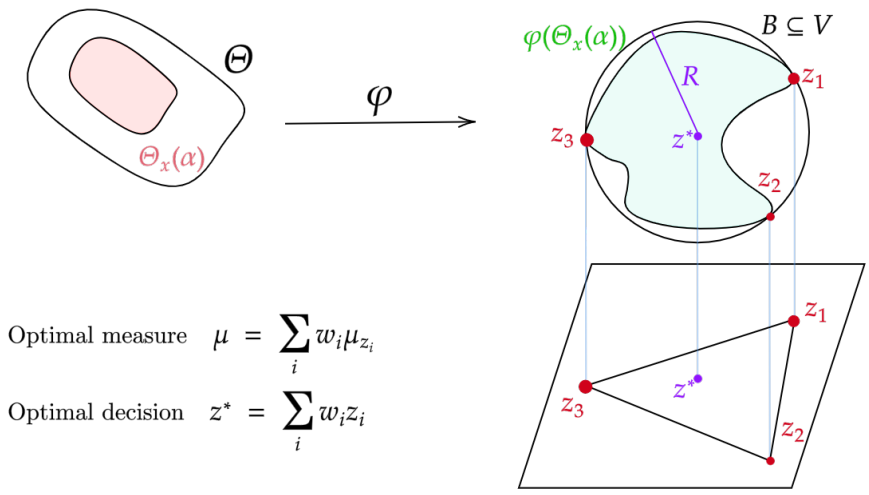


Figure 4.2: Example of the minimum enclosing ball B about the image $\varphi(\Theta_x(\alpha))$ (in green) with radius R and center $d = z^*$. An optimal discrete measure $\mu := \sum w_i \delta_{z_i}$ ($z_i = \varphi(\theta_i)$) on the range of φ for the maximum variance problem is characterized by the fact that it is supported on the intersection of $\varphi(\Theta_x(\alpha))$ and ∂B and $d = z^* = \sum w_i z_i$ is the center of mass of the measure μ. The size of the solid red balls indicates the size of the corresponding weights w_i.

of parameters $\theta \in \Theta$ whose relative likelihood

$$\bar{p}(x|\theta) := \frac{p(x|\theta)}{\sup_{\theta'} p(x|\theta')} \tag{4.6}$$

exceeds the threshold α, i.e.,

$$\Theta_x(\alpha) := \left\{\theta \in \Theta : \bar{p}(x|\theta) \geq \alpha\right\} \tag{4.7}$$

and let β_α be the maximum (over $\theta \in \Theta$) probability that θ does not belong to $\Theta_x(\alpha)$ when x is randomized according to the model $p(\cdot|\theta)$, i.e.,

$$\beta_\alpha := \sup_{\theta \in \Theta} P\left(\left\{x' \in X : \theta \notin \Theta_{x'}(\alpha)\right\} \middle| \theta\right). \tag{4.8}$$

β_α is interpreted as the significan/p-value of the hypothesis that $\theta^\dagger \in \Theta_x(\alpha)$. In particular, for α close to one $\Theta_x(\alpha)$ concentrates around the Maximum Likelihood Estimators of $\theta^\dagger$ and the probability β_α that the hypothesis is true goes to zero (which corresponds to accurate side of the tradeoff between accuracy and robustness). For α close to zero, $\Theta_x(\alpha)$ stretches over the whole set Θ, and the probability β_α that the hypothesis is true goes to one (which corresponds to the robust side of the tradeoff). The next step of this approach is to employ the game/decision theoretic approach with Θ replaced by the smaller set $\Theta_x(\alpha)$, i.e., replace (4.4) with

$$\mathcal{L}(\pi, d) = \mathbb{E}_{\theta\sim\pi, x\sim P(\cdot|\theta)} \mathbb{E}\left[\|\varphi(\theta) - d(x)\|^2\right], \quad \pi \in \mathcal{P}(\Theta_x(\alpha)), d : X \to V, \tag{4.9}$$

compute a saddle point (π^α, d^α) for (4.9) ($d^\alpha = d_{\pi^\alpha}$), and identify the optimal estimator as d^α and its risk/uncertainty $\mathcal{R}(d^\alpha)$ as the value of the game:

$$\mathcal{R}(d^\alpha) := \mathcal{L}(\pi^\alpha, d^\alpha)\,. \tag{4.10}$$

Main result

One of our main results (Theorems 4.3.3 and 4.3.5) is that this optimal decision and its associated risk/uncertainty (defined as the value of the game at the Nash equilibrium) can be identified as the center and the radius of the smallest ball enclosing the image of $\Theta_x(\alpha)$ under φ (see Fig. 4.2). Furthermore, we present rigorous and practical algorithms (Algorithms 7 and 8)[1] with approximation accuracy guarantees for computing that minimum enclosing ball based on the observation that optimal mixed strategies (priors) π for Player I can be restricted to be supported at a maximum of $\dim(V) + 1$ points located on the boundary of that ball.

Coin toss

At the cost of some forward referencing, we will now describe an application of our proposed problem to the estimation of the probability that a coin lands on heads based on the observation of n independent tosses of that coin.

***n* tosses of a single coin.**

In this example, we estimate the probability that a biased coin lands on heads from the observation of n independent tosses of that coin. Specifically, we consider flipping a coin Y which has an unknown probability $\theta^\dagger$ of coming heads ($Y = 1$) and probability $1 - \theta^\dagger$ coming up tails ($Y = 0$). Here $\Theta := [0, 1]$, $X = \{0, 1\}$, and the model $P : \Theta \to \mathcal{P}(\{0, 1\})$ is $P(Y = 1|\theta) = \theta$ and $P(Y = 0|\theta) = 1 - \theta$. We toss the coin n times, generating a sequence of i.i.d. Bernoulli variables $(Y_1, \ldots, Y_n)$ all with the same unknown parameter $\theta^\dagger \in [0, 1]$, and let $x := (x_1, \ldots, x_n) \in \{0, 1\}^n$ denote the outcome of the experiment. Let $h = \sum_{i=1}^n x_i$ denote the number of heads observed and $t = n - h$ the number of tails. Then the model for the n-fold toss is

$$P(x|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1 - \theta)^{1-x_i} = \theta^h(1 - \theta)^t \tag{4.11}$$

[1] Python implementation for these algorithms can be found in `https://github.com/JPLMLIA/UQ4K`

and, given an observation x, the MLE is $\theta = \frac{h}{n}$ so that the relative likelihood (4.6) is[2]

$$\bar{p}(x|\theta) = \frac{\theta^h(1-\theta)^t}{\left(\frac{h}{n}\right)^h\left(\frac{t}{n}\right)^t}. \tag{4.12}$$

We seek to estimate θ, so let $V = \mathbb{R}$ and let the quantity of interest $\varphi : \Theta \longrightarrow V$ be the identity function $\varphi(\theta) = \theta$. In this case, given $\alpha \in [0,1]$, the likelihood region

$$\Theta_x(\alpha) = \left\{\theta \in [0,1] : \frac{\theta^h(1-\theta)^t}{\left(\frac{h}{n}\right)^h\left(\frac{t}{n}\right)^t} \geq \alpha\right\} \tag{4.13}$$

constrains the support of priors to points with relative likelihood larger than α. Using Theorem 4.3.5 with $m = dim(V) + 1 = 2$, one can compute a saddle point (π^α, d^α) of the game (4.9) as

$$\pi^\alpha = w\delta_{\theta_1} + (1-w)\delta_{\theta_2} \text{ and } d^\alpha = w\theta_1 + (1-w)\theta_2, \tag{4.14}$$

where w, θ_1, θ_2 maximize the variance

$$\begin{cases} \text{Maximize} & w\theta_1^2 + (1-w)\theta_2^2 - (w\theta_1 + (1-w)\theta_2)^2 \\ \text{over} & 0 \leq w \leq 1, \quad \theta_1, \theta_2 \in [0,1] \\ \text{subject to} & \frac{\theta_i^h(1-\theta_i)^t}{\left(\frac{h}{n}\right)^h\left(\frac{t}{n}\right)^t} \geq \alpha, \qquad i = 1,2\,. \end{cases} \tag{4.15}$$

Equation (4.8) allows us to compute $\beta \in [0,1]$ as a function of $\alpha \in [0,1]$. The solution of the optimization problem can be found by finding the minimum enclosing ball of the set $\Theta_x(\alpha)$, which in this 1-D case is also subinterval of the interval $[0,1]$. For $n = 5$ tosses resulting in $h = 4$ heads and $t = 1$ tails, Figure 4.3 plots (1) β, the relative likelihood, its level sets and minimum enclosing balls as a function of α, and (2) The risk $\mathcal{R}(d^\alpha)$ =(4.10) and optimal decision d^α as a function of β. Three different points in the $\alpha - \beta$ curve are highlighted. Note that as α goes from 0 to 1, the relative likelihood region $\Theta_x(\alpha)$ gets smaller (it shrinks towards the MLE), the optimal estimator d^α goes from the center of the worst case interval to the MLE estimate, the risk (variance) of the estimator shrinks (which corresponds to an increase in accuracy), but the confidence $1 - \beta(\alpha)$ in that risk (the probability $\beta(\alpha)$ that $\theta^\dagger \in \Theta_x(\alpha)$) also shrinks towards zero (which corresponds to a loss of robustness).

[2]Although the fact that $\bar{p}(x|0) = \bar{p}(x|1) = 0$ violates our positivity assumptions (described in Sec. 4.2) on the model in our framework, in this case this technical restriction can be removed, so we can still use this example as an illustration.

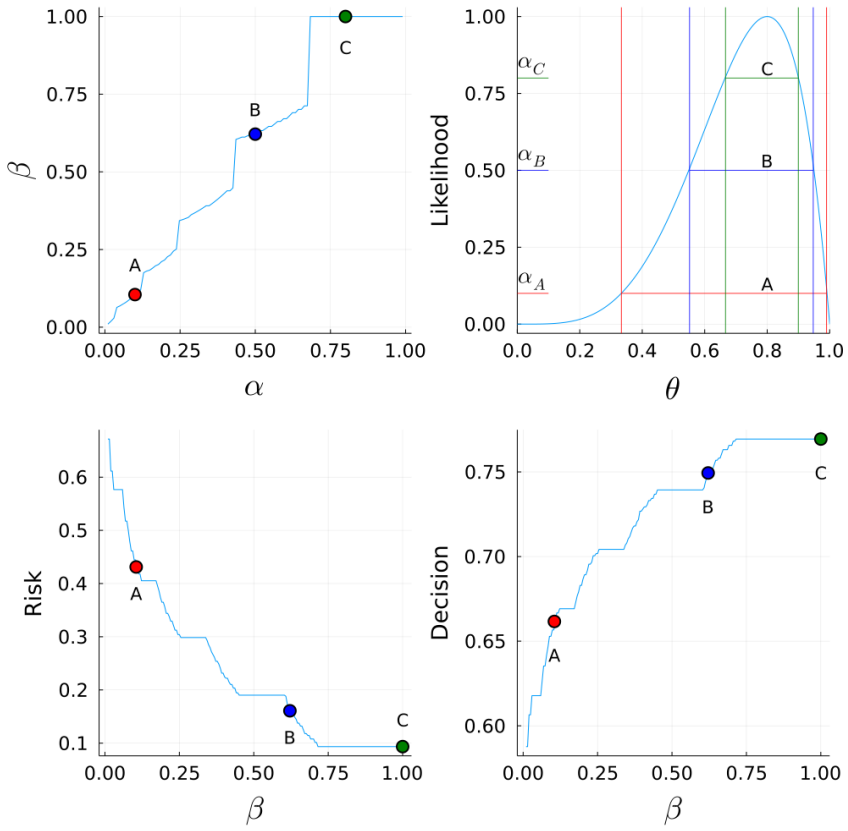


Figure 4.3: $\alpha - \beta$ relation, likelihood level sets, risk value and decision for different choices of α (and consequently β) for the 1 coin problem after observing four heads and one tail. Three different values in the $\alpha - \beta$ curve are highlighted across the plots

n_1 and n_2 tosses of two coins.

We now consider the same problem with two independent coins with unknown probabilities $\theta_1^\dagger, \theta_2^\dagger$. After tossing each coin i n_i times, the observation x consists of h_i heads and t_i tails for each i, produce a 2D relative likelihood function on $\Theta = [0,1]^2$ given by

$$\bar{p}(x|\theta_1,\theta_2) = \frac{\theta_1^{h_1}(1-\theta_1)^{t_1}}{\left(\frac{h_1}{n_1}\right)^{h_1}\left(\frac{t_1}{n_1}\right)^{t_1}} \frac{\theta_2^{h_2}(1-\theta_2)^{t_2}}{\left(\frac{h_2}{n_2}\right)^{h_2}\left(\frac{t_2}{n_2}\right)^{t_2}} . \tag{4.16}$$

Figure 4.4 illustrates the level sets $\bar{p}(x|\theta_1,\theta_2) \geq \alpha$ and their corresponding bounding balls for $h_1 = 1, t_1 = 3, h_2 = 5, t_2 = 1$ and different values of $\alpha \in [0,1]$.

Structure of the paper

This article is organized as follows: In Sec. 4.2, we formalize the UQ problem and review the three previous approaches to the problem, emphasizing the limitations

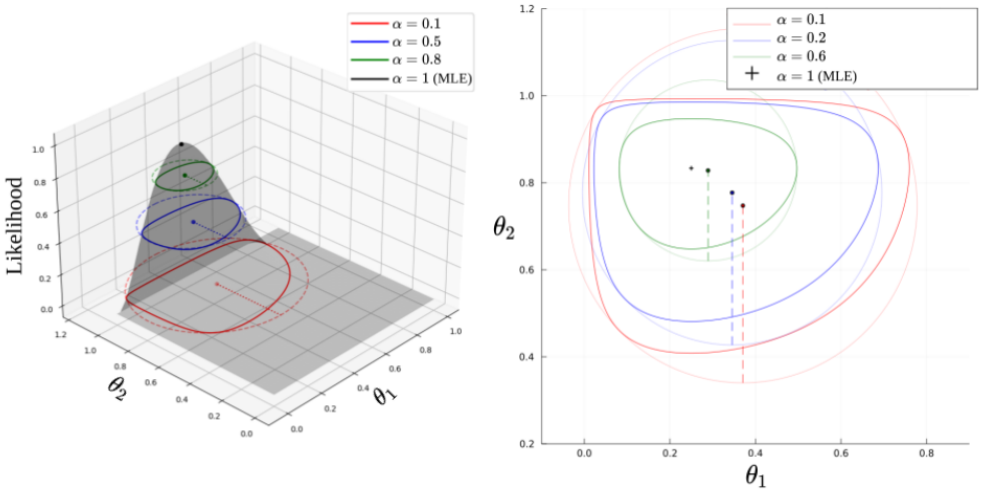


Figure 4.4: 2D likelihood level sets and minimum enclosing balls for different values of α, visualized as level sets of the likelihood function (left) and projected onto a 2D plane (right)

addressed with our method. In Sec. 4.3, we introduce a new kind of uncertainty quantification based on the minimum enclosing ball, and in Sections 4.4 and 4.5 we introduce the computational framework and our minimum enclosing ball algorithms. Sec. 4.6 presents numerical illustrations of the efficacy and scope of our approach. Sec. 4.7 generalizes the loss and rarity assumptions. Sec. 4.8 presents supporting theorems and proofs.

4.2 Previous approaches to UQ

We begin by formalizing the UQ problem introduced in the previous section. Let $\varphi : \Theta \to V$ be a quantity of interest, where V (the space of predictions) is a finite-dimensional vector space and Θ (the space of parameters) is a compact set. Let X (the space of data) be a measurable space and write $\mathcal{P}(X)$ for the set of probability distributions on X. Consider a model $P : \Theta \to \mathcal{P}(X)$ representing the dependence of the distribution of a data point $x \sim P(\cdot|\theta)$ on the value of the parameter $\theta \in \Theta$. Throughout, we use $\|\cdot\|$ to denote the Euclidean norm. We are then interested in solving the following problem.

Problem 1. Let $\theta^\dagger$ be an unknown element of Θ. Given an observation $x \sim P(\cdot|\theta^\dagger)$ of data, estimate $\varphi(\theta^\dagger)$ and quantify the uncertainty (accuracy/risk) of the estimate.

We assume that we can write a probability density function for any $P(\cdot|\theta)$. More

formally, we assume that P is a dominated model with positive densities, that is, for each $\theta \in \Theta$, $P(\cdot|\theta)$ is defined by a (strictly) positive density $p(\cdot|\theta) : X \to \mathbb{R}_{>0}$ with respect to a measure $\nu \in \mathcal{P}(X)$, such that, for each measurable subset A of X,

$$P(A|\theta) = \int_A p(x'|\theta) d\nu(x'), \quad \theta \in \Theta. \tag{4.17}$$

The three main approaches to UQ

Problem 1 is a fundamental Uncertainty Quantification (UQ) problem, and there are essentially three main approaches for solving it. We now describe them when V is a Euclidean space with the ℓ^2 loss function.

Worst-case

In a different setting, essentially where the set Θ consists of probability measures, the OUQ framework (H. Owhadi, C. Scovel, T. J. Sullivan, et al., 2013b) provides a worst-case analysis for providing rigorous uncertainty bounds. In the setting of this paper, in the absence of data (or ignoring the data x), the (vanilla) worst-case (or robust optimization) answer is to estimate $\varphi(\theta^\dagger)$ with the minimizer $d^* \in V$ of the worst-case error

$$\mathcal{R}(d) := \max_{\theta \in \Theta} \left[\|\varphi(\theta) - d\|^2 \right]. \tag{4.18}$$

In that approach, $(d^*, \mathcal{R}(d^*))$ are therefore identified as the center and squared radius of the minimum enclosing ball of $\varphi(\Theta)$.

Bayesian

The (vanilla) Bayesian (decision theory) approach (see e.g. Berger (J. O. Berger, 2013, Sec. 4.4)) is to assume that θ is sampled from a prior distribution $\pi \in \mathcal{P}(\Theta)$, and approximate $\varphi(\theta^\dagger)$ with the minimizer $d_\pi(x) \in V$ of the Bayesian posterior risk

$$\mathcal{R}_\pi(d) := \mathbb{E}_{\theta \sim \pi_x} \left[\|\varphi(\theta) - d\|^2 \right], \quad d \in V, \tag{4.19}$$

associated with the decision $d \in V$, where

$$\pi_x := \frac{p(x|\cdot)\pi}{\int_\Theta p(x|\theta) d\pi(\theta)} \tag{4.20}$$

is the posterior measure determined by the likelihood $p(x|\cdot)$, the prior π and the observation x. The minimizer $d_\pi(x)$ of (4.19) is the posterior distribution mean

$$d_\pi(x) := \mathbb{E}_{\theta \sim \pi_x} [\varphi(\theta)] \tag{4.21}$$

and the uncertainty is quantified by the posterior variance

$$\mathcal{R}_\pi(d_\pi(x)) := \mathbb{E}_{\theta\sim\pi_x}\left[\|\varphi(\theta) - d_\pi(x)\|^2\right]. \tag{4.22}$$

Game/decision theoretic

The Wald's game/decision theoretic approach is to consider a two-player zero-sum game where player I selects $\theta \in \Theta$, and player II selects a decision function $d : X \to V$ which estimates the quantity of interest $\varphi(\theta)$ (given the data $x \in X$), resulting in the loss

$$\mathcal{L}(\theta, d) := \mathbb{E}_{x\sim P(\cdot|\theta)}\left[\|\varphi(\theta) - d(x)\|^2\right], \qquad \theta \in \Theta,\ d : X \to V, \tag{4.23}$$

for player II. Such a game will normally not have a saddle point, so following von Neumann's approach (Neumann, 1928), one randomizes both players' plays to identify a Nash equilibrium. To that end, first observe that, for the quadratic loss considered here (for ease of presentation), because of the convexity of the loss in d, only the choice of player I needs to be randomized. Letting $\pi \in \mathcal{P}(\Theta)$ be a probability measure randomizing the play of player I, we consider the lift

$$\mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi}\mathbb{E}_{x\sim P(\cdot|\theta)}\left[\|\varphi(\theta) - d(x)\|^2\right], \qquad \pi \in \mathcal{P}(\Theta),\ d : X \to V, \tag{4.24}$$

of the game (4.23). A minmax optimal estimate of $\varphi(\theta^\dagger)$ is then obtained by identifying a Nash equilibrium (a saddle point) for (4.24), i.e. $\pi^* \in \mathcal{P}(\Theta)$ and $d^* : X \to V$ satisfying

$$\mathcal{L}(\pi, d^*) \le \mathcal{L}(\pi^*, d^*) \le \mathcal{L}(\pi^*, d), \qquad \pi \in \mathcal{P}(\Theta),\ d : X \to V. \tag{4.25}$$

Consequently, an optimal strategy of player II is then the posterior mean $d_{\pi^*}(x)$ of the form (4.21) determined by a worst-case measure and optimal randomized/mixed strategy for player I

$$\pi^* :\in \arg\max_{\pi\in\mathcal{P}(\Theta)} \mathbb{E}_{\theta\sim\pi, x\sim P(\cdot,\theta)}\left[\|\varphi(\theta) - d_\pi(x)\|^2\right]. \tag{4.26}$$

To connect with the Bayesian framework we observe (by changing the order of integration) that the Wald's risk (4.24) can be written as the average

$$\mathcal{L}(\pi, d) := \mathbb{E}_{x\sim X_\pi}\left[\mathcal{R}_\pi(d(x))\right] \tag{4.27}$$

of the Bayesian decision risk $\mathcal{R}_\pi(d(x))$ (=(4.19) for $d = d(x)$) determined by the prior π and decision $d(x)$ with respect to the X-marginal distribution

$$X_\pi := \int_\Theta P(\cdot|\theta) d\pi(\theta) \tag{4.28}$$

associated with the prior π and the model P. Therefore, the Wald framework identifies a worst-case prior (4.26), while the prior used in Bayesian decision theory is specified by the practitioner.

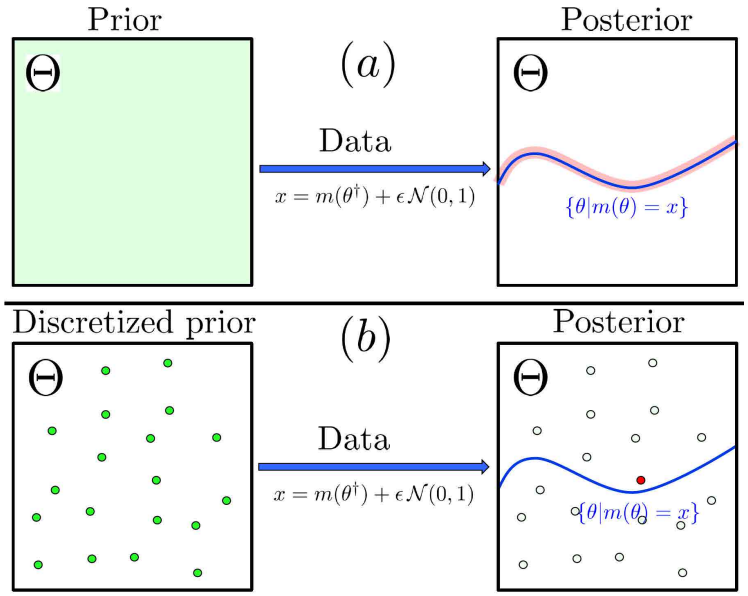


Figure 4.5: Curse of dimensionality in discretizing the prior. The data is of the form $x = m(\theta) + \epsilon\mathcal{N}(0,1)$ where m is deterministic and $\epsilon\mathcal{N}(0,1)$ is small noise. (a) For the continuous prior, the posterior concentrates around $\mathcal{M} := \{\theta \in \Theta | m(\theta) = x\}$. (b) For the discretized prior, the posterior concentrates on the delta Dirac that is the closest to $\mathcal{M}$.

Limitations of the three main approaches to UQ

All three approaches described in Section 4.2 have limitations in terms of accuracy, robustness, and computational complexity. Although the worst-case approach is robust, it appears unfavorable in terms of accuracy and data assimilation. The Bayesian approach, on the other hand, suffers from the computational complexity of estimating the posterior distribution and from brittleness (H. Owhadi, C. Scovel, and T. Sullivan, 2015b) with respect to the choice of prior along with Stark's admonition (P. Stark, 2020) "your prior can bite you on the posterior." Although Kempthorne (Kempthorne, 1987) develops a rigorous numerical procedure with convergence guarantees for solving the equations of Wald's statistical decision theory which appears amenable to computational complexity analysis, it suffers from the curse of dimensionality (see Fig. 4.5). This can be understood from the fact that the risk associated with the worst-case measure in the Wald framework is an average over the observational variable $x \in X$ of the conditional risk, conditioned on the observation x. Consequently, for a discrete approximation of a worst-case measure, after

an observation is made, there may be insufficient mass near the places where the conditioning will provide a good estimate of the appropriate conditional measure. Indeed, in the proposal (H. Owhadi and C. Scovel, 2017c) to develop Wald's statistical decision theory along the lines of Machine Learning, with its dual focus on performance and computation, it was observed that

> "Although Wald's theory of Optimal Statistical Decisions has resulted in many important statistical discoveries, looking through the three Lehmann symposia of Rojo and Pérez-Abreu (Rojo and Pérez-Abreu, 2004) in 2004, and Rojo (Rojo, 2006; Rojo, 2009) in 2006 and 2009, it is clear that the incorporation of the analysis of the computational algorithm, both in terms of its computational efficiency and its statistical optimality, has not begun."

Moreover, one might ask why, after seeing the data, one is choosing a worst-case measure which optimizes the average (4.27) of the Bayesian risk (4.22), instead of choosing it to optimize the value of the risk $\mathcal{R}_\pi(d_\pi(x))$ at the value of the observation x. It is therefore desirable for an approach to UQ to successfully assimilate the observed data, to avoid requiring having to manually select a prior and to have a data-dependent notion of risk. In Section 4.3, we will propose a framework with all these properties. A comparison of the properties of all the mentioned methods can be found in Table 4.1.

	Makes use of the observed data	No need to manually specify prior	Risk depends on the observed data
Worst case	×	✓	×
Bayesian	✓	×	✓
Decision Theory	✓	✓	×
UQ4K (Section 4.3)	✓	✓	✓

Table 4.1: Comparison of the three previous approaches to uncertainty quantification with our proposed method in Section 4.3

4.3 Uncertainty Quantification of the 4th Kind

Basic definitions

In this paper, we introduce a framework which is a hybrid between Wald's statistical decision theory (A. Wald and Wolfowitz, 1951), Bayesian decision theory (J. O. Berger, 2013, Sec. 4.4), robust optimization, and hypothesis testing. Here we de-

scribe its components for simplicity when the loss function is the ℓ^2 loss. Later in Section 4.7 we develop the framework for general loss functions.

Rarity assumption on the data

In (H. Owhadi, C. Scovel, and T. Sullivan, 2015b, Pg. 576) it was demonstrated that one could alleviate the brittleness of Bayesian inference (see (H. Owhadi, C. Scovel, and T. Sullivan, 2015a; H. Owhadi and C. Scovel, 2016)) by restricting to priors π for which the observed data x is not rare, that is,

$$p(x) := \int_\Theta p(x|\theta) d\pi(\theta) \geq \alpha \tag{4.29}$$

according to the density of the X-marginal determined by π and the model P, for some $\alpha > 0$. In the proposed framework, we consider playing a game *after observing the data* x whose loss function is defined by the Bayesian decision risk $\mathcal{R}_\pi(d)$ (4.19), where player I selects a prior π subject to a *rarity assumption* ($\pi \in \mathcal{P}_x(\alpha)$) and player II selects a decision $d \in V$. The rarity assumption considered here is

$$\mathcal{P}_x(\alpha) := \left\{ \pi \in \mathcal{P}(\Theta) : \text{support}(\pi) \subset \left\{ \theta \in \Theta : p(x|\theta) \geq \alpha \right\} \right\}. \tag{4.30}$$

Since $p(x|\theta) \geq \alpha$ for all θ in the support of any $\pi \in \mathcal{P}_x(\alpha)$ it follows that such a π satisfies (4.29) and therefore is sufficient to prevent Bayesian brittleness.

The relative likelihood for the rarity assumption

Observe in (4.20) that the map from the prior π to posterior π_x is scale-invariant in the likelihood $p(x|\cdot)$ and that the effects of scaling the likelihood in the rarity assumption can be undone by modifying α. Consequently, we scale the likelihood function

$$\bar{p}(x|\theta) := \frac{p(x|\theta)}{\sup_{\theta \in \Theta} p(x|\theta)}, \quad \theta \in \Theta, \tag{4.31}$$

to its *relative likelihood* function

$$\bar{p}(x|\cdot) : \Theta \to (0, 1]. \tag{4.32}$$

According to Sprott (Sprott, 2008, Sec. 2.4), the relative likelihood measures the plausibility of any parameter value θ relative to a maximum likely θ and summarizes the information about θ contained in the sample x. See Rossi (Rossi, 2018, p. 267) for its large sample connection with the $\mathcal{X}^2_1$ distribution and several examples of the relationship between likelihood regions and confidence intervals.

For $x \in X$ and $\alpha \in [0,1]$, let (4.7) denote the corresponding *likelihood region* and, updating (4.30), redefine the rarity assumption by

$$\mathcal{P}_x(\alpha) := \mathcal{P}\big(\Theta_x(\alpha)\big)\,. \tag{4.33}$$

That is, the rarity constraint $\mathcal{P}_x(\alpha)$ constrains priors to have support on the likelihood region $\Theta_x(\alpha)$. We will now define the confidence level of the family $\Theta_x(\alpha), x \in X$.

Significance/confidence level

For a given α, let the *significance* β_α at the value α be the maximum (over $\theta \in \Theta$) of the probability that a data $x' \sim P(\cdot|\theta)$ does not satisfy the rarity assumption $\bar{p}(x'|\theta) \geq \alpha$, i.e.,

$$\beta_\alpha \;:=\; \sup_{\theta\in\Theta} \int \mathbb{1}_{\{\bar{p}(\cdot|\theta)<\alpha\}}(x')p(x'|\theta)dv(x')\,, \tag{4.34}$$

where, for fixed θ, $\mathbb{1}_{\{\bar{p}(\cdot|\theta)<\alpha\}}$ is the indicator function of the set $\{x' \in X : \bar{p}(x'|\theta) < \alpha\}$. Observe that, in the setting of hypothesis testing, (1) β_α can be interpreted as the p-value associated with the hypothesis that the rarity assumption is not satisfied (i.e. the hypothesis that θ does not belongs to the set (4.7)), and (2) $1-\beta_\alpha$ can be interpreted as the confidence level associated with the rarity assumption (i.e. the smallest probability that θ belongs to the set (4.7)). Therefore, to select $\alpha \in [0,1]$, we set a *significance level* β^* (e.g. $\beta^* = 0.05$) and choose α to be the largest value such that the significance at α satisfies $\beta_\alpha \leq \beta^*$.

Connection to the Likelihood Ratio Test

The connection of the likelihood region we are defining and confidence sets can be more explicitly seen via the likelihood ratio test and the inversion of that test to produce a confidence set. Namely, we define a hypothesis test,

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0, \quad \theta \in \Theta. \tag{4.35}$$

The likelihood ratio test statistic is defined by

$$\Lambda(\theta_0, x) = \frac{p(x|\theta_0)}{\sup_{\theta\in\Theta} p(x|\theta)}. \tag{4.36}$$

Note, $\Lambda(\theta_0, x) = \bar{p}(x|\theta)$, i.e., Equation (4.31) and Equation (4.36) are equivalent. In the hypothesis testing setting, one would next define a critical value, $c > 0$, such that when $\Lambda(\theta_0, x) \leq c$, the null in test (4.35) is rejected. Ideally, c is chosen such

that the probability of false rejection under the null hypothesis is capped at some probability, $\alpha \in (0, 1)$, defining an α-level test. The type 1 error probability control is mathematically characterized as

$$P\left(\Lambda(\theta_0, x) \leq c|\theta_0\right) \leq \alpha. \tag{4.37}$$

The error control criterion of Equation (4.37) can be used to define an acceptance region in the sample space defined as follows:

$$A(\theta_0, c) = \{x \in X : \Lambda(\theta_0, x) \geq c\}. \tag{4.38}$$

The type 1 error control then implies $P(x \in A(\theta_0, c)|\theta_0) \geq 1 - \alpha$ and acts in a similar way to the control exerted by β_α in Equation (4.34). In that equation, α is chosen as the supremum over $\theta \in \Theta$ of $P\left(\Lambda(\theta, x) \leq c|\theta\right)$, thus making α a constant independent of θ. A viable alternative is to consider a curve $\alpha(\theta)$, where

$$P\left(\Lambda(\theta_0, x) \leq c|\theta_0\right) \leq \alpha(\theta_0) \quad \forall\ \theta_0, \tag{4.39}$$

and then redefining (4.7) accordingly. In this work we consider the fixed α model for simplicity.

In some cases, the false rejection control can be done cleanly. For instance, in the context of a noise model where $x = \theta + \varepsilon,\ \varepsilon \sim N(0, I),\ \theta \in \mathbb{R}^n$, it can be shown that the log-likelihood ratio follows the distribution,

$$-2 \log \Lambda(\theta_0, x) \sim \chi^2_n, \tag{4.40}$$

i.e., the log-likelihood ratio is distributed as a chi-squared distribution with n degrees of freedom, allowing c to be exactly chosen. In the event the test statistic distribution cannot be exactly known, asymptotic results such as that shown in Theorem 4.4.1 can provide similar results.

As discussed in Chapter 9 of Casella/Berger (G. Casella and R. L. Berger, 2002), one can think about inverting an α-level hypothesis test such as Test (4.35) to obtain a $1 - \alpha$ confidence set $C(x) \subset \Theta$, such that $P\left(\theta^* \in C(x)|\theta^*\right) \geq 1 - \alpha$, where θ^* is the true parameter value. The inverting is performed with the acceptance region of Equation (4.38) and the set is defined as follows:

$$C(x) = \{\theta_0 \in \Theta : x \in A(\theta_0, c)\}, \tag{4.41}$$

Note the equivalence between Equation (4.41) directly above and Equation (4.7) from the previous section. By the type 1 error control, we have

$$P(\theta^* \in C(x)|\theta^*) = P(x \in A(\theta^*, c)|\theta^*) \geq 1 - \alpha, \tag{4.42}$$

implying,

$$P(\theta^* \notin C(x)|\theta^*) \leq \alpha, \tag{4.43}$$

providing an additional view of the equivalence with Equation (4.34). As such, the *relative likelihood* and *rarity condition* can be seen through the more traditional statistical lens of the *likelihood ratio* and *type 1 error control* in the classical hypothesis testing setting.

Remark 4.3.1. For models where the maximum of the likelihood function

$$M(x') := \sup_{\theta \in \Theta} p(x'|\theta), \quad x' \in X,$$

is expensive to compute but for which there exists an efficiently computable upper approximation $M'(x') \geq M(x')$, $x' \in X$ available, the surrogate

$$\bar{p}'(x'|\theta) := \frac{p(x'|\theta)}{M'(x')}, \quad x' \in X, \tag{4.44}$$

to the relative likelihood may be used in place of (4.31). If we let β'_α denote the value determined in (4.34) using the surrogate (4.44) and $\Theta'_x(\alpha)$ denote the corresponding likelihood region, then we have $\beta_\alpha \leq \beta'_\alpha$ and $\Theta'_x(\alpha) \subset \Theta_x(\alpha)$, $\alpha \in [0, 1]$. Consequently, obtaining $\beta'_\alpha \leq \beta^*$ for significance level β^* implies that $\beta_\alpha \leq \beta^*$.

As an example, for an N-dimensional Gaussian model with $p(x'|\theta) = \frac{1}{(\sigma\sqrt{2\pi})^N} e^{-\frac{1}{2\sigma^2}\|x'-\theta\|^2}$ with $\Theta := [-\tau, \tau]^N$, the elementary upper bound

$$M(x') := \sup_{\theta \in \Theta} p(x|\theta) \leq \frac{1}{(\sigma\sqrt{2\pi})^N}$$

the surrogate relative likelihood defined in (4.44) becomes

$$\bar{p}'(x'|\theta) := e^{-\frac{1}{2\sigma^2}\|x'-\theta\|^2}.$$

Posterior game and risk

After observing $x \in X$, we now consider playing a game using the loss

$$\mathcal{L}(\pi, d) := \mathbb{E}_{\theta \sim \pi_x}\left[\|\varphi(\theta) - d\|^2\right], \quad \pi \in \mathcal{P}_x(\alpha), d \in V, \tag{4.45}$$

where π_x is the posterior (4.20). In (4.45), we think about the maximizing player as choosing the prior π and then the loss function depends on the posterior π_x. Since the likelihood $p(x|\cdot)$ is positive and the data x is fixed, we have support(π_x) =

support(π) and the map (4.20) mapping the prior π to the posterior π_x is bijective. Therefore one can equivalently consider the choice of the maximizing player to be directly maximizing the posterior π_x instead of the prior π that is later mapped to the posterior using the data. The optimal choices of these two games can then be mapped by (4.20) and its inverse. Using the invariance of posterior (4.20) under the scaling of the likelihood function $p(x|\cdot)$ we write the posterior in terms of the relative likelihood (4.31) as

$$\pi_x := \frac{\bar{p}(x|\cdot)\pi}{\int_\Theta \bar{p}(x|\theta)d\pi(\theta)}\,. \tag{4.46}$$

Therefore for simplicity one directly considers a game using the loss

$$\mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi}\left[\|\varphi(\theta) - d\|^2\right], \quad \pi \in \mathcal{P}_x(\alpha), d \in V. \tag{4.47}$$

Recall that a pair $(\pi^\alpha, d^\alpha) \in \mathcal{P}_x(\alpha) \times V$ is a saddle point of the game (4.47) if

$$\mathcal{L}(\pi, d^\alpha) \le \mathcal{L}(\pi^\alpha, d^\alpha) \le \mathcal{L}(\pi^\alpha, d), \qquad \pi \in \mathcal{P}_x(\alpha),\ d \in V.$$

We then have the following theorem.

Theorem 4.3.2. Consider $x \in X$, $\alpha \in [0, 1]$, and suppose that the relative likelihood $p(x|\cdot)$ and the quantity of interest $\varphi : \Theta \to V$ are continuous. The loss function $\mathcal{L}$ for the game (4.9) (=(4.86)) has saddle points and a pair $(\pi^\alpha, d^\alpha) \in \mathcal{P}_x(\alpha) \times V$ is a saddle point for $\mathcal{L}$ if and only if

$$d^\alpha := \mathbb{E}_{\pi^\alpha}[\varphi] \tag{4.48}$$

and

$$\pi^\alpha \in \arg\max_{\pi\in\mathcal{P}_x(\alpha)} \mathbb{E}_\pi\left[\|\varphi - \mathbb{E}_\pi[\varphi]\|^2\right]. \tag{4.49}$$

Furthermore the associated risk (the value of the two person game (4.9) (=(4.86)))

$$\mathcal{R}(d^\alpha) := \mathcal{L}(\pi^\alpha, d^\alpha) = \mathbb{E}_{\pi^\alpha}\left[\|\varphi - \mathbb{E}_{\pi^\alpha}[\varphi]\|^2\right] \tag{4.50}$$

is the same for all saddle points of $\mathcal{L}$. Moreover, the second component d^α of the set of saddle points is unique and the set $\mathcal{O}_x(\alpha) \subset \mathcal{P}_x(\alpha)$ of first components of saddle points is convex, providing a convex ridge $\mathcal{O}_x(\alpha) \times \{d^\alpha\}$ of saddle points.

Duality with the minimum enclosing ball

Although the Lagrangian duality between the maximum variance problem and the minimum enclosing ball problem on finite sets is known, see Yildirim (Yildirim,

2008), we now analyze the infinite case. Utilizing the recent generalization of the one-dimensional result of Popoviciu (Popoviciu, 1935) regarding the relationship between variance maximization and the minimum enclosing ball by Lim and McCann (Lim and McCann, 2021, Thm. 1), the following theorem demonstrates that essentially the maximum variance problem (4.49) determining a worst-case measure is the Lagrangian dual of the minimum enclosing ball problem on the image $\varphi(\Theta_x(\alpha))$. Let $\varphi_* : \mathcal{P}\big(\Theta_x(\alpha)\big) \to \mathcal{P}(\varphi(\Theta_x(\alpha))$ denote the pushforward map (change of variables) defined by $(\varphi_*\pi)(A) := \pi(\varphi^{-1}(A))$ for every Borel set A, mapping probability measures on $\Theta_x(\alpha)$ to probability measures on $\varphi\big(\Theta_x(\alpha)\big)$.

Theorem 4.3.3. For $x \in X$, $\alpha \in [0, 1]$, suppose the relative likelihood $\bar{p}(x|\cdot)$ and the quantity of interest $\varphi : \Theta \to V$ are continuous. Consider a saddle point (π^α, d^α) of the game (4.9) (=(4.86)). The optimal decision d^α and its associated risk $\mathcal{R}(d^\alpha)$ =(4.10) are equal to the center and squared radius, respectively, of the minimum enclosing ball of $\varphi(\Theta_x(\alpha))$, i.e. the minimizer z^* and the value R^2 of the minimum enclosing ball optimization problem

$$\begin{cases} \text{Minimize } r^2 \\ \text{Subject to } r \in \mathbb{R},\ z \in \varphi(\Theta_x(\alpha)), \\ \|x - z\|^2 \le r^2, \quad x \in \varphi(\Theta_x(\alpha)). \end{cases} \tag{4.51}$$

Moreover, the variance maximization problem on $\mathcal{P}_x(\alpha)$ (4.49) pushes forward to the variance maximization problem on the image of the likelihood region $\mathcal{P}(\varphi(\Theta_x(\alpha)))$ under φ, giving the identity

$$\mathbb{E}_\pi\left[\|\varphi - \mathbb{E}_\pi[\varphi]\|^2\right] = \mathbb{E}_{\varphi_*\pi}\left[\|v - \mathbb{E}_{\pi'}[v]\|^2\right], \quad \pi \in \mathcal{P}_x(\alpha),$$

and the latter is the Lagrangian dual to the minimum enclosing ball problem (4.51) on the image $\varphi(\Theta_x(\alpha))$. Finally, let B, with center z^*, denote the minimum enclosing ball of $\varphi(\Theta_x(\alpha))$. Then a measure $\pi^\alpha \in \mathcal{P}_x(\alpha)$ is optimal for the variance maximization problem (4.49) if and only if

$$\varphi_*\pi^\alpha\big(\varphi(\Theta_x(\alpha)) \cap \partial B\big) = 1$$

and

$$z^* = \int_V v d(\varphi_*\pi^\alpha)(v),$$

that is, all the mass of $\varphi_*\pi^\alpha$ lives on the intersection $\varphi(\Theta_x(\alpha)) \cap \partial B$ of the image $\varphi(\Theta_x(\alpha))$ of the likelihood region and the boundary ∂B of its minimum enclosing ball and the center of mass of the measure $\varphi_*\pi^\alpha$ is the center z^* of B.

Remark 4.3.4. Note that once α, and therefore $\Theta_x(\alpha)$, is determined that the computation of the risk and the minmax estimator is determined by the minimum enclosing ball about $\varphi(\Theta_x(\alpha))$, which is also determined by the worst-case optimization problem (4.18) for $\Theta := \Theta_x(\alpha)$.

Theorem 4.3.3 introduces the possibility of primal-dual algorithms, in particular, the availability of rigorous stopping criteria for the maximum variance problem (4.49). To that end, for a feasible measure $\pi \in \mathcal{P}_x(\alpha)$, let $\mathrm{Var}(\pi) := \mathbb{E}_\pi\left[\|\varphi - \mathbb{E}_\pi[\varphi]\|^2\right]$ denote its variance and denote by $\mathrm{Var}^* := \sup_{\pi\in\mathcal{P}_x(\alpha)} \mathrm{Var}(\pi)$ =(4.50) the optimal variance. Let (r, z) be feasible for the minimum enclosing ball problem (4.51). Then the inequality $\mathrm{Var}^* = R^2 \leq r^2$ implies the rigorous bound

$$\mathrm{Var}^* - \mathrm{Var}(\pi) \leq r^2 - \mathrm{Var}(\pi) \tag{4.52}$$

quantifying the suboptimality of the measure π in terms of known quantities r and $\mathrm{Var}(\pi)$.

Finite-dimensional reduction

Let $\Delta^m(\Theta)$ denote the set of convex sums of m Dirac measures located in Θ and, let $\mathcal{P}_x^m(\alpha) \subset \mathcal{P}_x(\alpha)$ defined by

$$\mathcal{P}_x^m(\alpha) := \Delta^m(\Theta) \cap \mathcal{P}_x(\alpha) \tag{4.53}$$

denote the finite-dimensional subset of the rarity assumption set $\mathcal{P}_x(\alpha)$ consisting of the convex combinations of m Dirac measures supported in $\Theta_x(\alpha)$.

Theorem 4.3.5. Let $\alpha \in [0, 1]$ and $x \in X$, and suppose that the likelihood function $p(x|\cdot)$ and quantity of interest $\varphi : \Theta \longrightarrow V$ are continuous. Then for any $m \geq dim(V) + 1$, the variance maximization problem (4.49) has the finite-dimensional reduction

$$\max_{\pi\in\mathcal{P}_x(\alpha)} \mathbb{E}_\pi\left[\|\varphi - \mathbb{E}_\pi[\varphi]\|^2\right] = \max_{\pi\in\mathcal{P}_x^m(\alpha)} \mathbb{E}_\pi\left[\|\varphi - \mathbb{E}_\pi[\varphi]\|^2\right]. \tag{4.54}$$

Therefore one can compute a saddle point (d^α, π^α) of the game (4.9) (=(4.86)) as

$$\pi^\alpha = \sum_{i=1}^m w_i \delta_{\theta_i} \text{ and } d^\alpha = \sum_{i=1}^m w_i \varphi(\theta_i), \tag{4.55}$$

where $w_i \geq 0, \theta_i \in \Theta$, $i = 1, \ldots, m$ maximize

$$\begin{cases} \text{Maximize } \sum_{i=1}^m w_i \left\|\varphi(\theta_i)\right\|^2 - \left\|\sum_{i=1}^m w_i \varphi(\theta_i)\right\|^2 \\ \text{Subject to } \ w_i \geq 0, \theta_i \in \Theta, i = 1, \ldots, m, \ \sum_{i=1}^m w_i = 1 \\ \bar{p}(x|\theta_i) \geq \alpha, \quad i = 1, \ldots, m\,. \end{cases} \tag{4.56}$$

As a consequence of Theorems 4.3.3 and 4.3.5, a measure with finite support $\mu := \sum w_i \delta_{z_i}$ on V is the pushforward under $\varphi : \Theta \to V$ of an optimal measure π^α for the maximum variance problem (4.49) if and only if, as illustrated in Figure 4.2, it is supported on the intersection of $\varphi(\Theta_x(\alpha))$ and the boundary ∂B of the minimum enclosing ball of $\varphi(\Theta_x(\alpha))$ and the center z^* of B is the center of mass $z^* = \sum w_i z_i$ of the measure μ.

Relaxing MLE with an accuracy/robustness tradeoff

For fixed $x \in X$, assume that the model P is such that the maximum likelihood estimate (MLE)

$$\theta^* := \arg\max_{\theta \in \Theta} p(x|\theta) \tag{4.57}$$

of $\theta^\dagger$ exists and is unique.

Observe that for α near one (1) the support of π^α and d^α concentrate around the MLE θ^* and $\varphi(\theta^*)$, (2) the risk $\mathcal{R}(d^\alpha)$ =(4.10) concentrates around zero, and (3) the confidence $1 - \beta_\alpha$ associated with the rarity assumption $\theta^\dagger \in \Theta_x(\alpha)$ is the smallest. In that limit, our estimator inherits the accuracy and lack of robustness of the MLE approach to estimating the quantity of interest.

Conversely for α near zero, since by (4.7) $\Theta_x(\alpha) \approx \Theta$, (1) the support of the pushforward of π^α by φ concentrates on the boundary of $\varphi(\Theta)$ and d^α concentrate around the center of the minimum enclosing ball of $\varphi(\Theta)$, (2) the risk $\mathcal{R}(d^\alpha)$ =(4.50) is the highest and concentrates around the worst-case risk (4.18), and (3) the confidence $1 - \beta_\alpha$ associated with the rarity assumption $\theta^\dagger \in \Theta_x(\alpha)$ is the highest. In that limit, our estimator inherits the robustness and lack of accuracy of the worst-case approach to estimating the quantity of interest.

For α between 0 and 1, the proposed game-theoretic approach induces a minmax optimal tradeoff between the accuracy of MLE and the robustness of the worst case.

4.4 Computational framework

The introduction developed this framework in the context of a model P with density p in terms of a single sample x. In Section 4.1, the single sample case was extended to N i.i.d. samples by defining the multisample $\mathcal{D} := (x_1, \ldots, x_N)$ and defining the product model density $p(\mathcal{D}|\theta) := \Pi_{i=1}^N p(x_i|\theta)$. Extensions incorporating correlations in the samples, such as Markov or other stochastic processes can easily be developed. Here we continue this development for the general model of the introduction for the ℓ^2 loss and also develop more fully a Gaussian noise model.

Later, in Sections 4.6 and 4.6 these models will be tested on estimating a quadratic function and a Lotka-Volterra predator-prey model based on noisy observations. In Section 4.7 the framework will be generalized to more general loss functions and rarity assumptions, which much of the current section generalizes to.

Let the possible states of nature be a compact subset $\Theta \subset \mathbb{R}^k$, the decision space be $V := \mathbb{R}^n$ and the elements of the N-fold multisample

$$\mathcal{D} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$$

lie in X, that is, $\mathcal{D}$ lies in the multisample space X^N. Let the n components of the quantity of interest $\varphi : \Theta \to \mathbb{R}^n$ be indicated by $\varphi_t : \Theta \to \mathbb{R}$, $t = 1, \ldots n$. Here, using the i.i.d. product model $P(\cdot|\theta)$ with density $p(\mathcal{D}|\theta) := \Pi_{i=1}^N p(x_i|\theta)$, the definition of β_α in (4.34) becomes

$$\begin{aligned} \beta_\alpha &:= \sup_{\theta\in\Theta} P\Big(\big\{\mathcal{D}' \in X^N : \theta \notin \Theta_{\mathcal{D}'}(\alpha)\big\}\Big|\theta\Big) \\ &= \sup_{\theta\in\Theta} \int \mathbb{1}_{\{\bar{p}(\cdot|\theta)<\alpha\}}(\mathcal{D}')p(\mathcal{D}'|\theta)d\nu^N(\mathcal{D}')\,, \end{aligned} \tag{4.58}$$

where, for fixed θ, $\mathbb{1}_{\{\bar{p}(\cdot|\theta)<\alpha\}}$ is the indicator function of the set $\{\mathcal{D}' \in X^N : \bar{p}(\mathcal{D}'|\theta) < \alpha\}$. We use boldface, such as $\boldsymbol{x}_i$ or $\boldsymbol{\theta}$, to emphasize the vector nature of variables and functions in the computational framework.

In this notation, the finite dimensional reduction guaranteed by Theorem 4.3.5 in (4.55) and (4.56) of the optimization problem (4.49) defining a worst-case measure of the form $\pi^\alpha := \sum_{i=1}^m w_i \delta_{\boldsymbol{\theta}_i}$ takes the form

$$\begin{array}{ll} \underset{\{(w_i,\boldsymbol{\theta}_i)\}_{i=1}^m}{\text{maximize}} & \sum_{t=1}^n \left(\sum_{i=1}^m \varphi_t^2\left(\boldsymbol{\theta}_i\right) w_i - \sum_{i,j=1}^m w_i \varphi_t\left(\boldsymbol{\theta}_i\right) \varphi_t\left(\boldsymbol{\theta}_j\right) w_j\right) \\ s.t. & \boldsymbol{\theta}_i \in \Theta,\, w_i \geqslant 0, \quad i = 1, \ldots, m \\ & \sum_{i=1}^m w_i = 1, \\ & \bar{p}\left(\mathcal{D}|\boldsymbol{\theta}_i\right) \geqslant \alpha, i = 1, \ldots, m, \end{array} \tag{4.59}$$

where the component of the objective function

$$var\left(\varphi_t\right) := \sum_{i=1}^m \varphi_t^2\left(\boldsymbol{\theta}_i\right) w_i - \sum_{i,j=1}^m w_i \varphi_t\left(\boldsymbol{\theta}_i\right) \varphi_t\left(\boldsymbol{\theta}_j\right) w_j$$

is the variance of the random variable $\varphi_t : \Theta \to \mathbb{R}$ under the measure $\pi := \sum_{i=1}^m w_i \delta_{\boldsymbol{\theta}_i}$.

Algorithm for solving the game

We are now prepared to develop an algorithm for player II (the decision maker) to play the game (4.47), using the saddle point Theorem 4.3.2 and the finite dimensional reduction Theorem 4.3.5 after selecting the rarity parameter α quantifying the rarity assumption (4.7) in terms of the relative likelihood (4.31) or a surrogate as described in Remark 4.3.1, to be the largest α such that the significance β_α (4.58) at α satisfies $\beta_\alpha \leq \beta^*$, the significance level.

At a high level the algorithm for computing a worst-case measure, its resulting risk (variance) and optimal estimator is as follows:

1. Observe a multisample $\mathcal{D}$.
2. Find the largest α such that β_α defined in (4.58) satisfies $\beta_\alpha \leq \beta^*$.
3. Solve (4.59) determining a worst-case measure $\pi^\alpha := \sum_{i=1}^m w_i \delta_{\boldsymbol{\theta}_i}$.
4. Output the Risk as the value of (4.59).
5. Output optimal decision $d^\alpha := \sum_{i=1}^m w_i \varphi(\boldsymbol{\theta}_i)$.

To solve (4.59) in Step 3 we apply the duality of the variance maximization problem with the minimum enclosing ball problem, Theorem 4.3.3, to obtain the following complete algorithm. It uses Algorithm 8 for computing the minimum enclosing ball about the (generally) infinite set $\varphi(\Theta_x(\alpha))$, which in turn uses a minimum enclosing ball algorithm *Miniball* applied to sets of size at most $dim(V) + 2$, see e.g. Welzl (Welzl, 1991), Yildirim (Yildirim, 2008) and Gartner (Gärtner, 1999). Here we use that of Welzl (Welzl, 1991). See Section 4.5 for a discussion and a proof in Theorem 4.5.1 of the convergence of Algorithm 8. Theorem 4.5.1 also establishes a convergence proof when the distance maximization Step 8a in Algorithm 7 is performed approximately. Note that the likelihood region $\Theta_{\mathcal{D}}(\alpha)$ is defined by

$$\Theta_{\mathcal{D}}(\alpha) := \left\{\boldsymbol{\theta} \in \Theta : p(\mathcal{D}|\boldsymbol{\theta}) \geq \alpha\, p(\mathcal{D}|\boldsymbol{\theta}^*)\right\},$$

where

$$\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}}\ p\big(\mathcal{D}|\boldsymbol{\theta}\big)$$

is a MLE.

Algorithm 7 UQ4K algorithm

1. Inputs:
   a) Multisample $\mathcal{D} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$
   b) ε_0
   c) Significance level β^*
2. Find MLE $\boldsymbol{\theta}^*$ by $\boldsymbol{\theta}^* \in \arg\max_{\boldsymbol{\theta}} \; p\left(\mathcal{D}|\boldsymbol{\theta}\right)$
3. Find the largest α such that β_α defined in (4.58) satisfies $\beta_\alpha \leq \beta^*$
4. $\boldsymbol{c} \leftarrow \boldsymbol{\varphi}\left(\boldsymbol{\theta}^*\right)$
5. $S \leftarrow \{\boldsymbol{\varphi}\left(\boldsymbol{\theta}^*\right)\}$
6. $\rho_0 \leftarrow 0$
7. $e \leftarrow 2\varepsilon_0$
8. While $e \geqslant \varepsilon_0$
   a) $$\begin{aligned} \bar{\boldsymbol{\theta}} \;\; &\in \;\; \arg\max_{\boldsymbol{\theta}} \left\|\boldsymbol{\varphi}\left(\boldsymbol{\theta}\right) - \boldsymbol{c}\right\|^2 \\ &s.t. \;\; p(\mathcal{D}|\boldsymbol{\theta}) \geq \alpha \, p(\mathcal{D}|\boldsymbol{\theta}^*) \end{aligned}$$
   b) if $\left\|\boldsymbol{\varphi}\left(\bar{\boldsymbol{\theta}}\right) - \boldsymbol{c}\right\| \geqslant \rho_0$
      i. $S \leftarrow S \cup \left\{\boldsymbol{\varphi}\left(\bar{\boldsymbol{\theta}}\right)\right\}$
   c) $\boldsymbol{c}, \rho \leftarrow Miniball\left(S\right)$
   d) $e = |\rho - \rho_0|$
   e) $\rho_0 \leftarrow \rho$
   f) if $|S| > n + 1$
      i. find subset $S' \subset S$ of size $n + 1$ such that $Miniball\left(S'\right) = Miniball\left(S\right)$
      ii. $S \leftarrow S'$
9. Find $\{w_i\}_{i=1}^{n+1}$ from $\underset{\{w_i\}_{i=1}^{n+1}}{\text{maximize}} \sum_{t=1}^{n} \left(\sum_{i=1}^{n+1} \varphi_t^2\left(\boldsymbol{\theta}_i\right) w_i - \sum_{i,j=1}^{n+1} w_i \varphi_t\left(\boldsymbol{\theta}_i\right) \varphi_t\left(\boldsymbol{\theta}_j\right) w_j\right)$

Large sample simplifications

Here we demonstrate that when the number of samples N is large, under classic regularity assumptions, the significance β_α is approximated by the value of a chi-squared distribution, substantially simplifying the determination of α in Step 3 of Algorithm 7.

Let $\Theta \subseteq \mathbb{R}^k$ and let data be generated by the model at the value $\theta \in \Theta$. Under standard regularity conditions, check (G. Casella and R. L. Berger, 2002, Sec. 10.6.2 & Thm. 10.1.12), the maximum likelihood estimator (MLE), $\widehat{\theta}_N$, is asymptotically efficient for θ. That is as the sample size $N \to \infty$

$$\sqrt{N}(\widehat{\theta}_N - \theta) \xrightarrow{\mathrm{d}} N(0, I(\theta)^{-1}), \tag{4.60}$$

where $I(\theta)$ is the Fisher information matrix. Therefore, standard arguments, see (G. Casella and R. L. Berger, 2002, Thm. 10.3.1), for the asymptotic distribution of the likelihood ratio test result in the following approximation of β_α.

Theorem 4.4.1. Let $\Theta \subseteq \mathbb{R}^k$ and assume that the model density p satisfies the regularity conditions of (G. Casella and R. L. Berger, 2002, Section. 10.6.2). Then

$$\beta_\alpha \to 1 - \chi_k^2\big(2 \ln \frac{1}{\alpha}\big) \tag{4.61}$$

as $N \to \infty$, where χ_k^2 is the chi-square distribution with k degrees of freedom.

Consequently, under these conditions Step 3 of Algorithm 7 can take the simple form

(Step 3): Solve for α satisfying $\beta_\alpha := 1 - \chi_k^2\big(2 \ln \frac{1}{\alpha}\big) = \beta^*$.

Algorithm 7 for a Gaussian noise model

Consider a Gaussian noise model where, $X = \mathbb{R}^r$ and for $\boldsymbol{\theta} \in \Theta$, the components of the multisample $\mathcal{D} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{rN}$ are i.i.d. samples from the Gaussian distribution $\mathcal{N}(\boldsymbol{m}(\boldsymbol{\theta}), \sigma^2 I_r)$, with mean $\boldsymbol{m}(\boldsymbol{\theta})$ and covariance $\sigma^2 I_r$, where $\boldsymbol{m} : \Theta \to \mathbb{R}^r$ is a *measurement function*, $\sigma > 0$ and I_r is the r-dimensional identity matrix. The measurement function $\boldsymbol{m}$ is a function such that the its value $\boldsymbol{m}(\theta)$ can be computed when the model parameter $\boldsymbol{\theta}$ is known. Therefore the i.i.d. multisample $\mathcal{D} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is drawn from $\big(\mathcal{N}(\boldsymbol{m}(\boldsymbol{\theta}), \sigma^2 I_r)\big)^N$ and so has the probability density

$$p(\mathcal{D}|\boldsymbol{\theta}) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{rN}} \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{N} \left\|\boldsymbol{x}_j - \boldsymbol{m}(\boldsymbol{\theta})\right\|^2\right) \tag{4.62}$$

with respect to the Lebesgue measure ν on $X := \mathbb{R}^{rN}$, and defining $\left(\sigma\sqrt{2\pi}\right)^{rN}$ times the maximum likelihood

$$M(\mathcal{D}) := \exp\left(-\frac{1}{2\sigma^2}\inf_{\boldsymbol{\theta}\in\Theta}\left(\sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2\right)\right), \tag{4.63}$$

the relative likelihood (4.31) is

$$\bar{p}\left(\mathcal{D}|\boldsymbol{\theta}\right) = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2\right)}{M(\mathcal{D})}. \tag{4.64}$$

Taking the logarithm of the constraint $\bar{p}\left(\mathcal{D}|\boldsymbol{\theta}_i\right) \geqslant \alpha$ defining the likelihood region $\Theta_{\mathcal{D}}(\alpha)$, using (4.64) we obtain

$$\Theta_{\mathcal{D}}(\alpha) = \left\{\boldsymbol{\theta}\in\Theta : \sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 \leqslant M_\alpha\right\} \tag{4.65}$$

in terms of

$$M_\alpha := \inf_{\boldsymbol{\theta}\in\Theta}\sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 + 2\sigma^2\ln\frac{1}{\alpha}. \tag{4.66}$$

Consequently, for the Gaussian case, the worst-case measure optimization problem (4.59) becomes

$$\begin{array}{ll}\underset{\{(w_i,\boldsymbol{\theta}_i)\}_{i=1}^m}{\text{maximize}} & \sum_{t=1}^{n}\left(\sum_{i=1}^{m}\varphi_t^2\left(\boldsymbol{\theta}_i\right)w_i - \sum_{i,j=1}^{m}w_i\varphi_t\left(\boldsymbol{\theta}_i\right)\varphi_t\left(\boldsymbol{\theta}_j\right)w_j\right)\\ s.t. & \boldsymbol{\theta}_i\in\Theta,\ w_i\geqslant 0,\quad i=1,\ldots,m\\ & \sum_{i=1}^{m}w_i = 1,\\ & \sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}_i\right)\right\|^2 \leqslant M_\alpha,\quad i=1,\ldots,m.\end{array} \tag{4.67}$$

Consequently, in the Gaussian noise case, Algorithm 7 appears with these modifications:

1. (Step 2): Find MLE $\boldsymbol{\theta}^*$ by $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}}\sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2$
2. (Step 8a): Solve

$$\begin{array}{rcl}\bar{\boldsymbol{\theta}} & \in & \arg\max_{\boldsymbol{\theta}}\left\|\boldsymbol{\varphi}\left(\boldsymbol{\theta}\right) - \boldsymbol{c}\right\|^2\\ & s.t. & \sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 \leqslant M_\alpha.\end{array} \tag{4.68}$$

Farthest point optimization in the Gaussian model

In Step 8a of Algorithm 7 we seek the farthest point $\bar{\boldsymbol{\theta}}$ from a center $\boldsymbol{c}$:

$$\begin{aligned} \bar{\boldsymbol{\theta}} \quad &\in \quad \arg\max_{\boldsymbol{\theta}} \|\boldsymbol{\varphi}(\boldsymbol{\theta}) - \boldsymbol{c}\|^2 \\ &\quad s.t. \quad \sum_{j=1}^{N} \left\|\boldsymbol{x}_j - \boldsymbol{m}(\boldsymbol{\theta})\right\|^2 \leqslant M_\alpha. \end{aligned} \tag{4.69}$$

To solve this optimization, we use the merit function technique (Nocedal and S. Wright, 2006) as follows:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} - \|\boldsymbol{\varphi}(\boldsymbol{\theta}) - \boldsymbol{c}\|^2 + \mu \max\left\{0, \sum_{j=1}^{N} \left\|\boldsymbol{x}_j - \boldsymbol{m}(\boldsymbol{\theta})\right\|^2 - M_\alpha\right\}. \tag{4.70}$$

In implementation, one should start with a small value of μ and increase it to find the optimum (Nocedal and S. Wright, 2006). The first term in (4.70) intends to increase the distance from the center $\boldsymbol{c}$ and the second term keeps the solution feasible. Any algorithm picked to solve (4.70) must be able to slide near the feasibility region $\sum_{j=1}^{N} \left\|\boldsymbol{x}_j - \boldsymbol{m}(\boldsymbol{\theta})\right\|^2 \leqslant M_\alpha$ to guarantee a better performance. Suggestions of such algorithms are the gradient descent (Kingma and Ba, 2014), if the gradients are available and differential evolution (Storn and Price, 1997) if gradients are not available.

Surrogate relative likelihoods

Although the computation of the maximum likelihood $\boldsymbol{\theta}^*$ in Step 2 of Algorithm 7 is only done once; for the observed data $\mathcal{D}$, the computation of β_α in Step 3 requires it to be computed for all data $\mathcal{D}'$ generated by the statistical model. Simplification of this computation can be obtained by large sample N approximations (see Section 4.4) or the utilization of a surrogate relative likelihood as discussed in Remark 4.3.1, which we now address.

Let the generic multisample be $\mathcal{D}' := (x'_1, \ldots, x'_N)$ in the computation of β_α in (4.58), and consider the upper bound on the maximum likelihood of the Gaussian

noise model (4.62)

$$\begin{aligned}\sup_{\theta\in\Theta} p\left(\mathcal{D}'|\boldsymbol{\theta}\right) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{rN}} \sup_{\theta\in\Theta} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2\right)\\ &\leq \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{rN}} \sup_{\boldsymbol{m}\in\mathbb{R}^r} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \boldsymbol{m}\right\|^2\right)\\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{rN}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}'_k\right\|^2\right),\end{aligned}$$

so that the resulting surrogate relative likelihood (using the same symbol as the relative likelihood) discussed in Remark 4.3.1 becomes

$$\bar{p}\left(\mathcal{D}'|\boldsymbol{\theta}\right) = \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}'_k\right\|^2\right)}, \tag{4.71}$$

and therefore the condition $\bar{p}\left(\cdot|\boldsymbol{\theta}\right) < \alpha$ in the computation of the surrogate significance $\beta'_\alpha \geq \beta_\alpha$ defined in (4.58) in terms of the surrogate relative likelihood (4.71) in Step 3 appears as

$$\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 - \sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}'_k\right\|^2 > 2\sigma^2 \ln\frac{1}{\alpha}. \tag{4.72}$$

Rewriting in terms of the $N(0, \sigma^2 I_r)$ Gaussian random variables

$$\epsilon_i := \boldsymbol{x}'_j - \boldsymbol{m}(\boldsymbol{\theta}), i = 1, \ldots, N$$

we obtain

$$\begin{aligned}\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 - \sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}'_k\right\|^2 &= \sum_{j=1}^{N}\left\|\epsilon_j\right\|^2 - \sum_{j=1}^{N}\left\|\epsilon_j - \frac{1}{N}\sum_{k=1}^{N}\epsilon_k\right\|^2\\ &= 2\sum_{j=1}^{N}\left\langle\epsilon_j, \frac{1}{N}\sum_{k=1}^{N}\epsilon_k\right\rangle - \left\|\frac{1}{N}\sum_{k=1}^{N}\epsilon_k\right\|^2\\ &= (2N-1)\left\|\frac{1}{N}\sum_{k=1}^{N}\epsilon_k\right\|^2,\end{aligned}$$

that is

$$\sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 - \sum_{j=1}^{N}\left\|\boldsymbol{x}'_j - \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{x}'_k\right\|^2 = (2N-1)\|v\|^2, \tag{4.73}$$

where

$$v := \frac{1}{N}\sum_{k=1}^{N} \epsilon_k$$

is Gaussian with mean zero and, since the ϵ_k are i.i.d, have covariance $\sigma^2 I_r$, that is $v \in N(0, \frac{\sigma^2}{N} I_r)$. Since Schott (Schott, 2016, Thm. 9.9) implies that $\frac{N}{\sigma^2}\|v\|^2$ is distributed as χ^2_r, it follows from (4.72), (4.73) and the definition of the surrogate significance β'_α (4.58) that

$$\beta'_\alpha = 1 - \chi^2_r\big(\frac{2N}{2N-1} \ln \frac{1}{\alpha}\big) \geq \beta_\alpha. \tag{4.74}$$

Consequently, removing the prime indicating the surrogate significance β'_α, denoting it as β_α, the modifications (4.68) to Algorithm 7 are augmented to

1. (Step 2): Find MLE $\boldsymbol{\theta}^*$ by $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^{N} \left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2$
2. (Step 3): Solve for α satisfying $\beta_\alpha := 1 - \chi^2_r\big(\frac{2N}{2N-1} \ln \frac{1}{\alpha}\big) = \beta^*$
3. (Step 8a): Solve

$$\begin{aligned} \bar{\boldsymbol{\theta}} \quad &\in \quad \arg\max_{\boldsymbol{\theta}} \|\boldsymbol{\varphi}\left(\boldsymbol{\theta}\right) - \boldsymbol{c}\|^2 \\ &\quad s.t. \quad \sum_{j=1}^{N} \left\|\boldsymbol{x}_j - \boldsymbol{m}\left(\boldsymbol{\theta}\right)\right\|^2 \leqslant M_\alpha. \end{aligned} \tag{4.75}$$

Stochastic processes

We now consider the case where $\Theta \subseteq \mathbb{R}^k$ and the data is the multisample $\mathcal{D} := (x_1, \ldots, x_N)$ with $x_i = (y_i, t_i) \in \mathcal{Y} \times \mathcal{T}$ where y_i corresponds to the observation of a stochastic process at time t_i. Letting θ parameterize the distribution of the stochastic process and assuming the y_i to be independent given θ and the t_i, the model density takes the form (A) $p(\mathcal{D}|\theta) := \Pi_{i=1}^N p(x_i|\theta, t_i) q(t_i)$ if the t_i are assumed to be i.i.d. with distribution Q (and density q with respect to some given base measure on $\mathcal{T}$), (B) and $p\big((x_1, \ldots, x_N)\big|\theta, (t_1, \ldots, t_N)\big) := \Pi_{i=1}^N p(x_i|\theta, t_i)$ if the t_i are assumed to be arbitrary. Observe that the model densities for cases (A) and (B) are proportional and, as a consequence, given the t_i (arbitrary or sampled), they share the same likelihood region $\Theta_{\mathcal{D}}(\alpha) = \big\{\theta \in \Theta : \Pi_{i=1}^N p(y_i|\theta, t_i) \geq \alpha \sup_{\theta'} \Pi_{i=1}^N p(y_i|\theta', t_i)\big\}$. Let $Q_N := \frac{\delta_{t_1} + \cdots + \delta_{t_N}}{N} \in \mathcal{P}(\mathcal{T})$ be the empirical probability distribution defined by $(t_1, \ldots, t_n)$, and assume $\mathcal{T}$ to be a compact subset of a finite dimensional Euclidean space. The following theorem indicates the result of Theorem 4.4.1 remains valid in case (B) if $Q_N \to Q$ (e.g. when $\mathcal{T} = [0, 1]$ and Q is the uniform distribution and the $t_i = i/N$).

Theorem 4.4.2. Assume that the model density $(y,t) \to p(y|\theta,t)q(t)$ satisfies the regularity conditions of (G. Casella and R. L. Berger, 2002, Section. 10.6.2), and that $Q_N \to Q$ (in the sense of weak convergence) as $N \to \infty$. Then in both cases (A) and (B) the limit $\beta_\alpha \to 1 - \chi^2_k\big(2\ln\frac{1}{\alpha}\big)$ holds true as $N \to \infty$.

4.5 Minimum enclosing ball algorithm

Let $K \subset \mathbb{R}^n$ be a compact subset and let $B \supset K$, with center z and radius R, be the smallest closed ball containing K. Together Theorem 4.3.3 and Theorem 4.7.2 demonstrate that the minimum enclosing ball exists and is unique. The problem of computing the minimum enclosing ball has received a considerable amount of attention, beginning with Sylvester (Sylvester, 1857) in 1857. Probably the most cited method is that of Welzl (Welzl, 1991), which, by (Welzl, 1991, Thm. 2), achieves the solution in expected $O((n+1)(n+1)!|K|)$ time, where $|K|$ is the cardinality of the set K. Yildirim (Yildirim, 2008) provides two algorithms which converge to an ϵ-approximate minimum enclosing ball in $O(\frac{|K|n}{\epsilon})$ computations and provides a historical review of the literature along with extensive references.

Although Yildirim does address the infinite K situation, we provide a new algorithm, Algorithm 8, based on that of Bădoiu, Har-Peled and Indyk (Bădoiu, Har-Peled, and Indyk, 2002, p. 251), to approximately compute the minimum enclosing ball B containing a (possibly infinite) compact set K in $\mathbb{R}^n$, using the approximate computation of maximal distances from the set K to fixed points in $\mathbb{R}^n$. To that end, let MINIBALL denote an existing algorithm for computing the minimum enclosing ball for sets of size $\leq n+2$. As we will demonstrate, the FOR loop in Algorithm 8 always gets broken at Step 10 for some x since, by Caratheodory's theorem (see e.g. (Rockafellar, 1970)) a minimum enclosing ball in n dimensions is always determined by $n+1$ points.

For $\delta \geq 0$, and a function $f : X \to \mathbb{R}$, let $\arg\max^\delta f$ denote a δ-approximate maximizer in the following sense; $x^* \in \arg\max^\delta f$ if

$$f(x^*) \geq \frac{1}{1+\delta} \sup_{x \in X} f(x).$$

For, $\epsilon > 0$, the ϵ-enlargement $B^{(1+\epsilon)}(x,r)$ of a closed ball $B(x,r)$ with center x and radius r is the closed ball $B(x,(1+\epsilon)r)$. In the following algorithm, δ is a parameter quantifying the degree of optimality of distance maximizations and ϵ is a parameter specifying the accuracy required of the produced estimate to the minimum enclosing ball. The following theorem demonstrates that Algorithm 8

Algorithm 8 Miniball algorithm

Inputs: $\epsilon \in [0, 1), \delta \geq 0$, K and MINIBALL for sets of size $\leq n + 2$
$(x_\alpha, x_\beta) \leftarrow \arg\max^{\delta}_{x,x' \in K} \|x - x'\|$
$A_0 \leftarrow \{x_\alpha, x_\beta\}$
$0 \leftarrow k$ (iteration counter)
repeat
  $B_k \leftarrow \text{MINIBALL}(A_k)$
  while $|A_k| > n + 1$ **do**
    for $x \in A_k$ **do**
      $B^x_k \leftarrow \text{MINIBALL}(A_k \setminus \{x\})$
      if $B^x_k = B_k$ then $A_k \leftarrow A_k \setminus \{x\}$ and break loop
    end for
  end while
  $z_k \leftarrow \text{Center}(B_k)$
  $x_{k+1} \leftarrow \arg\max^{\delta}_{x \in K} \|x - z_k\|$
  $A_{k+1} \leftarrow A_k \cup \{x_{k+1}\}$
  $k \leftarrow k + 1$
until $x_k \in B^{(1+\epsilon)}_{k-1}$
return $B^{(1+\epsilon)(1+\delta)}_{k-1}, A_{k-1}$

produces an approximation with guaranteed accuracy to the minimum enclosing ball in a quantified finite number of steps.

Theorem 4.5.1. For a compact subset $K \subset \mathbb{R}^n$, let R denote the radius of the minimum enclosing ball of K. Then, for $\epsilon \in [0, 1), \delta \geq 0$, Algorithm 8 converges to a ball B^*, satisfying

$$B^* \supset K$$

and

$$R(B^*) \leq (1 + \epsilon)(1 + \delta)R$$

in at most $\frac{16}{\epsilon^2}(1 + 2\delta)$ steps of the REPEAT loop. Moreover, the size of the working set A_k is bounded by

$$|A_k| \leq \min\left(2 + \frac{16}{\epsilon^2}(1 + 2\delta), n + 2\right)$$

for all k.

4.6 Examples

Gaussian Mean Estimation

Consider the problem of estimating the mean $\theta^\dagger$ of a Gaussian distribution $\mathcal{N}(\theta^\dagger, \sigma^2)$ with known variance $\sigma^2 > 0$ from the observation of one sample x from that

distribution and from the information that $\theta^\dagger \in [-\tau, \tau]$ for some given $\tau > 0$. Note that this problem can be formulated in the setting of Problem 1 by letting (1) $P(\cdot|\theta)$ be the Gaussian distribution on $X := \mathbb{R}$ with mean θ and variance σ^2, (2) $\Theta := [-\tau, \tau]$ and $V := \mathbb{R}$ and (3) $\varphi : \Theta \to V$ be the identity map $\varphi(\theta) = \theta$. The relative likelihood (4.31) is

$$\bar{p}(x|\theta) = \frac{e^{-\frac{1}{2\sigma^2}|x-\theta|^2}}{\sup_{\theta\in\Theta} e^{-\frac{1}{2\sigma^2}|x-\theta|^2}} \tag{4.76}$$

with the supremum in the denominator achieved at the closest $\theta \in \Theta$ to x (x itself if $x \in [-\tau, \tau]$). This defines the likelihood region $\Theta_x(\alpha) := \{\theta \in \Theta : \bar{p}(x|\theta) \geq \alpha\}$. A simple calculation yields, for the case $x \in [-\tau, \tau]$

$$\Theta_x(\alpha) = \left[\max\left(-\tau, x - \sqrt{2\sigma^2 \ln(1/\alpha)}\right),\ \min\left(\tau, x + \sqrt{2\sigma^2 \ln(1/\alpha)}\right)\right]. \tag{4.77}$$

Using Theorem 4.3.5 with $m = dim(V) + 1 = 2$, for $\alpha \in [0, 1]$, one can compute a saddle point (π^α, d^α) of the game (4.47) as

$$\pi^\alpha = w\delta_{\theta_1} + (1-w)\delta_{\theta_2} \text{ and } d^\alpha = w\theta_1 + (1-w)\theta_2, \tag{4.78}$$

where w, θ_1, θ_2 maximize the variance

$$\begin{cases} \text{Maximize} & w\theta_1^2 + (1-w)\theta_2^2 - (w\theta_1 + (1-w)\theta_2)^2 \\ \text{over} & 0 \leq w \leq 1, \quad \theta_1, \theta_2 \in [-\tau, \tau] \\ \text{subject to} & \frac{(x-\theta_i)^2}{2\sigma^2} \leq \ln\frac{1}{\alpha}, \qquad i = 1, 2, \end{cases} \tag{4.79}$$

where the last two constraints are equivalent to the rarity assumption $\theta_i \in \Theta_x(\alpha)$.

Hence for α near 0, $\Theta_x(\alpha) = \Theta = [-\tau, \tau]$, and by Theorem 4.3.3, the variance is maximized by placing each Dirac on each boundary point of the region Θ, each receiving half of the total probability mass, that is by $\theta_1 = -\tau$, $\theta_2 = \tau$ and $w = 1/2$, in which case $\operatorname{Var}\pi^\alpha = \tau^2$ and $d^\alpha = 0$. For $\alpha = 1$, the rarity constraint implies $\theta_1 = \theta_2 = x$ when $x \in [-\tau, \tau]$, leading to the MLE $d^\alpha = x$ with $\operatorname{Var}\pi^\alpha = 0$. Note that from (4.34) we have

$$\beta_\alpha = \sup_{\theta\in[-\tau,\tau]} \mathbb{P}_{x'\sim\mathcal{N}(\theta,\sigma^2)}\left[\bar{p}(x'|\theta) < \alpha\right]$$

which can be computed analytically for this example using 4.76 and separating into the three cases $x < -\tau$, $x \in [-\tau, \tau]$ and $x > \tau$. We illustrate in Figure 4.6 the different results of solving the optimization problem (4.79) in the case $\sigma^2 = 1$,

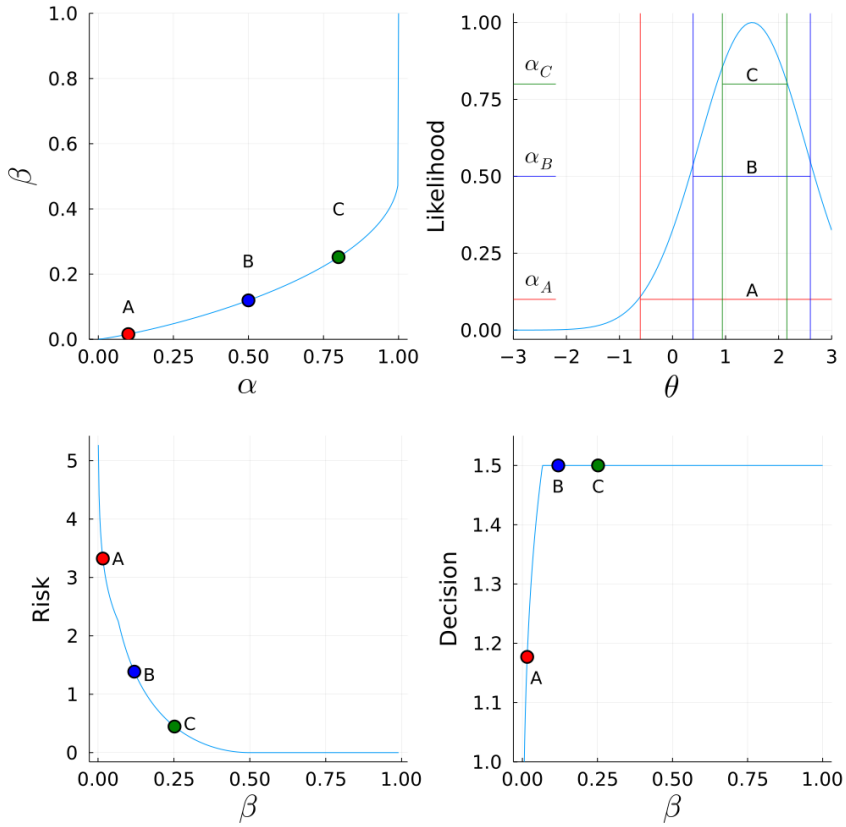


Figure 4.6: $\alpha - \beta$ relation, likelihood level sets, risk value and decision for different choices of α (and consequently β) for the normal mean estimation problem with $\tau = 3$ and observed value $x = 1.5$. Three different values in the $\alpha - \beta$ curve are highlighted across the plots

$x = 1.5$ and $\tau = 3$. We plot the $\alpha - \beta$ curve (top left), the likelihood of the model in $[-\tau, \tau]$ and the $\alpha-$level sets (top right), and the evolutions of the risk with β (bottom left), and the optimal decision with β (bottom right). Since, by Theorem 4.3.3, the optimal decision is the midpoint of the interval with extremes in either the $\alpha-$level sets or $\pm\tau$, we observe that for low β, our optimal decision does not coincide with the MLE.

Estimation of a quadratic function

The measurement function $\boldsymbol{m}$ of Section 4.6, being defined as the solution of the Lotka-Volterra predator-prey model as a function of its parameters $\boldsymbol{\theta} \in \Theta$, does not appear simple to differentiate and therefore `SciPy`'s version of Storn and Price's (Storn and Price, 1997) Differential Evolution optimizer (Virtanen et al., 2020) was used to perform the farthest point optimization problem in Step 8a in Algorithm 7. In this section, we test this framework on a problem which does not possess this

complication: estimating the parameters $\boldsymbol{\theta} := (\theta_0, \theta_1, \theta_2)$ of a quadratic function

$$\boldsymbol{m}(t; \boldsymbol{\theta}) := \theta_0 + \theta_1 t + \theta_2 t^2$$

on a uniform grid T of the interval $(0, 5)$ consisting of 100 points, using noisy observational data. In this case, we can use automatic differentiation in the merit function technique of Section 4.4 to perform the farthest point optimization problem in Step 8a using gradient descent methods via automatic differentiating modules available in packages like autograd, or computing the gradient and applying a gradient descent method.

We proceed as in Section 4.6 with $\Theta := [-30, 30]^3$ and assume that, given $\boldsymbol{\theta} \in \Theta$, a single sample path $\mathcal{D} := \{\boldsymbol{x}\} := \left(\boldsymbol{x}^{(t)}\right)_{t=1}^{100}$ is generated on the grid T to the stochastic process

$$\boldsymbol{x}^{(t)} = \boldsymbol{m}(t; \boldsymbol{\theta}) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \sigma^2), \quad t \in T, \tag{4.80}$$

with $\sigma^2 := 10$. Consequently $X := \mathbb{R}^{100}$. We let the decision space be $V = \mathbb{R}^3$ and the quantity of interest $\varphi : \Theta \to \mathbb{R}^3$ be the identity function.

For the experiment, we generate a single ($N := 1$) full sample path $\mathcal{D} := \{\boldsymbol{x}\} := \left(\boldsymbol{x}^{(t)}\right)_{t=1}^{100}$ according to (4.80) at the unknown values $\boldsymbol{\theta}^* := (\theta_1^*, \theta_2^*, \theta_3^*) = (1, .5, 1)$. As discussed in Section 4.4, we tune μ in the merit function (4.70), and set gradient descent with adaptive moment estimation optimizer (Paszke et al., 2019) with parameters (e.g. learning rate = 0.001, max epochs=50,000) to achieve full convergence. We observe the convergence plots for the increment of θs from the ball center as diagnostic.

Figure 4.7 shows the β vs α relationship defined by the surrogate significance (4.74) derived from the surrogate likelihood method with $N := 1$ and $r := 100$, the risk as function of α, and the likelihood regions, their minimum enclosing balls and the optimal decisions (centers of the balls), for $\alpha \in [0, 1]$. As can be seen the optimal decisions, being the centers of the minimum enclosing balls, do not move and only the size of the minimum enclosing balls change, resulting in various risk values associated with the same optimal estimates.

Finally, Figure 4.7 shows the results for the maximum likelihood solution and the two supporting points of the minimum enclosing balls for the case that $\beta_\alpha = \beta^* = 0.05$. From this experiment, we obtained the optimal decision $d^* = (0.24, 1.22, 0.89)$ along with the two support points $S = \{(-2.27, 3.52, 0.48), (2.75, -1.10, 1.31)\}$ of the minimum enclosing ball. For the sake of comparison, we also performed the

same experimentation with `SciPy`'s version of Storn and Price's (Storn and Price, 1997) Differential Evolution optimizer (Virtanen et al., 2020) at the default settings, to perform the farthest point optimization problem in Step 8a in Algorithm 7, using the merit function (4.70) of Section 4.4, and obtained similar results.

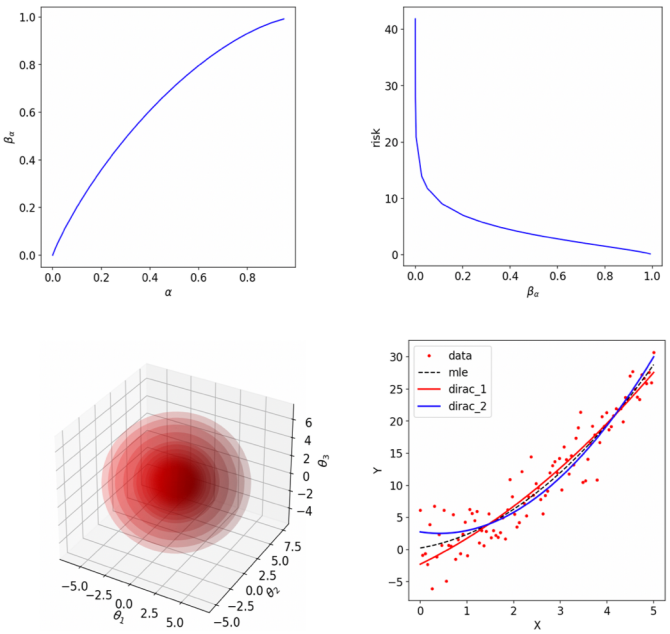


Figure 4.7: Quadratic model results: (left-top) β_α vs α, (right-top) risk vs β_α, (left-bottom) the supporting surfaces with the decision points at the center of each surface and (right-bottom) maximum-likelihood solution and supporting points of the minimum enclosing ball for $\beta_\alpha = \beta^* = .05$

Estimation of a Lotka-Volterra predator-prey model

Here we implement Algorithm 7 for the Gaussian noise model of Section 4.4, where the measurement function $(\theta_1, \theta_2) \mapsto \boldsymbol{m}(\theta_1, \theta_2)$ is defined as the solution map of the Lotka-Volterra (Lotka, 1920) predator-prey model

$$\frac{dx}{dt} = \theta_1 x - \eta xy \tag{4.81}$$

$$\frac{dy}{dt} = \xi xy - \theta_2 y, \tag{4.82}$$

evaluated on the uniform time grid $T := \{t_i\}_{i=1}^{200}$ such that $t_0 = 0$ and $t_{200} = 20$, with fixed and known parameters η, ξ and initial data x_0, y_0, describing the evolution of a prey population with variable x and a predator population with variable y. As such, denoting $\boldsymbol{\theta} := (\theta_1, \theta_2) \in \Theta$, we denote the solution map

$$\boldsymbol{\theta} \mapsto \boldsymbol{m}(t; \boldsymbol{\theta}), t \in T,$$

by $\boldsymbol{m} : \Theta \to (\mathbb{R}^2)^T$. For the probabilistic model, we let $\Theta := [-5, 5]^2$ and assume the Gaussian model (4.64) with $N = 1$, where the data $\mathcal{D}$ consists of a single sample path $\mathcal{D} := \{\boldsymbol{x}\} := \left(\boldsymbol{x}^{(t)}\right)_{t=1}^{200}$ of the T-indexed stochastic process

$$\boldsymbol{x}^{(t)} = \begin{bmatrix} x_t \\ y_t \end{bmatrix} = \boldsymbol{m}(t; \boldsymbol{\theta}) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \quad t \in T, \tag{4.83}$$

where $\mathbf{I} \in \mathbb{R}^{2\times 2}$ is the identity matrix and $\sigma = 5$. Note that in the notation of (4.64) we have

$$\|\boldsymbol{x} - \boldsymbol{m}\,(\boldsymbol{\theta})\,\|^2 = \sum_{t\in T} \|\boldsymbol{x}^{(t)} - \boldsymbol{m}\,(t; \boldsymbol{\theta})\,\|^2.$$

Let the decision space be $V := \mathbb{R}^2$ and let the quantity of interest $\varphi : \Theta \to \mathbb{R}^2$ be the identity. For the experiment, we generate one sample path $\mathcal{D} := \{\boldsymbol{x}\} := \left(\boldsymbol{x}^{(t)}\right)_{t=1}^{200}$ according to (4.83) at the unknown values $\boldsymbol{\theta}^* := (\theta_1^*, \theta_2^*) = (0.55, 0.8)$, with $x_0 = 30$, $y_0 = 10$, $\eta = 0.025$ and $\xi = 0.02$ known. We consider the evolution $t \mapsto \boldsymbol{m}(t; \boldsymbol{\theta}^*)$, $t \in T$, the *true predator-prey* values and the sample path $\left(\boldsymbol{x}^{(t)}\right)_{t=1}^{200}$ as noisy observations of it. The resulting time series $\mathcal{D}$ is shown in the top image of Figure 4.8.

For $\alpha \in [0, 1]$, by taking the logarithm of the defining relation (4.7) of the likelihood region $\Theta_{\mathcal{D}}(\alpha)$, we obtain the representation (4.65),

$$\Theta_{\mathcal{D}}(\alpha) = \left\{\boldsymbol{\theta} \in \Theta : \sum_{t\in T} \left\|\boldsymbol{x}^{(t)} - \boldsymbol{m}\,(t; \boldsymbol{\theta})\right\|^2 \leqslant M_\alpha\right\}$$

in terms of

$$M_\alpha := \inf_{\boldsymbol{\theta}\in\Theta} \sum_{t\in T} \left\|\boldsymbol{x}^{(t)} - \boldsymbol{m}\,(t; \boldsymbol{\theta})\right\|^2 + 2\sigma^2 \ln \frac{1}{\alpha}\,,$$

for the likelihood region $\Theta_{\mathcal{D}}(\alpha)$ in terms of the data $\mathcal{D}$.

To determine α at significance level $\beta^* := .05$, we approximate the significance β_α defined in (4.34) using the chi-squared approximation (4.74) and then select α to be the value such that this approximation yields $\beta_\alpha = \beta^* = .05$. The validity of this approximation for this example is additionally demonstrated in the right image of Figure 4.10, which shows via Monte Carlo simulation that the $1 - \beta_\alpha$ versus α curve is well characterized by the χ_2^2 distribution.

Having selected α, to implement Algorithm 7, we need to select an optimizer for Step 8a. Instead of computing the Jacobian of the solution map $\boldsymbol{m}$, here we utilize the gradient-free method of `SciPy`'s version of Storn and Price's (Storn and

Price, 1997) Differential Evolution optimizer (Virtanen et al., 2020) at the default settings. Given the data generating value $\boldsymbol{\theta}^* = (0.55, 0.8)$, the primary feasible region $\Theta := [-5, 5]^2$ is sufficiently non-suggestive of the the data generating value $\boldsymbol{\theta}^*$. Finally, since $dim(V) = 2$, Algorithm 7 produces: a set S of at most three boundary points of $\Theta_{\mathcal{D}}(\alpha)$, the minimum enclosing ball B of $\Theta_{\mathcal{D}}(\alpha)$, its center as the optimal estimate of $\boldsymbol{\theta}^*$, and the weights of the set S corresponding to a worst-case measure, optimal for the variance maximization problem (4.49). The results are displayed in Figure 4.9.

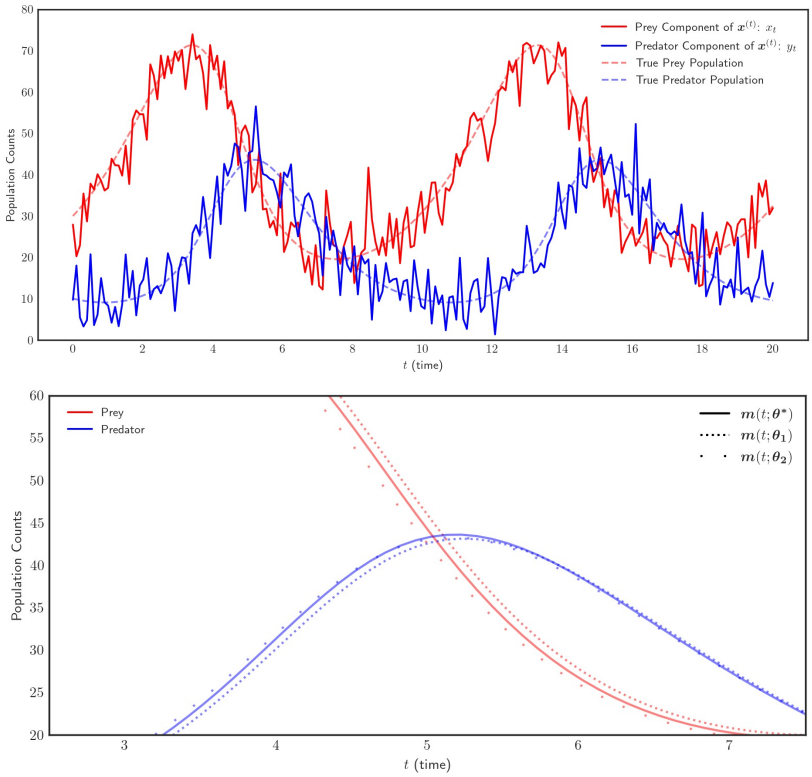


Figure 4.8: **(Top image)** Data $\mathcal{D} := \boldsymbol{x}^{(t)}, t \in T$, generated, according to the Gaussian noise model (4.83): solid red is the prey component and solid blue the predator component of the generated data $\boldsymbol{x}^{(t)}$, the dotted red and dotted blue are the prey-predator components of the Lotka-Volterra solution $\boldsymbol{m}(t, \boldsymbol{\theta}^*)$ for $t \in T$. **(Bottom image)** Uncertainty in the population dynamics corresponding to the worst-case measure: (1) red is prey and blue is predator, (2) solid line is the Lotka-Volterra evolution $\boldsymbol{m}(t, \boldsymbol{\theta}^*), t \in T$, fine dots $\boldsymbol{m}(t, \boldsymbol{\theta}_1), t \in T$, and coarse dots $\boldsymbol{m}(t, \boldsymbol{\theta}_2)$, where $S := \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ is the set of support points of the worst case (posterior) measure (located on the boundary of the minimum enclosing ball).

To get a sense of the output uncertainty of $\boldsymbol{m}(\cdot)$ with these optimized results, we plot the predator and prey population dynamics associated with each optimized

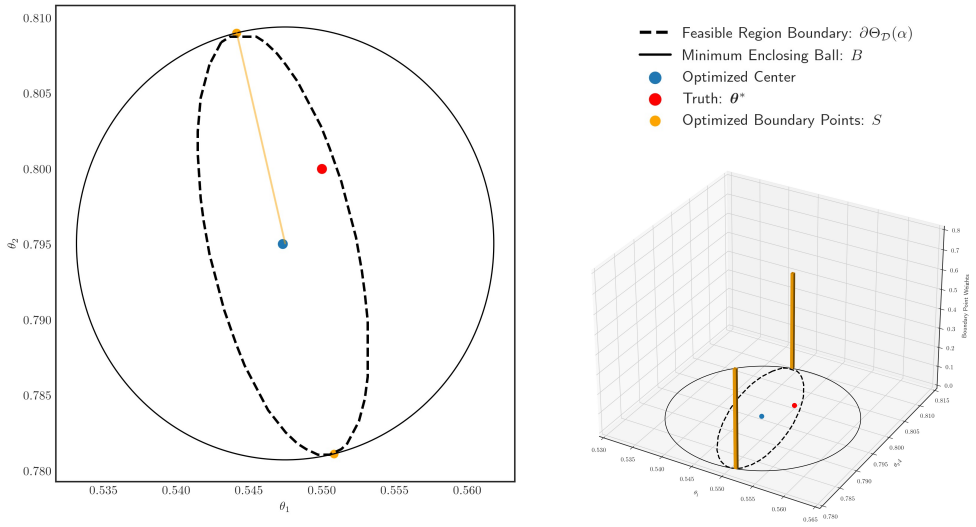


Figure 4.9: Minimum enclosing ball for the Lotka-Volterra Model: (left) the dashed line indicates the boundary of the likelihood region $\Theta_{\mathcal{D}}(\alpha)$, the solid circle its minimum enclosing ball, the red point the data generating value $\boldsymbol{\theta}^*$ and the blue point the center of the minimum enclosing ball and the optimal estimate of $\boldsymbol{\theta}^*$. The two yellow points comprise the set $S := \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$. (right) a projected view with the yellow columns indicating the weights $(.5, .5)$ of the set S in their determination of a worst-case (posterior) measure.

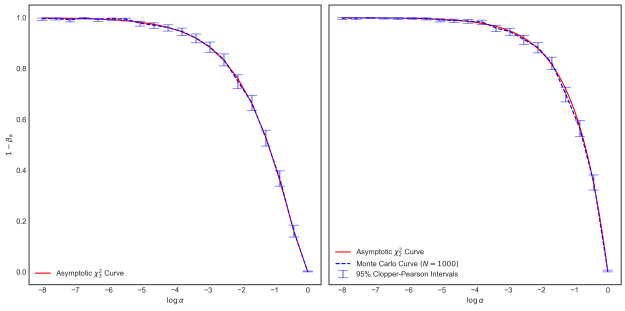


Figure 4.10: Monte Carlo numerically confirms the result from Theorem 4.4.1 in both the quadratic function estimation **(left image)** and Lotka-Volterra **(right image)** examples.

boundary point of $\Theta_{\mathcal{D}}(\alpha)$ in Figure 4.8. This figure shows that with significance value $\beta_\alpha = 0.05$, the optimized boundary points create population dynamics in a tight band around the true population dynamics.

4.7 General loss functions and rarity assumptions

Here we generalize the framework introduced in Section 4.2 to allow more general loss functions than the ℓ^2 loss, in for example Equations (4.18), (4.19), and (4.23), and more general rarity assumptions than (4.33).

The discussions of worst-case, robust Bayes, and Wald's statistical decision theory generalize in a straightforward manner, so we focus on generalizing the current UQ of the 4th kind. Let $\ell : V \times V \to \mathbb{R}_+$ be a loss function. In addition to the pointwise rarity assumption (4.33), consider an *integral rarity assumption* $\mathcal{P}_x^\Psi(\alpha) \subset \mathcal{P}(\Theta)$ determined by a real-valued function Ψ, defined by

$$\mathcal{P}_x^\Psi(\alpha) := \left\{\pi \in \mathcal{P}(\Theta) : \int_\Theta \Psi\big(\bar{p}(x|\theta)\big) d\pi(\theta) \geq \alpha\right\} \tag{4.84}$$

generalizing (4.29). Note that for Ψ the identity function $\mathcal{P}_x(\alpha) \subset \mathcal{P}_x^\Psi(\alpha)$ and by comparison with (4.29) it follows from the following remark that such integral rarity assumptions can also alleviate the brittleness of Bayesian inference.

Remark 4.7.1. Jensen's inequality implies that

$$\Psi\Big(\int_\Theta \bar{p}(x|\theta) d\pi(\theta)\Big) \leq \int_\Theta \Psi\big(\bar{p}(x|\theta))\big) d\pi(\theta)$$

when the function Ψ is convex, and

$$\Psi\Big(\int_\Theta \bar{p}(x|\theta) d\pi(\theta)\Big) \geq \int_\Theta \Psi\big(\bar{p}(x|\theta))\big) d\pi(\theta)$$

when the function Ψ is concave, such as when Ψ is a logarithm. Consequently, when Ψ is concave and strictly increasing, the assumption

$$\int_\Theta \Psi\big(\bar{p}(x|\theta)\big) d\pi(\theta) \geq \alpha$$

implies that the denominator in the conditional measure (4.46) satisfies

$$\int_\Theta \bar{p}(x|\theta) d\pi(\theta) \geq \Psi^{-1}(\alpha)\,.$$

Consequently, such constraints, by keeping the denominator in the conditional measure bound away from zero, stabilize the numerical computation of the conditional measure in the numerical computation of a worst-case measure.

The following applies equally as well for the pointwise rarity assumption (4.33) and the integral rarity assumption (4.84). For simplicity of exposition, we restrict to the pointwise rarity assumption. Generalizing (4.45), consider playing a game using the loss

$$\mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi_x}\big[\ell(\varphi(\theta), d)\big], \quad \pi \in \mathcal{P}_x(\alpha), d \in V\,. \tag{4.85}$$

For the pointwise rarity assumption, the same logic following (4.45) implies that this game is equivalent to the generalization (4.47) to a game using the loss

$$\mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi}\big[\ell(\varphi(\theta), d)\big], \quad \pi \in \mathcal{P}_x(\alpha), d \in V\,. \tag{4.86}$$

For the integral rarity assumption, we maintain the form (4.85). β_α is defined before as in (4.34) and the selection of $\alpha \in [0, 1]$ is as before.

Under the mild conditions of Theorem 4.8.1, we can show that, for each $\alpha \in [0, 1]$, a maxmin optimal solution π^α of $\max_{\pi\in\mathcal{P}_x(\alpha)} \min_{d\in V} \mathcal{L}(\pi, d)$ can be computed. Moreover, by Theorem 4.8.1, it also follows that the saddle function $\mathcal{L}$ in (4.86) satisfies the conditions of Sion's minmax theorem (Sion, 1958), resulting in a minmax result for the game (4.86):

$$\min_{d\in V} \max_{\pi\in\mathcal{P}_x(\alpha)} \mathcal{L}(\pi, d) = \max_{\pi\in\mathcal{P}_x(\alpha)} \min_{d\in V} \mathcal{L}(\pi, d)\,. \tag{4.87}$$

Consequently if, for each $\pi \in \mathcal{P}_x(\alpha)$, we select

$$d_\pi \in \arg\min_{d\in V} \mathcal{L}(\pi, d), \tag{4.88}$$

then a worst-case measure for the game (4.86), solving the maxmin problem on the right-hand side of (4.87), satisfies

$$\pi^\alpha \in \arg\max_{\pi\in\mathcal{P}_x(\alpha)} \mathcal{L}(\pi, d_\pi)\,. \tag{4.89}$$

Moreover let $d^\alpha \in \arg\min_{d\in V} \max_{\pi\in\mathcal{P}_x(\alpha)} \mathcal{L}(\pi, d)$ denote any solution to the minmax problem on the left-hand side of (4.87). Then it is well known that the minmax equality (4.87) implies that the pair (π^α, d^α) is a saddle point of $\mathcal{L}$ in that

$$\mathcal{L}(\pi, d^\alpha) \le \mathcal{L}(\pi^\alpha, d^\alpha) \le \mathcal{L}(\pi^\alpha, d), \qquad \pi \in \mathcal{P}_x(\alpha),\ d \in V\,. \tag{4.90}$$

Since the solution to

$$d_{\pi^\alpha} := \arg\min_{d\in V} \mathcal{L}(\pi^\alpha, d) \tag{4.91}$$

is uniquely defined under the conditions of Theorem 4.7.2 (identical to those of Theorem 4.8.1), it follows from the right-hand side of the saddle equation (4.3) that $d_{\pi^\alpha} = d^\alpha$ and $(\pi^\alpha, d_{\pi^\alpha})$ is a saddle point of $\mathcal{L}$, that is we have

$$\mathcal{L}(\pi, d_{\pi^\alpha}) \leq \mathcal{L}(\pi^\alpha, d_{\pi^\alpha}) \leq \mathcal{L}(\pi^\alpha, d), \qquad \pi \in \mathcal{P}_x(\alpha),\ d \in V\,. \tag{4.92}$$

Moreover, its associated risk is the value

$$\mathcal{R}(d_{\pi^\alpha}) := \mathcal{L}(\pi^\alpha, d_{\pi^\alpha}) \tag{4.93}$$

in (4.92) of the two person game defined in (4.86), which is the same for all saddle points of $\mathcal{L}$.

Finite-dimensional reduction

Let $\Delta^m(\Theta)$ denote the set of convex sums of m Dirac measures located in Θ and, let $\mathcal{P}_x^m(\alpha) \subset \mathcal{P}_x(\alpha)$ defined by

$$\mathcal{P}_x^m(\alpha) := \left\{\pi \in \Delta^m(\Theta) \cap \mathcal{P}_x(\alpha)\right\}, \tag{4.94}$$

denote the finite-dimensional subset of the rarity assumption set consisting of the convex combinations of m Dirac measures in $\mathcal{P}_x(\alpha)$. Then the following reduction Theorem 4.7.2 asserts that

$$\max_{\pi \in \mathcal{P}_x(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d) = \max_{\pi \in \mathcal{P}_x^m(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d), \tag{4.95}$$

for any $m \geq dim(V) + 2$. We note that the improvement to $m \geq dim(V) + 1$ when the loss is the ℓ^2 loss follows from the Lagrangian duality, Theorem 4.3.3, of the maximization problem (4.95) with the minimum enclosing ball and Caratheodory's theorem.

In the following theorem, applicable to both pointwise and integral rarity assumptions, we provide sufficient conditions that the computation of a worst-case measure in the optimization problem (4.89) can be reduced to a finite-dimensional one.

Theorem 4.7.2. Let Θ be compact, X be a measurable space, $\mathbb{P}$ be a positive dominated model such that, for each x, its likelihood function is continuous, and let $\varphi : \Theta \to \mathbb{R}$ be continuous. Let V be a finite-dimensional Euclidean space and let $\ell : V \times V \to \mathbb{R}_+$ be continuously differentiable, strictly convex and coercive in its second variable and vanishing along the diagonal. Let $\Delta^m(\Theta)$ denote the set of convex sums of m Dirac measures located in Θ and, for $\Psi : (0, 1] \to \mathbb{R}$ upper

semicontinuous and $x \in X$, consider both the pointwise rarity assumption subset $\mathcal{P}_x^m(\alpha) \subset \mathcal{P}_x(\alpha)$ defined by

$$\mathcal{P}_x^m(\alpha) := \left\{\pi \in \Delta^m(\Theta) \cap \mathcal{P}_x(\alpha)\right\}, \tag{4.96}$$

and the integral rarity assumption subset $\mathcal{P}_x^{\Psi,m}(\alpha) \subset \mathcal{P}_x^{\Psi}(\alpha)$, defined by

$$\mathcal{P}_x^{\Psi,m}(\alpha) := \left\{\pi \in \Delta^m(\Theta) : \int_\Theta \Psi\big(p(x|\theta)\big) d\pi(\theta) \geq \alpha\right\}, \tag{4.97}$$

where $p(x|\cdot) : \Theta \to (0, 1]$ is the relative likelihood. Then we have

$$\max_{\pi \in \mathcal{P}_x(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d) = \max_{\pi \in \mathcal{P}_x^m(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d), \tag{4.98}$$

for any $m \geq dim(V) + 2$, and

$$\max_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d) = \max_{\pi \in \mathcal{P}_x^{\Psi,m}(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d), \tag{4.99}$$

for any $m \geq dim(V) + 3$, unless Ψ is the identity function, when $m \geq dim(V) + 2$.

Remark 4.7.3. Theorem 4.7.2 easily generalizes to vector integral functions Ψ of more general form than (4.97), where the number of Diracs required is then $m \geq dim(V) + \dim(\Psi) + 1$ and $m \geq dim(V) + \dim(\Psi) + 2$ for the pointwise and integral cases, respectively.

(A. Shapiro and Kleywegt, 2002, Thm.2.1) implies one can generalize Theorem 4.7.2 to the more general class of loss functions ℓ which are coercive and convex in the second argument, but require more, $(3\big(\dim(V) + 1\big))$ Dirac measures in the integral case. See (A. Shapiro and Kleywegt, 2002, Prop. 3.1).

The following theorem generalizes the duality Theorem 4.3.3 to more general convex loss functions.

Theorem 4.7.4. Let $\varphi : \Theta_x(\alpha) \to V$ be continuous and suppose that the loss function satisfies function $\ell(v_1, v_2) = W(v_1 - v_2)$, where $W : V \to \mathbb{R}_+$ is non-negative, convex, and coercive. For $x \in X$ and $\alpha \in [0, 1]$, suppose that $\Theta_x(\alpha)$ is compact. Then $\varphi\big(\Theta_x(\alpha)\big) \subset V$ is compact. Let $\lambda \geq 0$ be the smallest value such that there exists a $z \in V$ with

$$\varphi(\Theta_x(\alpha)) + z \subset W^{-1}([0, \lambda]).$$

Then λ is the value of the maxmin problem defined by the game (4.86)

$$\lambda = \max_{\pi \in \mathcal{P}_x(\alpha)} \min_{d \in V} \mathbb{E}_{\theta \sim \pi}\left[\ell(\varphi(\theta), d)\right], \tag{4.100}$$

and π^* is maxmin optimal for it if and only if there exists

$$d_* \in \operatorname{argmin}_{d\in V} \mathbb{E}_{\theta\sim\pi^*}\left[\ell(\varphi(\theta), d)\right]$$

with

$$\operatorname{support}(\varphi_*\pi^*) \subset W^{-1}(\lambda) - d_*.$$

Remark 4.7.5. Pass (Pass, 2020) generalizes these results to the case where the decision space V is a not-necessarily affine metric space and the ℓ^2 distance is replaced by the metric.

4.8 Supporting theorems and proofs

Minmax theorem

Theorem 4.8.1. Consider the saddle function (4.86) for the pointwise and (4.85) for the integral rarity assumptions, respectively. Given the assumptions of Theorem 4.7.2 we have

$$\min_{d\in V} \max_{\pi\in\mathcal{P}_x(\alpha)} \mathcal{L}(\pi, d) = \max_{\pi\in\mathcal{P}_x(\alpha)} \min_{d\in V} \mathcal{L}(\pi, d)$$

and

$$\min_{d\in V} \max_{\pi\in\mathcal{P}_x^\Psi(\alpha)} \mathcal{L}(\pi, d) = \max_{\pi\in\mathcal{P}_x^\Psi(\alpha)} \min_{d\in V} \mathcal{L}(\pi, d)$$

for the pointwise and integral rarity assumptions, respectively.

Proof. We prove the result for the integral rarity assumption case only, the pointwise case being much simpler. The assumptions imply $\Psi(p(x|\cdot))$ is upper semicontinuous, implying that $\mathcal{P}_x^\Psi(\alpha) \subset \mathcal{P}(\Theta)$ is closed, see e.g. (Aliprantis and Border, 2006, Thm. 15.5). Since $\mathcal{P}(\Theta)$ is a compact subset of the space of signed measures in the weak topology, see e.g. (Aliprantis and Border, 2006, Thm. 15.22), it follows that $\mathcal{P}_x^\Psi(\alpha) \subset \mathcal{P}(\Theta)$ is compact. Consequently, to apply Sion's minmax theorem (Sion, 1958) it is sufficient to establish that the map $\mathcal{L}(\cdot, d) : \mathcal{P}(\Theta) \to \mathbb{R}$ is upper semicontinuous and quasiconcave for each $d \in V$ and the map $\mathcal{L}(\pi, \cdot) : V \to \mathbb{R}$ is lower semicontinuous and quasiconvex for each $\pi \in \mathcal{P}(\Theta)$. To that end, observe that since ϕ is continuous and Θ compact, the function $\ell(\phi(\cdot), d)$ is bounded and continuous for each $d \in V$. Fixing d, observe the positivity of the likelihood function implies that the set

$$\left\{\pi \in \mathcal{P}(\Theta) : \mathbb{E}_{\theta\sim\pi_x}[\ell(\phi(\theta), d)] \geq r\right\} = \left\{\pi \in \mathcal{P}(\Theta) : \mathbb{E}_{\theta\sim\pi}[p(x|\cdot)\ell(\phi(\theta), d)] \geq r\mathbb{E}_{\theta\sim\pi}[p(x|\cdot)]\right\}$$

is closed by the continuity of $\ell(\phi(\cdot), d)$ and $p(x|\cdot)$, see e.g. (Aliprantis and Border, 2006, Thm. 15.5). Moreover, one can show that the reverse inequality also produces a closed set. Additionally, since it is a linear condition it is convex and therefore the

function $\mathcal{L}(\cdot, d) : \mathcal{P}(\Theta) \to \mathbb{R}$ is upper and lower semicontinuous and quasiconcave for each $d \in V$. Moreover, fixing $\pi \in \mathcal{P}(\Theta)$, since the function ℓ is continuous and Θ is compact and ϕ is continuous, it follows that the function $\mathcal{L}(\pi, \cdot) : V \to \mathbb{R}$ is continuous and convex and therefore lower semicontinuous and quasiconvex. Consequently, Sion (Sion, 1958, Cor. 3.3) implies that

$$\inf_{d \in V} \sup_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \mathcal{L}(\pi, d) = \sup_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \inf_{d \in V} \mathcal{L}(\pi, d).$$

Since, for fixed d, inner optimization $\sup_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \mathcal{L}(\pi, d)$ is over of an upper semicontinuous function over a compact set, it achieves its supremum. Since have established in the proof of Theorem 4.7.2 that the inner optimization $\inf_{d \in v} \mathcal{L}(\pi, d)$ achieves its infimum, we can write

$$\inf_{d \in V} \max_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \mathcal{L}(\pi, d) = \sup_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \min_{d \in V} \mathcal{L}(\pi, d)$$

Since the inner minimum $\min_{d \in v} \mathcal{L}(\pi, d)$ is the minimum of a family upper semicontinuous functions, it produces an upper semicontinuous function, see e.g. (Aliprantis and Border, 2006, Lem. 2.41), since the maximization of the outer loop is over the compact set $\mathcal{P}_x^{\Psi}(\alpha)$, we conclude that the supremum on the righthand side is attained. Moreover, since the inner maximum $\max_{\pi \in \mathcal{P}_x^{\Psi}(\alpha)} \mathcal{L}(\pi, d)$ is the maximum over a family of continuous functions, and therefore lower semicontinuous functions, it follows that it produces a lower semicontinuous function. Restricting to the compact subset V^* from the proof of Theorem 4.7.2 we conclude the outer infimum is attained, thus establishing the assertion. □

Proof of Theorem 4.3.2

Since the ℓ^2 loss $(v_1, v_2) \mapsto \|v_1 - v_2\|^2$ is strictly convex and coercive in its second argument and vanishes on the diagonal, the assumptions imply that the saddle function $\mathcal{L}(\pi, d) := \mathbb{E}_{\theta \sim \pi}\big[\|\varphi(\theta) - d\|^2\big]$, $\pi \in \mathcal{P}_x(\alpha)$, $d \in V$, (4.47) satisfies the conditions of Theorem 4.7.2, so that it follows from Theorem 4.8.1 that $\mathcal{L}$ satisfies the minmax equality, in particular establishing the existence of a worst-case measure

$$\pi^* \in \arg \max_{\pi \in \mathcal{P}_x(\alpha)} \inf_{d \in V} \mathcal{L}(\pi, d) \tag{4.101}$$

and a worst-case decision

$$d^* \in \arg \min_{d \in V} \sup_{\pi \in \mathcal{P}_x(\alpha)} \mathcal{L}(\pi, d). \tag{4.102}$$

In addition to establishing the existence of saddle points, where a pair $\left(\pi^*, d^*\right) \in \mathcal{P}_x(\alpha) \times V$ is a saddle point of $\mathcal{L}$ if we have

$$\mathcal{L}(\pi, d^*) \leq \mathcal{L}(\pi^*, d^*) \leq \mathcal{L}(\pi^*, d), \quad \pi \in \mathcal{P}_x(\alpha),\, d \in V, \tag{4.103}$$

observe that Bertsekas et al. (Bertsekas, Nedić, and Ozdaglar, 2003, Prop. 2.6.1) assert that a pair $(\pi^*, d^*) \in \mathcal{P}_x(\alpha) \times V$ is a saddle point if and only if they are a worst-case measure and worst-case decision, respectively, as defined in (4.101) and (4.102).

Now let (π^*, d^*) be a saddle point. As demonstrated in the proof of Theorem 4.7.2, since the function $d \mapsto \|\varphi(\theta) - d\|^2$ is strictly convex for all θ, it follows that its expectation $d \mapsto \mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi}\left[\|\varphi(\theta) - d\|^2\right]$ is strictly convex. Moreover a minimizer

$$d^{**} \in \arg\min_{d\in V} \mathcal{L}(\pi^*, d)$$

exists and by strict convexity it is necessarily unique; see e.g. (Rockafellar and Wets, 1998). Consequently, by the right-hand side of the definition (4.103) of a saddle point it follows that $d^{**} = d^*$. Since π^* satisfying (4.101) is equivalent to it satisfying (4.48) and d^* satisfying the right-hand side of the definition (4.103) of a saddle point is equivalent to it satisfying (4.49), the assertion regarding the form (4.48) and (4.49) for saddle points is proved.

Let (π_1, d_1) and (π_2, d_2) be two saddle points. Then by the saddle relation (4.103) we have

$$L(\pi_2, d_2) \leq \mathcal{L}(\pi_2, d_1) \leq \mathcal{L}(\pi_1, d_1) \leq \mathcal{L}(\pi_1, d_2) \leq \mathcal{L}(\pi_2, d_2),$$

establishing equality of the value of the risk (4.50) for all saddle points.

Finally, since $d \mapsto \mathcal{L}(\pi, d)$ is strictly convex it follows that its maximum $d \mapsto \sup_{\pi\in\mathcal{P}_x(\alpha)} \mathcal{L}(\pi, d)$ is strictly convex, demonstrating the uniqueness of solutions to (4.102). Moreover, since $\mathcal{P}_x(\alpha)$ is convex, the mapping $\pi \mapsto \mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi}\left[\|\varphi(\theta) - d\|^2\right]$ is affine and therefore concave for all $d \in V$, and therefore its minimum $\pi \mapsto \inf_{d\in V} \mathcal{L}(\pi, d)$ is also concave. Consequently, the set of all worst-case measures, that is, maximizers of (4.101), is convex, establishing the final assertion.

Proof of Theorem 4.3.3

Since the relative likelihood is continuous, the likelihood region $\Theta_x(\alpha)$ is closed and therefore compact, and since φ is continuous, it follows that $\varphi\big(\Theta_x(\alpha)\big)$ is compact

and therefore measurable. According to Bonnans and Shapiro (Bonnans and A. Shapiro, 2000, Sec. 5.4.1), because the constraint function $(r, z, x) \mapsto \|x - z\|^2 - r^2$ is continuous, the Lagrangian of the minimum enclosing ball problem (4.51) is

$$L(r, z; \mu) := r^2 + \int_{\varphi(\Theta_x(\alpha))} \left(\|x - z\|^2 - r^2\right) d\mu(x), \quad r \in \mathbb{R},\ z \in V,\ \mu \in \mathcal{M}\left(\varphi(\Theta_x(\alpha))\right). \tag{4.104}$$

Define

$$\Psi(\mu) := \inf_{r \in \mathbb{R}, z \in V} L(r, z; \mu)$$

and observe that

$$\inf_{r \in \mathbb{R}} L(r, z; \mu) = \begin{cases} \int_{\varphi(\Theta_x(\alpha))} \|x - z\|^2 d\mu(x), & \int d\mu = 1 \\ -\infty, & \int d\mu \neq 1 \end{cases}$$

so that

$$\Psi(\mu) := \inf_{r \in \mathbb{R}, z \in V} L(r, z; \mu) = \begin{cases} \int_{\varphi(\Theta_x(\alpha))} \left\|x - \mathbb{E}_\mu[x]\right\|^2 d\mu(x), & \int d\mu = 1 \\ -\infty, & \int d\mu \neq 1 \end{cases}$$

and therefore the dual problem to the minimum enclosing ball problem (4.51) is

$$\max_{\mu \in \mathcal{M}(\varphi(\Theta_x(\alpha)))} \Psi(\mu) = \max_{\mu \in \mathcal{P}(\varphi(\Theta_x(\alpha)))} \mathbb{E}_\mu\left[\|x - \mathbb{E}_\mu[x]\|^2\right],$$

establishing the Lagrangian duality assertion.

Moreover, since $\Theta_x(\alpha)$ is compact it is Polish, that is Hausdorff and completely metrizable, and since $\varphi : \Theta_x(\alpha) \to \varphi\left(\Theta_x(\alpha)\right)$ is continuous (Aliprantis and Border, 2006, Thm. 15.14) asserts that $\varphi_* : \mathcal{P}\left(\Theta_x(\alpha)\right) \to \mathcal{P}(\varphi(\Theta_x(\alpha)))$ is surjective. The change of variables formula (Aliprantis and Border, 2006, Thm. 13.46) establishes that the objective function of (4.49) satisfies

$$\mathbb{E}_\pi\left[\|\varphi - \mathbb{E}_\pi[\varphi]\|^2\right] = \mathbb{E}_{\varphi_* \pi}\left[\|v - \mathbb{E}_{\varphi_* \pi}[v]\|^2\right],$$

so that the surjectivity of φ_* implies that the value of (4.49) is equal to

$$\max_{\nu \in \mathcal{P}(\varphi(\Theta_x(\alpha)))} \mathbb{E}_\nu\left[\|v - \mathbb{E}_\nu[v]\|^2\right]. \tag{4.105}$$

The primary assertions then follow from Lim and McCann's (Lim and McCann, 2021, Thm. 1) generalization of the one-dimensional result of Popoviciu (Popoviciu, 1935) regarding the relationship between variance maximization and the minimum enclosing ball of the domain $\varphi(\Theta_x(\alpha)) \subset V$.

Proof of Theorem 4.3.5

The proof of Theorem 4.7.2 using rarity assumptions of the current form (4.33) does not require that the likelihood function $p(x', \cdot)$ be continuous for all $x' \in X$ but only at x. It asserts the finite-dimensional reduction (4.54) for $m \geq dim(V) + 2$ Dirac measures. On the other hand, the duality Theorem 4.3.3 implies that the optimality of such a measure $\pi := \sum_{i=1}^{m} w_i \delta_{\theta_i}$ is equivalent to the images of these Diracs $\delta_{\varphi(\theta_i)}, i = 1, \ldots, m$ lying on the intersection $\varphi(\Theta_x(\alpha)) \cap \partial B$ of the image of the likelihood region and the boundary of its minimum enclosing ball B, and that the weights of these Diracs determine that the center of mass of this image measure $\varphi_* \pi$ is the center d^α of the ball B, expressed as the right-hand side of (4.55). Consequently, the center d^α is in the convex hull of the m points $\varphi(\theta_i), i = 1, \ldots, m$ and by Caratheodory's theorem, see e.g. (Rockafellar, 1970), d^α is in the convex hull of $dim(V) + 1$ of these points. Let $S \subset \{1, \ldots, m\}$ correspond to such a subset. Then by the if and only if characterization of duality Theorem 4.3.3 it follows that the subset $\varphi(\theta_i), i \in S$ of $dim(V) + 1$ image points, using the weights $w_i', i \in S$ defining this convex combination to be the center d^α, corresponds to an optimal measure $\pi' := \sum_{i \in S} w_i' \delta_{\theta_i}$, thus establishing the assertion.

Proof of Theorem 4.4.1

The following are standard results in the statistics literature, see (G. Casella and R. L. Berger, 2002). The idea is to write the Taylor expansion of the log-likelihood around the MLE, then consider properties of the MLE, and finally apply the law of large numbers and Slutsky's theorem.

For simplicity, first let $\theta \in \Theta \subseteq \mathbb{R}$. Since the second term on the right-hand side in the Taylor expansion

$$\sum_{i=1}^{N} \ln p(x_i|\theta) = \sum_{i=1}^{N} \ln p(x_i|\widehat{\theta}_N) + \sum_{i=1}^{N} \frac{\partial}{\partial \theta} \ln p(x_i|\theta)_{\theta=\widehat{\theta}_N} (\theta - \widehat{\theta}_N) + \frac{1}{2} \sum_{i=1}^{N} \frac{\partial^2}{\partial \theta^2} \ln p(x_i|\theta)_{\theta=\widehat{\theta}_N} (\theta - \widehat{\theta}_N)^2 + o_p(1)$$

of the log-likelihood around the MLE, $\widehat{\theta}_N$, vanishes by the first order condition of the MLE, we obtain

$$\sum_{i=1}^{N} \ln p(x_i|\theta) - \sum_{i=1}^{N} \ln p(x_i|\widehat{\theta}_N) = \frac{1}{2} \sum_{i=1}^{N} \frac{\partial^2}{\partial \theta^2} \ln p(x_i|\theta)_{\theta=\widehat{\theta}_N} (\theta - \widehat{\theta}_N)^2 + o_p(1),$$

which we write as

$$-2\sum_{i=1}^{N}\ln\big(\frac{p(x_i|\theta)}{p(x_i|\widehat{\theta}_N)}\big) = -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\theta^2}\ln p(x_i|\theta)_{\theta=\widehat{\theta}_N}(\sqrt{N}(\theta-\widehat{\theta}_N))^2 + o_p(1).$$

By the law of large numbers and the consistency of the MLE we have

$$-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2}{\partial\theta^2}\ln p(x_i|\theta)_{\theta=\widehat{\theta}_N} \xrightarrow{\mathrm{P}} -E_{X\sim P(\cdot|\theta)}\frac{\partial^2}{\partial\theta^2}\ln p(X|\theta) = I(\theta), \tag{4.106}$$

where $I(\theta) := -E_{X\sim P(\cdot|\theta)}\frac{\partial^2}{\partial\theta^2}\ln p(X|\theta)$ is the Fisher information and $\xrightarrow{\mathrm{P}}$ represents convergence in probability. By the asymptotic efficiency of the MLE (under our regularity assumptions), the variance $V_0(\theta)$ of the MLE $\widehat{\theta}_N$ is the inverse of the Fisher information. That is

$$I(\theta) = (V_0(\theta))^{-1}.$$

Moreover, under the regularity conditions of (G. Casella and R. L. Berger, 2002, Section. 10.6.2), we have

$$\frac{\sqrt{N}(\widehat{\theta}_N-\theta)}{\sqrt{V_0(\theta)}} \xrightarrow{\mathrm{d}} N(0,1),$$

where $\xrightarrow{\mathrm{d}}$ represents convergence in distribution, and therefore Slutsky's theorem implies

$$-2\ln\bar{p}(\mathcal{D}|\theta) = -2\sum_{i=1}^{N}\ln\left(\frac{p(x_i|\theta)}{p(x_i|\widehat{\theta}_N)}\right) \xrightarrow{\mathrm{d}} \chi^2_1.$$

In the more general case $\Theta\subset\mathbb{R}^k$, under the regularity conditions of (G. Casella and R. L. Berger, 2002, Section. 10.6.2), we have

$$\sqrt{N}(\widehat{\theta}_N-\theta) \xrightarrow{\mathrm{d}} N(0, I(\theta)^{-1}), \tag{4.107}$$

where $I(\theta)$ is the Fisher information matrix, and the same argument goes through obtaining

$$-2\ln\bar{p}(\mathcal{D}|\theta) \overset{a}{=} N(\theta-\widehat{\theta}_N)I(\theta)^{-1}(\theta-\widehat{\theta}_N) \xrightarrow{\mathrm{d}} \chi^2_k,$$

where $\overset{a}{=}$ represents asymptotic equality. Therefore, for large sample sizes, one may use the following approximation:

$$-2\ln\bar{p}(\mathcal{D}|\theta) \approx \chi^2_k.$$

Finally, the condition $\bar{p}(\mathcal{D}|\theta) < \alpha$ in the computation of β_α is the same as

$$-2\ln \bar{p}(\mathcal{D}|\theta) > 2\ln(\frac{1}{\alpha}) ,$$

so that consequently, for large sample sizes, β_α may be approximated by

$$\beta_\alpha \approx 1 - \chi^2_k\big(2\ln(\frac{1}{\alpha})\big) .$$

Proof of Theorem 4.4.2

For case (A) the proof is an application of Theorem 4.4.1. Case (B) follows from a direct adaptation of the proof of Theorem 4.4.1. The first main step (in this adaptation) is to use the convergence of Q_N and the Independence of the y_i given the t_i to replace (4.106) by

$$-\frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2}{\partial\theta^2} \ln p(y_i|\theta, t_i)_{\theta=\widehat{\theta}_N} \xrightarrow{\text{P}} -E_{t\sim Q, y\sim P(\cdot|\theta,t)} \frac{\partial^2}{\partial\theta^2} \ln p(y|\theta, t) . \tag{4.108}$$

The second main step is to derive the asymptotic consistency and normality (4.107) of the MLE in case (B). This can be done by adapting the proofs of (Dudley, 2009, Lec. 5).

Proof of Theorem 4.7.2

We prove the theorem for the integral rarity case only, the pointwise rarity case being much simpler. First note that Theorem 4.8.1 asserts that the saddle function $\mathcal{L}$ satisfies a minmax equality, in particular, that the inner loop $\min_{d\in V} \mathcal{L}(\pi, d)$ of the primary assertion (4.99) indeed has a solution. To analyze such a solution, recall, by (4.85), that $\mathcal{L}(\pi, d)$ is the expectation

$$\mathcal{L}(\pi, d) := \mathbb{E}_{\theta\sim\pi_x}[\ell(\phi(\theta), d)] . \tag{4.109}$$

It is easy to show that the expectation of a family of strictly convex functions is strictly convex, so that it follows, for fixed π, that $\mathcal{L}(\pi, d)$ is strictly convex. Since ϕ is continuous and Θ is compact, it follows that the image $\phi(\Theta) \subset V$ is compact and since ℓ is coercive in its second variable, continuous in its first and $\phi(\Theta)$ is compact, it follows that ℓ is uniformly coercive in θ; that is, for every $y \in \mathbb{R}_+$, there exists an $R \in \mathbb{R}_+$ such that $|d| \geq R \implies \ell(\phi(\theta), d) \geq y,\ \theta \in \Theta$. It follows that $\mathcal{L}(\pi, d) \geq y, |d| \geq R, \pi \in \mathcal{P}(\Theta)$. Since y is arbitrary and $\mathcal{L}$ is convex, it follows that $\mathcal{L}$ achieves its minimum and since it is strictly convex this minimum is achieved at a unique point d_π; see e.g. (Rockafellar and Wets, 1998).

Since ℓ is continuous in its first variable, ϕ is continuous, and Θ compact, it follows that $\ell(\phi(\theta), 0)$ is uniformly bounded in θ and therefore implies a uniform bound on $\mathcal{L}(\pi, 0), \pi \in \mathcal{P}(\Theta)$. Consequently, the coerciveness of $\mathcal{L}$ implies that we can uniformly bound the unique optima $d_\pi, \pi \in \mathcal{P}(\Theta)$.

Let $V^* \supset \phi(\Theta)$ be a closed cube in V containing an open neighborhood of the image $\phi(\Theta)$ and this feasible set of optima just discussed. Since ϕ is continuous, Θ is compact, and ℓ is continuously differentiable, it follows that $\nabla_d \ell(\phi(\theta), d)$ is uniformly bounded in both θ and $d \in V^*$. Consequently, the Leibniz theorem for differentiation under the integral sign, (see e.g. (Aliprantis and Burkinshaw, 1998, Thm. 24.5)), implies, for fixed π, that

$$\nabla_d \mathcal{L}(\pi, d) := \mathbb{E}_{\theta \sim \pi_x}\left[\nabla_d \ell(\phi(\theta), d)\right], \quad d \in V^* . \tag{4.110}$$

Consequently, the relation $0 = \nabla_d \mathcal{L}(\pi, d_\pi)$ at the unique minimum d_π of $\min_{d \in V} \mathcal{L}(\pi, d)$ implies that

$$\mathbb{E}_{\theta \sim \pi_x}\left[\nabla_d \ell(\phi(\theta), d_\pi)\right] = 0, \quad \pi \in \mathcal{P}(\Theta) . \tag{4.111}$$

The formula (4.46) for the conditional measure and the positivity of its denominator imply that we can write (4.111) as

$$\frac{1}{\int_\Theta p(x|\cdot) d\pi} \mathbb{E}_{\theta \sim \pi}\left[p(x|\theta) \nabla_d \ell(\phi(\theta), d_\pi)\right] = 0 \tag{4.112}$$

which is equivalent to

$$\mathbb{E}_{\theta \sim \pi}\left[p(x|\theta) \nabla_d \ell(\phi(\theta), d_\pi)\right] = 0 . \tag{4.113}$$

Consequently, adding the constraint (4.113), equivalent to the minimization problem, the maxmin problem on the left-hand side of (4.99) can be written

$$\begin{cases} \text{Maximize } \mathcal{L}(\pi, d) \\ \text{Subject to } \ \pi \in \mathcal{P}_x^\Psi(\alpha), \ d \in V^* \\ \mathbb{E}_{\theta \sim \pi}\left[p(x|\theta) \nabla_d \ell(\phi(\theta), d)\right] = 0 \end{cases} \tag{4.114}$$

which again using the conditional formula (4.46) and the definition of $\mathcal{L}(\pi, d)$ can be written

$$\begin{cases} \text{Maximize } \frac{1}{\int_\Theta p(x|\theta) d\pi(\theta)} \mathbb{E}_{\theta \sim \pi}\left[p(x|\theta) \ell(\phi(\theta), d)\right] \\ \text{Subject to } \ \pi \in \mathcal{P}_x^\Psi(\alpha), \ d \in V^* \\ \mathbb{E}_{\theta \sim \pi}\left[p(x|\theta) \nabla_d \ell(\phi(\theta), d)\right] = 0 \end{cases} \tag{4.115}$$

which, introducing a new variable, can be written

$$\begin{cases} \text{Maximize } \frac{1}{\epsilon}\mathbb{E}_{\theta\sim\pi}\left[p(x|\theta)\ell(\phi(\theta),d)\right] \\ \text{Subject to } \ \pi \in \mathcal{P}_x^{\Psi}(\alpha),\ d \in V^*, \epsilon > 0 \\ \mathbb{E}_{\theta\sim\pi}\left[p(x|\theta)\nabla_d\ell(\phi(\theta),d)\right] = 0, \quad \epsilon = \mathbb{E}_{\theta\sim\pi}[p(x|\theta)]\ . \end{cases} \tag{4.116}$$

Now fix $\epsilon > 0$ and $d \in V^*$ and consider the inner maximization loop

$$\begin{cases} \text{Maximize } \frac{1}{\epsilon}\mathbb{E}_{\theta\sim\pi}\left[p(x|\theta)\ell(\phi(\theta),d)\right] \\ \text{Subject to } \ \pi \in \mathcal{P}_x^{\Psi}(\alpha), \\ \mathbb{E}_{\theta\sim\pi}\left[p(x|\theta)\nabla_d\ell(\phi(\theta),d)\right] = 0, \quad \epsilon = \mathbb{E}_{\theta\sim\pi}[p(x|\theta)]\ . \end{cases} \tag{4.117}$$

Since this is linear optimization of the integration of a non-negative, and thus integrable function, with possible integral value $+\infty$ for all π, over the full simplex of probability measures subject to $\dim(V)+1$ linear equality constraints defined by integration against measurable functions, plus one linear inequality constraint defined by integration against a measurable function, (H. Owhadi, C. Scovel, T. J. Sullivan, et al., 2013b, Thm. 4.1), which uses von Weizsacker and Winkler (Weizsäcker and Winkler, 1979, Cor. 3) (see also (Karr, 1983)) which is applicable under more assumptions on the model $\mathbb{P}$, implies this optimization problem can be reduced to optimization over the convex combination of $dim(V)+3$ Dirac measures supported on Θ. Since the full problem is the supremum of such problems, using the compactness of the space $\mathcal{P}_x^{\Psi,m}(\alpha)$ in the weak topology, the primary assertion follows. When Ψ is the identity function one of the constraints disappears, and the assertion in that case follows.

Proof of Theorem 4.7.4

The proof follows from the invariance of the variance under φ_* and the equality of their maximum variance problems established in the proof of Theorem 4.3.3 and Lim and McCann's (Lim and McCann, 2021, Thm. 2) generalization of their ℓ^2 result (Lim and McCann, 2021, Thm. 1).

Proof of Theorem 4.5.1

First consider the $\epsilon > 0$ case. Our proof will use results from Bădoiu, Har-Peled and Indyk (Bădoiu, Har-Peled, and Indyk, 2002). Consider the REPEAT loop. As previously mentioned, the FOR loop always gets broken at Step 10 for some x since, by Theorem 4.3.3, the center of the ball must lie in the convex hull of the $n+2$ points, but by Caratheodory's theorem this center also lies in the convex hull of $n+1$

of those points, and Theorem 4.3.3 then asserts that this ball is also the minimum enclosing ball of those $n + 1$ points. Clearly, the breaking of the step implies that the elimination of the point does not change the current ball. Consequently, the only change in the current ball is through the discovery in Step 14 of a distant point and its addition to the working set A_k followed by the calculation in Step 6 of a new minimum ball containing this enlarged working set. Let $B(A_k)$ denote the minimum enclosing ball of A_k, $R(A_k)$ denote its radius, and, overloading notation, let us denote $R(A_k) := R(B_k)$. Then when a new point is added to A_k to obtain A_{k+1}, it follows from $A_k \subset A_{k+1}$ that $B(A_{k+1}) \supset A_{k+1} \supset A_k$ and therefore $R(A_{k+1}) \geq R(A_k)$. Likewise $A_k \subset K$ implies that $R(A_k) \leq R$, the radius of the minimum enclosing ball B of K. Consequently the sequence $R(A_k) \leq R$ of the radii of the balls is monotonically increasing and bounded by R. Moreover, (Bădoiu, Har-Peled, and Indyk, 2002, pp. Clm. 2.4), using (Bădoiu, Har-Peled, and Indyk, 2002, Lem. 2.2) from Goel et al. (Goel, Indyk, and Varadarajan, 2001), implies that, until the stopping criterion in Step 17 is satisfied, we have

$$R(A_{k+1}) \geq \big(1 + \frac{\epsilon^2}{16}\big)R(A_k), \tag{4.118}$$

and when the stopping criterion is satisfied it follows from Step 14 that the output in Step 18 satisfies

$$B_{k-1}^{(1+\epsilon)(1+\delta)} \supset K. \tag{4.119}$$

Observe that we have $R \leq \Delta$ where $\Delta := \mathrm{diam}(K)$ and the initialization implies that $R(A_0) \geq \frac{1}{2(1+\delta)}\Delta$. Consequently (4.118) implies that the radius $R(A_k)$ increases by at least $\frac{\epsilon^2}{32((1+\delta)}\Delta$ at each step. Since the sequence is bounded by $R \leq \Delta$ it follows that at most $\frac{16}{\epsilon^2}(1 + 2\delta)$ of the REPEAT loop can be taken before terminating at Step 17. Upon termination the returned ball $B^* := B_{k-1}^{(1+\epsilon)(1+\delta)}$ in Step 18, by (4.119), satisfies

$$B^* \supset K,$$

and since

$$R(B_{k-1}^{(1+\epsilon)(1+\delta)}) = (1 + \epsilon)(1 + \delta)R(B_{k-1}) \leq (1 + \epsilon)(1 + \delta)R$$

we obtain

$$R\big(B^*\big) \leq (1 + \epsilon)(1 + \delta)R,$$

establishing the primary assertion. Since each step in the REPEAT loop adds at most one new point to the working set A_k, it follows that the working set size is bounded

by 2 plus the number of steps in the REPEAT loop, that is $2 + \frac{16}{\epsilon^2}(1 + 2\delta)$. Since the WHILE loop keeps the bound $\leq n + 2$, the proof is finished.

Now consider the $\epsilon = 0$ case. First let $\delta = 0$. By compactness of the set K, there exists a sub-sequence (A_{t_1}), indexed by $T_1 \subseteq \mathbb{N}$, of (A_t) such that $C(A_{t_1}) \to C_1$, in our notation whenever the algorithm stops or we reach to a fixed point the sequence repeats the last set of points. For every $t_1 \in T_1$, let y^{t_1} be the point selected as a furthest point from $B(A_{t_1})$ in the algorithm to form A_{t_1+1}. Again by compactness of the set K, there exists a sub-sequence (y^{t_2}), indexed by $T_2 \subseteq T_1 \subseteq \mathbb{N}$, of (y^{t_1}) such that $y^{t_2} \to y^*$. By another application of compactness of K, there exists a sub-sequence (A_{t_3}), indexed by $T_3 \subseteq T_2 \subseteq T_1 \subseteq \mathbb{N}$, of (A_{t_2}) such that $C(A_{t_3} \cup \{y^{t_3}\}) \to C_2$. Since $C(A_{t_3}) \to C_1$ and $R(A_{t_3}) \uparrow R_0 \leq R(K)$ as $t_3 \in T_3 \to \infty$, it follows that $B(A_{t_3}) \to B(C_1, R_0)$.

To complete the proof it is sufficient to show that $K \subseteq B(C_1, R_0)$, since then $R_0 \leq R(K)$ implies that $B(C_1, R_0) = B(K)$. To that end, we demonstrate that

$$d := \max_{x \in K} dist(x, B(C_1, R_0)) = 0.$$

Since $dist()$ is a continuous function in both of its arguments, $B(A_{t_3}) \to B(C_1, R_0)$ as $t_3 \in T_3 \to \infty$, $y^{t_3^k} \in \text{argmax}_{x \in K}\, dist(x, B(A_{t_3^k}))$ for every $t_3^k \in T_3$ and $y^{t_3^k} \to y^*$, it follows that

$$dist(y^*, B(C_1, R_0)) = d\,. \tag{4.120}$$

By the choice of T_3, $y^{t_3} \to y^*$, $B(A_{t_3}) \to B(C_1, R_0)$, and $B(A_{t_3} \cup \{y^{t_3}\}) \to B(C_2, R_0)$. Therefore, for $\epsilon_1 > 0$, there exists a large number $N_{\epsilon_1} \in \mathbb{N}$, such that for all $t_3 \in T_3$ with $t_3 \geq N_{\epsilon_1}$, we have

$$A_{t_3} \subseteq B(C_1, R_0 + \epsilon_1) \text{ and } A_{t_3} \cup \{y^{t_3}\} \cup y^* \subseteq B(C_2, R_0 + \epsilon_1). \tag{4.121}$$

Consequently (4.120), (4.121), and the triangle inequality imply

$$dist(C_1, C_2) \geq dist(y^*, C_1) - dist(y^*, C_2) \geq (d + R_0) - (R_0 + \epsilon_1) = d - \epsilon_1$$

and

$$A_{t_3} \subseteq B(C_1, R_0 + \epsilon_1) \cap B(C_2, R_0 + \epsilon_1), \quad t_3 \in T_3, t_3 \geq N_{\epsilon_1}.$$

Let $\bar{C} = \frac{C_1 + C_2}{2}$ and consider the hyperplane orthogonal to the vector $C_1 - C_2$ passing through $\bar{C}$. By Pythagoras' Theorem, $B(C_1, R_0 + \epsilon_1) \cap B(C_2, R_0 + \epsilon_1) \subseteq B(\bar{C}, \sqrt{(R_0 + \epsilon_1)^2 - (\frac{d - \epsilon_1}{2})^2})$ and therefore, $A_{t_3} \subseteq B(\bar{C}, \sqrt{(R_0 + \epsilon_1)^2 - (\frac{d - \epsilon_1}{2})^2})$,

which implies that $R(A_{t_3}) \leq \sqrt{(R_0+\epsilon_1)^2 - (\frac{d-\epsilon_1}{2})^2}$ for all $t_3 \in T_3$ with $t_3 \geq N_{\epsilon_1}$. By sending ϵ_1 to zero, we obtain that $R(A_{t_3}) \leq \sqrt{(R_0)^2 - (\frac{d}{2})^2}$ as $t_3 \in T_3 \rightarrow \infty$, but $R(A_{t_3}) \uparrow R_0$ implies $d = 0$, which completes the proof.

For the general case $\delta \geq 0$, we can use the same technique. Let $K' = (y^t)_{t\in\mathbb{N}}$ be a sequence of points selected as the furthest point in Step 14 (with the relative error size of δ) in one complete execution of the algorithm. Using the result and the language of the case $\delta = 0$ applied to the set K', we obtain that $B(K') = B(C_1, R_0)$, where C_1 and R_0 are the center and radius returned by the algorithm as $t \rightarrow \infty$, with the convention that whenever the algorithm stops we repeat the last set of points up to infinity.

Note that $\max_{x\in K} dist(C_1, x) \leq (1+\delta)R_0$, since otherwise the algorithm would have not converged to C_1, and therefore, $K \subseteq B(C_1, (1+\delta)R_0)$ which implies that $R(K) \leq (1+\delta)R_0$. Moreover, by $K' \subseteq K$ we have $R_0 = R(K') \leq R(K)$ and therefore $R_0 \leq R(K) \leq (1+\delta)R_0$, completing the proof.

Chapter 5

KERNEL METHODS ARE COMPETITIVE FOR OPERATOR LEARNING

We present a general kernel-based framework for learning operators between Banach spaces along with a priori error analysis and comprehensive numerical comparisons with popular neural net (NN) approaches such as Deep Operator Networks (DeepONet) (L. Lu, Jin, et al., 2021) and Fourier Neural Operator (FNO) (Li, Kovachki, Azizzadenesheli, B. Liu, et al., 2020). We consider the setting where the input/output spaces of target operator $\mathcal{G}^\dagger : \mathcal{U} \to \mathcal{V}$ are reproducing kernel Hilbert spaces (RKHS), the data comes in the form of partial observations $\phi(u_i), \varphi(v_i)$ of input/output functions $v_i = \mathcal{G}^\dagger(u_i)$ $(i = 1, \dots, N)$, and the measurement operators $\phi : \mathcal{U} \to \mathbb{R}^n$ and $\varphi : \mathcal{V} \to \mathbb{R}^m$ are linear. Writing $\psi : \mathbb{R}^n \to \mathcal{U}$ and $\chi : \mathbb{R}^m \to \mathcal{V}$ for the optimal recovery maps associated with ϕ and φ, we approximate $\mathcal{G}^\dagger$ with $\bar{\mathcal{G}} = \chi \circ \bar{f} \circ \phi$ where $\bar{f}$ is an optimal recovery approximation of $f^\dagger := \varphi \circ \mathcal{G}^\dagger \circ \psi : \mathbb{R}^n \to \mathbb{R}^m$. We show that, even when using vanilla kernels (e.g., linear or Matérn), our approach is competitive in terms of cost-accuracy trade-off and either matches or beats the performance of NN methods on a majority of benchmarks. Additionally, our framework offers several advantages inherited from kernel methods: simplicity, interpretability, convergence guarantees, a priori error estimates, and Bayesian uncertainty quantification. As such, it can serve as a natural benchmark for operator learning.

5.1 Introduction

Operator learning is a well-established field going back at least to the 1970s with the articles (Almroth, Stern, and Brogan, 1978; Noor and J. M. Peters, 1980) who introduced the reduced basis method as a way speeding up expensive model evaluations. In the most broad sense operator learning arises in the solution of stochastic PDEs (Ghanem and Spanos, 2003), emulation of computer codes (Kennedy and O'Hagan, 2001), reduced order modeling (ROM) (Lucia, Beran, and Silva, 2004), and numerical homogenization (Houman Owhadi and Clint Scovel, 2019a). In recent years, and with the rise of machine learning, operator learning has become the focus of extensive research with the development of neural net (NN) methods such as Deep Operator Nets (L. Lu, Jin, et al., 2021) and Fourier Neural Nets (Li, Kovachki,

Azizzadenesheli, B. Liu, et al., 2020) among many others. While these NN methods are often benchmarked against each other (L. Lu, X. Meng, et al., 2022), they are rarely compared with the aforementioned classical approaches. Furthermore, the theoretical analysis of NN methods is often limited to density/universal approximation results, showing the existence of a network of a requisite size achieving a certain error rate, without guarantees whether this network is computable in practice (see for example (Deng et al., 2022; Kovachki, Lanthaler, and Mishra, 2021)).

In order to alleviate the aforementioned shortcomings we present a mathematical framework for approximation of mappings between Banach spaces using the theory of operator valued reproducing Kernel Hilbert spaces (RKHS) and Gaussian Processes (GPs). Our abstract framework is (1) mathematically simple and interpretable, (2) convenient to implement, (3) encompasses some of the classical approaches such as linear methods, and (4) comes with a priori error analysis and convergence theory. We further present extensive benchmarking of our kernel method with the DeepONet and FNO approaches and show that the kernel approach either matches or outperforms NN methods in most benchmark examples.

In the remainder of this section we give a summary of our methodology and results: we pose the operator learning problem in Section 5.1 before presenting a running example in Section 5.1 which is used to outline our proposed framework and main theoretical results in Section 5.1 as well as brief numerical results in Section 5.1. Our main contributions are summarized in Section 5.1 followed by a literature review in Section 5.1.

The operator learning problem

Let $\mathcal{U}$ and $\mathcal{V}$ be two (possibly infinite-dimensional) separable Banach spaces and suppose that

$$\mathcal{G}^\dagger : \mathcal{U} \to \mathcal{V} \tag{5.1}$$

is an arbitrary (possibly nonlinear) operator. Then, broadly speaking, the goal of operator learning is to approximate $\mathcal{G}^\dagger$ from a finite number N of input/output data on $\mathcal{G}^\dagger$. For our framework, we consider the setting where the input/output data are only partially observed through a finite collection of linear measurements which we formalize as follows:

Problem 2. Let $\{u_i, v_i\}_{i=1}^N$ be N elements of $\mathcal{U} \times \mathcal{V}$ such that

$$\mathcal{G}^\dagger(u_i) = v_i, \quad \text{for } i = 1, \dots, N\,. \tag{5.2}$$

Let $\phi : \mathcal{U} \to \mathbb{R}^n$ and $\varphi : \mathcal{V} \to \mathbb{R}^m$ be bounded linear operators. Given the data $\{\phi(u_i), \varphi(v_i)\}_{i=1}^N$ approximate $\mathcal{G}^\dagger$.

Running example

To give context to the above problem and our solution method we briefly outline a running example to which the reader can refer to throughout the rest of this section. Consider the following elliptic PDE, which is of broad interest in geosciences and material science:

$$\begin{cases} -\operatorname{div} e^u \nabla v = w, & \text{in} \quad \Omega, \\ \qquad\qquad v = 0, & \text{on} \quad \partial\Omega\,, \end{cases} \tag{5.3}$$

where $\Omega = (0,1)^2$, $u \in H^3(\Omega)$, $w \in H^1(\Omega)$ and $v \in H^3(\Omega) \cap H_0^1(\Omega)$. For a fixed forcing term w, we wish to approximate the nonlinear operator mapping the diffusion coefficient u to the solution v, i.e., $\mathcal{G}^\dagger : u \mapsto v$. In this case we may take $\mathcal{U} \equiv H^3(\Omega)$ and $\mathcal{V} \equiv H^3(\Omega) \cap H_0^1(\Omega)$. We further assume that a training data set is available in the form of limited observations of input-out pairs. As a canonical example, consider the evaluation bounded and linear operators

$$\phi : u \mapsto (u(X_1), u(X_2), \ldots, u(X_n))^T \quad \text{and} \quad \varphi : v \mapsto (v(Y_1), v(Y_2), \ldots, v(Y_m))^T\,, \tag{5.4}$$

where the $\{X_j\}_{j=1}^n$ and $\{Y_j\}_{j=1}^m$ are distinct collocation points in the domain Ω as well as pairs $\{u_i, v_i\}_{i=1}^N$ that satisfy the PDE (5.3). Then our goal is to approximate $\mathcal{G}^\dagger$ from the training data set $\{\phi(u_i), \varphi(v_i)\}_{i=1}^N$ [1].

The proposed solution

Our setup naturally gives rise to a commutative diagram depicted in Figure 5.1. Here the map $f^\dagger : \mathbb{R}^n \to \mathbb{R}^m$ explicitely defined as

$$f^\dagger := \varphi \circ \mathcal{G}^\dagger \circ \psi \tag{5.5}$$

is a mapping between finite-dimensional Euclidean spaces, and is therefore amenable to numerical approximation. However, in order to approximate $\mathcal{G}^\dagger$ we also need the reconstruction maps $\psi : \mathbb{R}^n \to \mathcal{U}$ and $\chi : \mathbb{R}^m \to \mathcal{V}$.

Our proposed solution is to endow $\mathcal{U}$ and $\mathcal{V}$ with an RKHS structure and use kernel/GP regression to identify the maps ψ and χ. As a prototypical example we consider the situation where $\mathcal{U}$ is an RKHS of functions $u : \Omega \to \mathbb{R}$ defined by a

[1] Choosing ϕ, φ as pointwise evaluation functionals is common to many applications, although our abstract framework readily accommodates other choices such as integral operators and basis projections

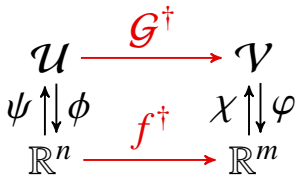


Figure 5.1: Commutative diagram of our operator learning setup.

kernel $Q : \Omega \times \Omega \to \mathbb{R}$ and $\mathcal{V}$ is an RKHS of functions $u : D \to \mathbb{R}$ defined by a kernel $K : D \times D \to \mathbb{R}$. For our running example, we have $D = \Omega$, and we can take Q and K to be Matérn like kernels, e.g., the Green's function of elliptic PDEs (possibly on Ω or restricted to Ω) with appropriate regularity. One can also choose Q, K to be smoother kernels such that their RKHSs are embedded in $\mathcal{U}$ and $\mathcal{V}$.

We then define ψ and χ as the following optimal recovery maps [2] :

$$\begin{aligned} \psi(U) &:= \operatorname{argmin}_{w \in \mathcal{U}} \|w\|_Q \quad \text{s.t.} \quad \phi(w) = U, \\ \chi(V) &:= \operatorname{argmin}_{w \in \mathcal{V}} \|w\|_K \quad \text{s.t.} \quad \varphi(w) = V, \end{aligned} \tag{5.6}$$

where $\|\cdot\|_Q$ and $\|\cdot\|_K$ are the RKHS norms arising from their pertinent kernels.

In the case where ϕ and φ are pointwise evaluation maps ($\phi(u) = (u(X_1), \ldots, u(X_n))$ and $\varphi(v) = (v(Y_1), \ldots, v(Y_m))$ where the X_i and Y_j are pairwise distinct collocation points in Ω and D), our optimal recovery maps can be expressed in closed form using standard representer theorems for kernel interpolation (Schölkopf, Herbrich, and Alex J. Smola, 2001):

$$\psi(U)(x) = Q(x, X)Q(X, X)^{-1}U, \qquad \chi(V)(y) = K(y, Y)K(Y, Y)^{-1}V, \tag{5.7}$$

where $Q(X, X)$ and $K(Y, Y)$ are kernel matrices with entries $Q(X, X)_{ij} = Q(X_i, X_j)$ and $K(Y, Y)_{ij} = K(Y_i, Y_j)$ respectively, while $Q(x, X)$ and $K(y, Y)$ denote row-vector fields with entries $Q(x, X)_i = Q(x, X_i)$ and $K(y, Y)_i = K(y, Y_i)$.

We further propose to approximate $f^\dagger$ by optimal recovery in a vector-valued RKHS. Let $\Gamma : \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^m)$ be a matrix valued kernel (Alvarez, Rosasco, Lawrence, et al., 2012); here $\mathcal{L}(\mathbb{R}^m)$ is the space of $m \times m$ matrices) with RKHS $\mathcal{H}_\Gamma$ equipped with the norm $\|\cdot\|_\Gamma$ [3] and proceed to approximate $f^\dagger$ by the map $\bar{f}$ defined as

$$\bar{f} := \operatorname{argmin}_{f \in \mathcal{H}_\Gamma} \|f\|_\Gamma \quad \text{s.t.} \quad f(\phi(u_i)) = \varphi(v_i) \quad \text{for} \quad i = 1, \ldots, N.$$

[2] It is possible to define the optimal recovery maps ψ, χ in the setting where ϕ and ψ are nonlinear, following the general framework of (Yifan Chen, Hosseini, et al., 2021a; Houman Owhadi, 2022; Houman Owhadi, 2023a). However, in this setting the closed form formulae (5.7) no longer hold.

[3] See Section 5.6 for a review of operator-valued kernels or the reference (Kadri et al., 2016).

A simple and practical choice for Γ is the diagonal kernel

$$\Gamma(U, U') = S(U, U')I, \tag{5.8}$$

where $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is an arbitrary scalar-valued kernel, such as RBF, Laplace, or Matérn, and I is the $m \times m$ identity matrix. More complicated choices, such as sums of kernels or replacing the identity matrix for a fixed positive definite matrix, implying correlations between various input or output correlations, are also possible. However, these may lead to greater computational cost and we observe empirically that the simple choice of the identity matrix already provides good performance. Then we can approximate the components of $\bar{f}$ via the independent optimal recovery problems

$$\bar{f}_j := \operatorname{argmin}_{g \in \mathcal{H}_S} \|g\|_S \quad \text{s.t.} \quad g(\phi(u_i)) = \varphi_j(v_i), \quad \text{for} \quad i = 1, \dots, N \tag{5.9}$$

for $j = 1, \dots, m$. Here we wrote $\varphi_j(v_i)$ for the entry j of the vector $\varphi(v_i)$ and, as our notation suggests, $\mathcal{H}_S$ is the RKHS of S equipped with the norm $\|\cdot\|_S$. Since (5.9) is a standard optimal recovery problem, each $\bar{f}_j$ can be identified by the usual representer formula:

$$\bar{f}_j(U) = S(U, \mathbf{U})S(\mathbf{U}, \mathbf{U})^{-1}\mathbf{V}_{\cdot,j}, \tag{5.10}$$

where $\mathbf{U} := (\phi(u_1), \dots, \phi(u_N))$ and $\mathbf{V}_{\cdot,j} := (\varphi_j(v_1), \dots, \varphi_j(v_N))^T$ and $S(U, \mathbf{U})$ is a block-vector and $S(\mathbf{U}, \mathbf{U})$ is a block-matrix defined in an analogous manner to those in (5.7). By combining equations (5.7) and (5.10) we obtain the operator

$$\bar{\mathcal{G}} := \chi \circ \bar{f} \circ \phi \tag{5.11}$$

as an approximation to $\mathcal{G}^\dagger$. We provide further details and generalize the proposed framework in Section 5.2 to the setting where ϕ and φ are obtained from arbitrary linear measurements (e.g., integral operators as in tomography) and $\mathcal{U}$ and $\mathcal{V}$ may not be spaces of continuous functions.

Convergence guarantee

Under suitable regularity assumptions on $\mathcal{G}^\dagger$, our method comes with worst-case convergence guarantees as the number of data points N , i.e., input-output pairs and the number of collocations points n and m go to infinity. We present here a condensed version of this result and defer the proof to Section 5.3. Below we write $B_R(\mathcal{H})$ for the ball of radius $R > 0$ in a normed space $\mathcal{H}$.

Theorem 5.1.1 (Condensed version of Thm. 5.3.4)**.** Suppose it holds that:

(5.1.1.1) (*Regularity of the domains* Ω *and* D) Ω and D are compact sets of finite dimensions d_Ω and d_D and with Lipschitz boundary.

(5.1.1.2) (*Regularity of the kernels* Q *and* K). Assume that $\mathcal{H}_Q \subset H^s(\Omega)$ and $\mathcal{H}_K \subset H^t(D)$ for some $s > d_\Omega/2$ and some $t > d_D/2$ with inclusions indicating continuous embeddings.

(5.1.1.3) (Space filling property of collocation points) The fill distance between the collocation points $\{X_i\}_{i=1}^n \subset \Omega$ and the $\{Y_j\}_{j=1}^m \subset D$ goes to zero as $n \to \infty$ and $m \to \infty$.

(5.1.1.4) (*Regularity of the operator* $\mathcal{G}^\dagger$) The operator $\mathcal{G}^\dagger$ is continuous from $H^{s'}(\Omega)$ to $\mathcal{H}_K$ for some $s' \in (0, s)$ as well as from $\mathcal{U}$ to $\mathcal{V}$ and all its Fréchet derivatives are bounded on $B_R(\mathcal{H}_Q)$ for any $R > 0$.

(5.1.1.5) *(Regularity of the kernels* S^n*)* Assume that for any $n \geq 1$ and any compact subset Υ of $\mathbb{R}^n$, the RKHS of S^n restricted to Υ is contained in $H^r(\Upsilon)$ for some $r > n/2$ and contains $H^{r'}(\Upsilon)$ for some $r' > 0$ that may depend on n.

(5.1.1.6) *(Resolution and space-filling property of the data)* Assume that for n sufficiently large, the data points $(u_i)_{i=1}^N \subset B_R(\mathcal{H}_Q)$ belong to the range of ψ^n and are space filling in the sense that they become dense in $\phi^n(B_R(\mathcal{H}_Q))$ as $N \to \infty$.

Then, for all $t' \in (0, t)$,

$$\lim_{n,m\to\infty} \lim_{N\to\infty} \sup_{u\in B_R(\mathcal{H}_Q)} \|\mathcal{G}^\dagger(u) - \chi^m \circ \bar{f}_N^{m,n} \circ \phi^n(u)\|_{H^{t'}(D)} \to 0\,, \tag{5.12}$$

where our notation makes the dependence of ψ, ϕ, χ, S, and $\bar{f}$ on n, m, and N explicit.

We note that Assumptions (5.1.1.1)–(5.1.1.3) are standard, and concern the accuracy of the optimal recovery maps ϕ^n and χ^m as $n, m \to \infty$. Assumptions (5.1.1.4)–(5.1.1.5) are less standard and amount to regularity assumptions on the map $\mathcal{G}^\dagger$ while Assumption (5.1.1.6) concerns the acquisition and regularity of the training data set.

In Section 5.3 we also present Theorem 5.3.3 as the quantitative analogue of the above result which characterizes how the speed of convergence depends on the regularity of the operator $\mathcal{G}^\dagger$ and the choice of ϕ and φ in the setting of pointwise

measurement operators. We also comment on how this analysis could be extended to other linear measurements.

Numerical Framework

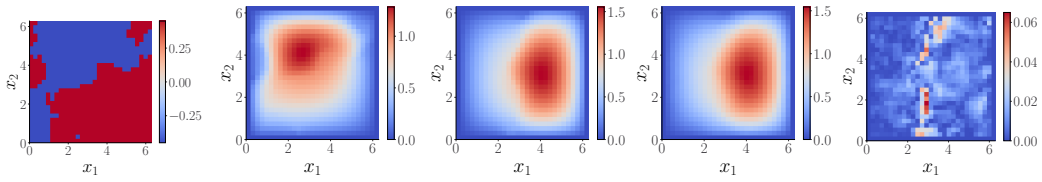
Returning to our running example, we implement the proposed framework for learning the non-linear operator mapping u to v in (5.3). We consider 1,000 inputs and outputs of u and v. The data is taken from (L. Lu, X. Meng, et al., 2022) and the experimental setup is discussed further in Remark 8. We take φ to be of the form (5.4) with $m = 841$ while we define ϕ through a PCA pre-processing step. More precisely, let $\phi_{\text{pointwise}}$ be of the form (5.4) with $n = 841$. Choose $n_{\text{PCA}} = 202$ (this value captures 95% of the empirical variance of our training data) and define

$$\phi(u) = \Pi_{\text{PCA}} \circ \phi_{\text{pointwise}}(u) \in \mathbb{R}^{202}. \tag{5.13}$$

In other words, we take our ϕ map to be the linear map that computes the first 202 PCA coefficients of the input functions u given on a uniform grid; observe that we do not use PCA pre-processing on the output data here, although we do this for some of our other examples in Section 5.4 for better performance.

With ϕ and φ identified (recall Figure 5.1) we proceed to implement our kernel method using the simple choice of a diagonal kernel $S(U, U')I$ where S is a rational quadratic (RQ) kernel (see Section 5.8). This choice transforms the problem into 841 independent kernel regression problems, each corresponding to one component of $f^\dagger$ (i.e., the $f_j^\dagger$'s).

We used the PCA and kernel regression modules of the `scikit-learn` Python library (Pedregosa et al., 2011) to implement our algorithm. This implementation automatically selects the best kernel parameters by maximizing the marginal likelihood function (Rasmussen and Williams, 2006) jointly for all problems. Our proposed method can therefore be implemented conveniently using off-the-shelf software. Figure 5.2 illustrates examples of the inputs and outputs of our operator learning problem. Despite the simple implementation of our method, we are able to obtain competitive accuracy as shown in Table 5.1 where the relative testing L^2 loss of our method is compared to other popular algorithms. Moreover, our approach is amenable to well-known numerical analysis techniques, such as sparse or low-rank approximation of kernel matrices, to reduce its complexity. For the present example (and those in Section 5.4) we only consider "vanilla" kernel methods which compute (5.10) by computing the full Cholesky factors of the matrix $S(\mathbf{U}, \mathbf{U})$.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

Figure 5.2: Example of training data and test prediction and pointwise errors for the Darcy flow problem (5.3).

Method	Accuracy
DeepONet	2.91 %
FNO	2.41 %
POD-DeepONet	2.32 %
Linear	6.74 %
Rational quadratic	2.87%

Table 5.1: The L^2 relative test error of the Darcy flow problem in our running example. The kernel approach is compared with variations of DeepONet and FNO. Results of our kernel method are presented below the dashed line with the pertinent choice of the kernel S.

Summary of contributions

The main results of the article concern the properties, performance, and error analysis of the map $\bar{\mathcal{G}}$ defined in (5.11). Our contributions can be summarized under four categories:

1. **An abstract kernel framework for operator learning:** In Section 5.2, we propose a framework for operator learning using kernel methods with several desirable properties. A family of methods of increasing complexity is proposed that includes linear models and diagonal kernels as well as non-diagonal kernels which capture output correlations. These properties make our approach ideal for benchmarking purposes. Furthermore, the methodology is (i) applicable to any choice of the linear functionals φ and ϕ, (ii) minimax optimal with respect to an implicitly defined operator-valued kernel, and (iii) is mesh-invariant. We emphasize in remark 7 that our optimal recovery maps can be applied to *any* operator learning after training to obtain a mesh-invariant pipeline.

2. **Error analysis and convergence rates for $\bar{\mathcal{G}}$:** In Section 5.3, we develop rigorous worst-case a priori error bounds and convergence guarantees for

our method: Theorem 5.3.3 provides quantitative error bounds while Theorem 5.3.4 (the detailed version of Theorem 5.3.3) shows the convergence of $\bar{\mathcal{G}} \to \mathcal{G}$ under appropriate conditions.

3. **A simple to use vanilla kernel method:** While our abstract kernel method is quite general, our numerical implementation in Section 5.4 focuses on a simple, easy-to-implement version using diagonal kernels of the form (5.8). Off-the-shelf software, such as the kernel regression modules of `scikit-learn`, can be employed for this task. We empirically observe low training times and robust choice of hyperparameters. These properties further suggest that kernel methods are a good baseline for benchmarking of more complex methods.

4. **Competitive performance.** In Section 5.4 we present a series of numerical experiments on benchmark PDE problems from the literature and observe that our simple implementation of the kernel approach is competitive in terms of complexity-accuracy tradeoffs in comparison to several NN-based methods. Since kernel methods can be interpreted as an infinite-width, one-layer NN, the results raise the question of how much of a role the depth of a deep NN plays in the performance of algorithms for the purposes of operator learning.

Review of relevant literature

In the most broad sense, operator learning is the problem of approximating a mapping between two infinite-dimensional function spaces (Bhattacharya et al., 2021; De Hoop et al., 2022). In recent years, this problem has become an area of intense research in the scientific machine learning community with a particular focus on parametric or stochastic PDEs. However, the approximation of such a parameter to solution maps has been an area of intense research in the computational mathematics and engineering communities, going back at least to the reduced basis method introduced in the 1970s (Almroth, Stern, and Brogan, 1978; Noor and J. M. Peters, 1980) as a way of speeding up the solution of families of parametric PDEs in applications that require many PDE solves such as design (Economon et al., 2016; Martins and Lambe, 2013; Bendsoe and Sigmund, 2003; Boncoraglio and Farhat, 2021), uncertainty quantification (UQ) (Sudret, Marelli, and Wiart, 2017; Martin et al., 2012; Huang, Schneider, and Andrew M Stuart, 2022), and multi-scale modeling (Weinan, 2011; Fish et al., 1997; Feyel and Chaboche, 2000; Kovachki, B. Liu, et al., 2022). In what follows we give a brief summary of the various areas and methodologies that overlap with operator learning; we cannot provide an exhaustive

list of references due to space, but refer the reader to key contributions and surveys where further references can be found.

Deep learning techniques The use of NNs for operator learning goes back at least to the 90s and the seminal works of Chen and Chen (T. Chen and H. Chen, 1995b; T. Chen and H. Chen, 1995a) who proved a universal approximation theorem for NN approximations to operators. The use and design of NNs for operator learning has become popular in the last five years as a consequence of growing interest in NNs for scientific computing starting with the article (Zhu and Zabaras, 2018) which used autoencoders to build surrogates for UQ of subsurface flow models. Since then many different approaches have been proposed, some of which use specific architectures or target particular families of PDEs (J. Hesthaven and Ubbiali, 2018; Khoo and Ying, 2019; Li, Kovachki, Azizzadenesheli, B. Liu, et al., 2020; Khoo, J. Lu, and Ying, 2021; L. Lu, Jin, et al., 2021; Gin et al., 2021; Boullé and Townsend, 2022; Kröpfl, Maier, and Peterseim, 2022; Kissas et al., 2022). The most relevant of among these methods to our proposed framework are the DeepONet family (L. Lu, Jin, et al., 2021; S. Wang, H. Wang, and Perdikaris, 2021; L. Lu, X. Meng, et al., 2022; S. Wang, H. Wang, and Perdikaris, 2022), FNO (Li, Kovachki, Azizzadenesheli, B. Liu, et al., 2020), and PCA-Net (J. Hesthaven and Ubbiali, 2018; Bhattacharya et al., 2021) where the main novelty appears to be the use of novel, flexible, and expressive NN architectures that allow the algorithm to learn and adapt the bases that are selected for the input and outputs of the solution map as well as possible nonlinear dependencies between the basis coefficients. Although not part of our comparisons, we note that (Fan, Bohorquez, and Ying, 2019; Fan, Feliu-Faba, et al., 2019; Fan, Lin, et al., 2019) obtained competitive accuracy by using deep neural networks with architectures inspired by conventional fast solvers.

Classical numerical approximation methods Operator learning has been the subject of intense research in the computational mathematics literature in the context of stochastic Galerkin methods (Ghanem and Spanos, 2003; Xiu and J. Shen, 2009), polynomial chaos (Xiu, 2010; Xiu and George Em Karniadakis, 2002), reduced basis methods (Noor and J. M. Peters, 1980; Maday, Patera, and Turinici, 2002) and numerical homogenization (Houman Owhadi and L. Zhang, 2007; Houman Owhadi, 2015a; Houman Owhadi and Clint Scovel, 2019a; Altmann, Henning, and Peterseim, 2021). In the setting of stochastic and parametric PDEs, the the goal is often to approximate the solution of a PDE as a function of a random or

uncertain parameter. The well-established approach to such problems is to pick or construct appropriate bases for the input parameter and the solution of the PDE and then construct a parametric, high-dimensional map that transforms the input basis coefficients to the output coefficients. Well-established methods such as polynomial chaos, stochastic finite element methods, and reduced basis methods (Ghanem and Spanos, 2003; Xiu, 2010; Cohen and DeVore, 2015; J. S. Hesthaven, Rozza, Stamm, et al., 2016; Lucia, Beran, and Silva, 2004) fall within this category. A vast amount of literature in applied mathematics exists on this subject, and the theoretical analysis of these methods is extensive; see for example (Beck et al., 2012; Chkifa, Cohen, DeVore, et al., 2012; Chkifa, Cohen, and Schwab, 2014; Nobile, Raúl Tempone, and Webster, 2008; Nobile, Raul Tempone, and Webster, 2008; Gunzburger, Webster, and G. Zhang, 2014) and references therein.

Operator compression For solving PDEs, the objectives of operator learning are also similar to those of operator compression (Feischl and Peterseim, 2020; Kröpfl, Maier, and Peterseim, 2022) as formulated in numerical homogenization (Houman Owhadi and Clint Scovel, 2019a; Altmann, Henning, and Peterseim, 2021) and reduced order modeling (ROM) (Amsallem and Farhat, 2008; Lucia, Beran, and Silva, 2004), i.e., the approximation of the solution operator from pairs of solutions and source/forcing terms. While both ROM and numerical homogenization seek operator compression through the identification of reduced basis functions that are as accurate as possible (this translates into low-rank approximations with SVD and its variants (Boullé and Townsend, 2022)), numerical homogenization also requires those functions to be as localized as possible (Målqvist and Peterseim, 2014) and in turn leverages both low rank and sparse approximations. These localized reduced basis functions are known as Wannier functions in the physics literature (Marzari et al., 2012), and can be interpreted as linear combinations of eigenfunctions that are localized in both frequency space and the physical domain, akin to wavelets. The hierarchical generalization of numerical homogenization (Houman Owhadi, 2017) (gamblets) has led to the current state-of-the-art for operator compression of linear elliptic (Schaäfer, Katzfuss, and Houman Owhadi, 2021; Florian Schäfer, Timothy John Sullivan, and Houman Owhadi, 2021a) and parabolic/hyperbolic PDEs (Houman Owhadi and L. Zhang, 2017). In particular, for arbitrary (and possibly unknown) elliptic PDEs (Florian Schäfer and Houman Owhadi, 2021) shows that the solution operator (i.e., the Green's function) can be approximated in near-linear complexity to accuracy ϵ from only $\tilde{O}(\log^{d+1}(\frac{1}{\epsilon}))$ solutions of the PDE.

GP emulators In the case where the range of the operator of interest is finite dimensional, then operator learning coincides with surrogate modeling techniques that were developed in the UQ literature, such as GP surrogate modeling/emulation (Kennedy and O'Hagan, 2001; Bastos and O'hagan, 2009). When the kernels of the underlying GPs are also learned from data (Houman Owhadi and Yoo, 2019a; Yifan Chen, Houman Owhadi, and A. Stuart, 2021b), GP surrogate modeling has been shown to offer a simple, low-cost, and accurate solution to learning dynamical systems (Hamzi and Houman Owhadi, 2021), geophysical forecasting (Hamzi, Maulik, and Houman Owhadi, 2021a), and radiative transfer emulation (Susiluoto et al., 2021), and the inference of the structure of convective storms from passive microwave observations (Prasanth et al., 2021). Indeed, our proposed kernel framework for operator learning can be interpreted as an extension of these well-established GP surrogates to the setting where the range of the operator is a function space.

Outline of the article

The remainder of the article is organized as follows: we present our operator learning framework in Section 5.2 for the generalized setting where ϕ, φ can be any collection of bounded and linear operators along with an interpretation of our method from the GP perspective. Our convergence analysis and quantitative error bounds are presented in Section 5.3 where we present the full version of Theorem 5.3.4. Our numerical experiments, implementation details, and benchmarks against FNO and DeepONet are collected in Section 5.4. We discuss future directions and open problems in Section 5.5. The appendix collects a review of operator valued kernels and GPs along with other auxiliary details.

5.2 The RKHS/GP framework for operator learning

We now present our general kernel framework for operator learning, i.e., the proposed solution to Problem 2. We emphasize that here we do not require the spaces $\mathcal{U}$ and $\mathcal{V}$ to be spaces of continuous functions and in particular, we do not require the maps ϕ and φ to be obtained from pointwise measurements. To describe this, we will introduce the dual spaces of $\mathcal{U}$ and $\mathcal{V}$ to define optimal recovery with respect to kernel operators rather than just kernel functions.

Write $\mathcal{U}^*$ and $\mathcal{V}^*$ for the duals of $\mathcal{U}$ and $\mathcal{V}$, and write $[\cdot, \cdot]$ for the pertinent duality pairings. Assume that $\mathcal{U}$ is endowed with a quadratic norm $\|\cdot\|_Q$, i.e., there exists a linear bijection $Q : \mathcal{U}^* \to \mathcal{U}$ that is symmetric ($[\phi_a, Q\phi_b] = [\phi_b, Q\phi_a]$), positive ($[\phi_a, Q\phi_a] > 0$ for $\phi_a \neq 0$), and such that $\|u\|_Q^2 = [Q^{-1}u, u]$, $\forall u \in \mathcal{U}$.

As in (Houman Owhadi and Clint Scovel, 2019a, Ch. 11), although $\mathcal{U}$ and $\mathcal{U}^*$ are also Hilbert spaces under $\|\cdot\|_Q$ and its dual norm $\|\cdot\|_Q^*$ (with inner products $\langle u, v\rangle_Q = [Q^{-1}u, v]$ and $\langle \phi_a, \phi_b\rangle_Q^* = [\phi_a, Q\phi_b]$), we will keep using the Banach space terminology to emphasize the fact that our dual pairings will not be based on the inner product through the Riesz representation theorem, but on a different realization of the dual space, as this setting is more practical.

If $\mathcal{U}$ is a space of continuous functions on a subset $\Omega \subset \mathbb{R}^{d_\Omega}$ then $\mathcal{U}^*$ contains delta Dirac functions and, to simplify notations, we also write $Q(x, y) := [\delta_x, Q\delta_y]$ for $x, y \in \mathbb{R}^{d_\Omega}$ to denote the kernel induced by the operator Q. Note that in that case, $\mathcal{U}$ is a RKHS with norm $\|\cdot\|_Q$ induced by the kernel Q. Since ϕ is bounded and linear, its entries ϕ_i (write $\phi := (\phi_1, \ldots, \phi_n)$) must be elements of $\mathcal{U}^*$. We assume those elements to be linearly independent. Write $\psi : \mathbb{R}^n \to \mathcal{U}$ for the linear operator defined by

$$\psi(Y) := (Q\phi)\, Q(\phi, \phi)^{-1} Y \text{ for } Y \in \mathbb{R}^n, \tag{5.14}$$

where we write $Q(\phi, \phi)$ for the $n \times n$ symmetric positive definite (SPD) matrix with entries $Q(\phi_i, \phi_j) := [\phi_i, Q\phi_j]$ [4] and $Q\phi$ for $(Q\phi_1, \ldots, Q\phi_n) \in \mathcal{U}^n$. As described in (Houman Owhadi and Clint Scovel, 2019a, Chap. 11), for $u \in \mathcal{U}$, given $\phi(u) = Y$, $\psi(Y)$ is the minmax optimal recovery of u when using the relative error in $\|\cdot\|_Q$-norm as a loss.

Similarly, assume that $\mathcal{V}$ is endowed with a quadratic norm $\|\cdot\|_K$, defined by the symmetric positive linear bijection $K : \mathcal{V}^* \to \mathcal{V}$. Write $\varphi := (\varphi_1, \ldots, \varphi_m)$ and assume the entries of φ to be linearly independent elements of $\mathcal{V}^*$. Using the same notations as in (5.14) write $\chi : \mathbb{R}^m \to \mathcal{V}$ for the linear operator defined by

$$\chi(Z) := (K\varphi)\, K(\varphi, \varphi)^{-1} Z \text{ for } Z \in \mathbb{R}^m\,. \tag{5.15}$$

Then, as above, for $v \in \mathcal{V}$, given $\varphi(v) = Z$, $\chi(Z)$ is the minmax optimal recovery of v when using the relative error in $\|\cdot\|_K$-norm as a loss.

Write $\mathcal{L}(\mathbb{R}^m)$ for the space of bounded linear operators mapping $\mathbb{R}^m$ to itself , i.e., $m \times m$ matrices. Let $\Gamma : \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^m)$ be a matrix-valued kernel (Alvarez, Rosasco, Lawrence, et al., 2012) defining an RKHS $\mathcal{H}_\Gamma$ of continuous functions $f : \mathbb{R}^n \to \mathbb{R}^m$ equipped with an RKHS norm $\|\cdot\|_\Gamma$. For $i \in \{1, \ldots, N\}$, write $U_i := \phi(u_i)$ and $V_i := \varphi(v_i)$. Write $\mathbf{U}$ and $\mathbf{V}$ for the block-vectors with entries U_i and

[4] For linear measurements involving derivatives the computation of these kernel matrices requires the computation of derivatives of the kernels; see (Yifan Chen, Hosseini, et al., 2021a) for practical examples and considerations.

V_i. Write $\Gamma(\mathbf{U}, \mathbf{U})$ for the $N \times N$ block-matrix with entries $\Gamma(U_i, U_j)$ and assume $\Gamma(U, U)$ to be invertible (which is satisfied if Γ is non-degenerate and $U_i \neq U_j$ for $i \neq j$). Let $f^\dagger$ be an element of $\mathcal{H}_\Gamma$ and write $f^\dagger(\mathbf{U})$ for the block vector with entries $f^\dagger(U_i)$. Then given $f^\dagger(\mathbf{U}) = \mathbf{V}$ it follows that

$$\bar{f}(U) := \Gamma(U, \mathbf{U})\Gamma(\mathbf{U}, \mathbf{U})^{-1}\mathbf{V}\,, \tag{5.16}$$

is the minimax optimal recovery of $f^\dagger$, where $\Gamma(\cdot, \mathbf{U})$ is the block-vector with entries $\Gamma(\cdot, U_i)$.

To this end, we propose to approximate the ground truth operator $\mathcal{G}^\dagger$ with

$$\bar{\mathcal{G}} := \chi \circ \bar{f} \circ \phi\,, \tag{5.17}$$

also recall Figure 5.1. Combining (5.15) and (5.16) we further infer that $\bar{\mathcal{G}}$ admits the following explicit representer formula

$$\bar{\mathcal{G}}(u) = (K\varphi)\, K(\varphi, \varphi)^{-1}\Gamma(\phi(u), \mathbf{U})\Gamma(\mathbf{U}, \mathbf{U})^{-1}\mathbf{V}. \tag{5.18}$$

In the remainder of this section we will provide more details and observations regarding our approximate operator $\bar{\mathcal{G}}$ that is useful later in Section 5.3 and of independent interest.

The kernel and RKHS associated with $\bar{\mathcal{G}}$

The explicit formula (5.18) suggests that the operator $\bar{\mathcal{G}}$ is an element of an RKHS defined by an operator-valued kernel, which we now characterize. For $u_1, u_2 \in \mathcal{U}$ and $v \in \mathcal{V}$ write

$$G(u_1, u_2)v := (K\varphi)\,(K(\varphi, \varphi))^{-1}\Gamma(\phi(u_1), \phi(u_2))(K(\varphi, \varphi))^{-1}\varphi(v). \tag{5.19}$$

It turns out that $G : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V})$ is a well-defined operator-valued kernel whose RKHS contains operators of the form $\bar{\mathcal{G}}$.

Proposition 5.2.1. The kernel G in (5.19) is an operator-valued kernel. Write $\mathcal{H}_G$ for its RKHS and $\|\cdot\|_G$ for the associated norm. Then it holds that $\mathcal{G} \in \mathcal{H}_G$ if and only if $\mathcal{G} = \chi \circ f \circ \phi$ for $f = \varphi \circ \mathcal{G} \circ \psi \in \mathcal{H}_\Gamma$ and $\|\mathcal{G}\|_G = \|f\|_\Gamma$.

Proof. Since G is Hermitian and positive, we deduce that G is an operator-valued kernel. Indeed for $\widetilde{u}_1, \ldots, \widetilde{u}_m \in \mathcal{U}$ and $\widetilde{v}_1, \ldots, \widetilde{v}_m \in \mathcal{V}$, using $\left\langle \widetilde{v}_i, K\varphi_s \right\rangle_K = \varphi_s(\widetilde{v}_i)$ and the fact that Γ is a matrix-valued kernel we have

$$\begin{aligned}\left\langle \widetilde{v}_i, G(\widetilde{u}_i, \widetilde{u}_j)\widetilde{v}_j \right\rangle_K &= \varphi(\widetilde{v}_i)^T (K\varphi)\,(K(\varphi, \varphi))^{-1}\Gamma(\phi(\widetilde{u}_i), \phi(\widetilde{u}_j))(K(\varphi, \varphi))^{-1}\varphi(\widetilde{v}_j) \\ &= \left\langle G(\widetilde{u}_j, \widetilde{u}_i)\widetilde{v}_i, \widetilde{v}_j \right\rangle_K,\end{aligned} \tag{5.20}$$

where we used $\left\langle \widetilde{v}_i, K\varphi_s \right\rangle_K = \varphi_s(\widetilde{v}_i)$ and the fact that Γ is a matrix-valued kernel. Furthermore, summing (5.20), we deduce that $\sum_{i,j=1}^m \left\langle \widetilde{v}_i, G(\widetilde{u}_i, \widetilde{u}_j)\widetilde{v}_j \right\rangle_K \geq 0$. From (5.19) we infer

$$\sum_{j=1}^m G(u, \widetilde{u}_j)\widetilde{v}_j = \chi \circ f \circ \phi(u) \tag{5.21}$$

with the function

$$f(U) = \sum_{j=1}^m \Gamma(U, \phi(\widetilde{u}_j))(K(\varphi, \varphi))^{-1}\varphi(\widetilde{v}_j)\,. \tag{5.22}$$

Furthermore using the reproducing property of G and (5.20) we have $\left\| \sum_{j=1}^m G(u, \widetilde{u}_j)\widetilde{v}_j \right\|_G^2 = \|f\|_\Gamma^2$. Therefore the closure of the space of operators of the form (5.21) with respect to the RKHS norm induced by G is the space of functions of the form $\chi \circ f \circ \phi$ where f lives in the closure of functions of the form (5.22) with respect to the RKHS norm induced by Γ. We deduce that $\mathcal{H}_G = \{\chi \circ f \circ \phi \mid f \in \mathcal{H}_\Gamma\}$. The uniqueness of f in the representation $\mathcal{G} = \chi \circ f \circ \phi$ for $f \in \mathcal{H}_G$ follows from $f = \varphi \circ \mathcal{G} \circ \psi$ following the identities $\varphi \circ \chi = I_d$ and $\phi \circ \psi = I_d$. □

Using the above result we can further characterize $\bar{\mathcal{G}}$ and $\bar{f}$ via optimal recovery problems in $\mathcal{H}_G$ and $\mathcal{H}_\Gamma$ respectively. In what follows we will write $\mathbf{u}$ for the N vector whose entries are the u_i, and $\mathcal{G}(\mathbf{u})$ for the N vector whose entries are $\mathcal{G}^\dagger(u_i)$.

Proposition 5.2.2. The operator $\bar{\mathcal{G}}$ is the minimizer of

$$\begin{cases} \text{Minimize} & \|\mathcal{G}\|_G^2 \\ \text{Over} & \mathcal{G} \in \mathcal{H}_G \text{ such that } \varphi \circ \mathcal{G}(\mathbf{u}) = \varphi \circ \mathcal{G}^\dagger(\mathbf{u}) \end{cases} \tag{5.23}$$

while the map $\bar{f}$ is the minimizer of

$$\begin{cases} \text{Minimize} & \|f\|_\Gamma^2 \\ \text{Over} & f \in \mathcal{H}_\Gamma \text{ such that } f \circ \phi(\mathbf{u}) = \varphi \circ \mathcal{G}^\dagger(\mathbf{u})\,. \end{cases} \tag{5.24}$$

Proof. By Proposition 5.2.1 $\bar{G}$ is completely identified by $\bar{f}$ and $\|\bar{G}\|_G = \|\bar{f}\|_\Gamma$. Then solving (5.23) is equivalent to solving (5.24). The statement regarding $\bar{f}$ follows directly from representer formulae for optimal recovery with matrix-valued kernels. □

Regularizing $\bar{G}$ by operator regression

As is often the case with optimal recovery/kernel regression the estimator for $\bar{f}$ in (5.16) is susceptible to numerical error due to ill-conditioning of the kernel matrix $\Gamma(\mathbf{U}, \mathbf{U})$. To overcome this issue we regularize our estimator by adding a small diagonal perturbation to this matrix. More precisely, let $\gamma > 0$ and write I for the identity matrix. We then define the regularized map

$$\bar{f}_\gamma(U) := \Gamma(U, \mathbf{U})\big(\Gamma(\mathbf{U}, \mathbf{U}) + \gamma I\big)^{-1}\mathbf{V}\,. \tag{5.25}$$

This regularized map gives rise to the regularized approximate operator

$$\bar{\mathcal{G}}_\gamma := \chi \circ \bar{f}_\gamma \circ \phi\,,$$

which admits the following representer formula:

$$\bar{\mathcal{G}}_\gamma(u) = (K\varphi)\, K(\varphi, \varphi)^{-1}\Gamma(\phi(u), \mathbf{U})\big(\Gamma(\mathbf{U}, \mathbf{U}) + \gamma I\big)^{-1}\mathbf{V}\,. \tag{5.26}$$

We can further characterize this operator as the solution to an operator regression problem.

Proposition 5.2.3. $\bar{\mathcal{G}}_\gamma$ is the solution to

$$\text{Minimize}_{\,\mathcal{G}\in\mathcal{H}_G} \|\mathcal{G}\|_G^2 + \gamma^{-1}|\varphi \circ \mathcal{G}(\mathbf{u}) - \varphi \circ \mathcal{G}^\dagger(\mathbf{u})|^2. \tag{5.27}$$

Proof. By Proposition 5.2.1, $\mathcal{G} = \chi \circ f \circ \phi$ solves (5.27) if and if f solves

$$\text{Minimize}_{\,f\in\mathcal{H}_\Gamma} \|f\|_\Gamma^2 + \gamma^{-1}|f(\mathbf{U}) - \mathbf{V}|^2\,. \tag{5.28}$$

It then follows by standard representer theorems for matrix-valued kernel regression (see Section 5.6) that $\bar{f}_\gamma$ is the minimizer of (5.28). □

Interpretation as conditioned operator valued GPs

Our kernel approach to operator learning has a natural GP regression interpretation that is compatible with Bayesian inference and UQ pipelines. We present some facts and observations in this direction.

Write $\xi \sim \mathcal{N}(0, G)$ for the centered operator-valued GP with covariance kernel G [5] and $\zeta \sim \mathcal{N}(0, \Gamma)$ for a centered vector valued GP with covariance kernel Γ. Then it is straightforward to show that the law of ξ is equivalent to that of $\chi \circ \zeta \circ \phi$. Let

[5] See Section 5.6 for a review of operator valued GPs.

$Z = (Z_1, \ldots, Z_N)$ be a random block-vector, independent from ξ, with i.i.d. entries $Z_j \sim \mathcal{N}(0, \gamma I_m)$ for $j = 1, \ldots, N$; here $\gamma \geq 0$ and I_m is the $m \times m$ identity matrix.

Then ξ conditioned on $\varphi \circ \xi(\mathbf{u}) = \varphi(\mathbf{v}) + Z$ is an operator-valued GP with mean $\bar{\mathcal{G}}_\gamma$, as in (5.26), and conditional covariance kernel

$$G^\perp(u, u')v = (K\varphi)\,(K(\varphi,\varphi))^{-1}\Gamma(\phi(u_1), \phi(u_2))$$
$$\big(\Gamma(\phi(u), \phi(u')) - \Gamma(\phi(u), \mathbf{U})(\Gamma(\mathbf{U}, \mathbf{U}) + \gamma I)^{-1}\Gamma(\mathbf{U}, \phi(u'))\big)(K(\varphi,\varphi))^{-1}\varphi(v).$$

Furthermore, the law of ξ conditioned on $\varphi \circ \xi(\mathbf{u}) = \varphi(\mathbf{v}) + Z$ is equivalent to that of $\chi \circ \zeta^\perp \circ \phi$ where $\zeta^\perp \sim \mathcal{N}(\bar{f}_\gamma, \Gamma^\perp)$ is the GP ζ conditioned on $\zeta(\mathbf{U}) = \mathbf{V} + Z'$, whose mean is $\bar{f}_\gamma$ as in (5.25) and conditional covariance kernel is

$$\Gamma^\perp(U, U') = \Gamma(U, U') - \Gamma(U, \mathbf{U})(\Gamma(\mathbf{U}, \mathbf{U}) + \gamma I)^{-1}\Gamma(\mathbf{U}, U').$$

We also use the GP approach to derive an alternative regularization of (5.27) in Section 5.7.

Measurement and mesh invariance

As argued in (Li, Kovachki, Azizzadenesheli, B. Liu, et al., 2020), mesh invariance is a key property for operator learning methods, i.e, the learned operator should be generalizable at test time beyond the specific discretization that was used during training. In our framework, this translates to being able to predict the output of a test input function $\widetilde{u}$ given only a linear measurement $\widetilde{\phi}(\widetilde{u})$, where $\widetilde{\phi}$ was unknown at training time. For example $\widetilde{\phi}$ could be of the same form as ϕ (say (5.4)) but on a finer or coarser grid. Similarly, we may choose to output with an operator $\widetilde{\varphi}$ which is a coarse/fine version of φ. Our proposed framework can easily provide mesh invariance using additional optimal recovery and measurement operators at the input and outputs of the operator $\bar{\mathcal{G}}$ as depicted in Figure 5.3. In fact, we can not only accommodate modification of the grid but completely different measurement operators at testing time. For example, while ϕ, φ may be of the form (5.4) we may take $\widetilde{\phi}$ and $\widetilde{\varphi}$ to be integral operators such as Fourier or Radon transforms.

Let us describe our approach to mesh invariance in detail. Given bounded and linear operators $\widetilde{\phi} : \mathcal{U} \to \mathbb{R}^{\widetilde{n}}$ and $\widetilde{\varphi} : \mathcal{V} \to \mathbb{R}^{\widetilde{m}}$ we can approximate $\widetilde{\varphi}(\mathcal{G}^\dagger(\widetilde{u}))$ using the map $\bar{f}$ obtained from (5.16) defined in terms of our training. To achieve mesh invariance we simply need a consistent approach to interpolate/extend the testing measurment operators to those used for training and we achieve this using the optimal recovery map $\widetilde{\psi}$ that is defined from $\widetilde{\phi}$ analogously to ψ in (5.14).

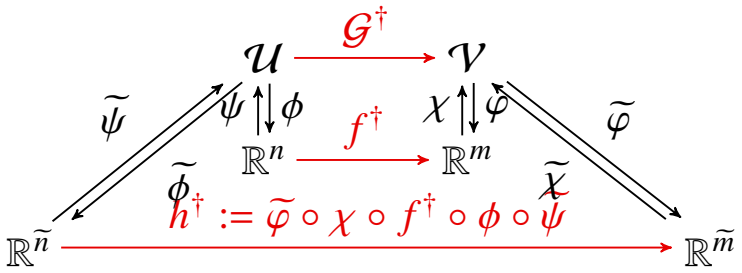


Figure 5.3: Generalization of fig. 5.1 to the mesh invariant setting where the measurement functionals are different at test time.

This setup gives rise to a natural approximation of $\mathcal{G}^\dagger$ in terms of the function $h^\dagger : \mathbb{R}^{\widetilde{n}} \to \mathbb{R}^{\widetilde{m}}$ depicted in Figure 5.3 which in turn can be approximated with $\bar{h} := \widetilde{\varphi} \circ \chi \circ \bar{f} \circ \phi \circ \widetilde{\psi} \equiv \widetilde{\varphi} \circ \bar{\mathcal{G}} \circ \widetilde{\psi}$. This expression further gives rise to another approximation to $\mathcal{G}^\dagger$ given by the operator $\widetilde{\mathcal{G}} = \widetilde{\chi} \circ \bar{h} \circ \widetilde{\phi}$.

Remark 7. Observe that the definition of $\bar{h}$ (and consequently $\bar{\mathcal{G}}$) is independent of the fact that $\bar{f}$ is constructed using the kernel approach. Thus, the optimal recovery maps χ and $\widetilde{\psi}$ can be used to retrofit any fixed-mesh operator learning algorithm, to become mesh-invariant and able to use arbitrary linear measurements of the function $\widetilde{u}$ at test time.

5.3 Convergence and error analysis

In this section, we present convergence guarantees and rigorous a priori error bounds for our proposed kernel method for operator learning and give a detailed statement and proof of Theorem 5.3.3. We assume that $\mathcal{H}_Q$ is a space of continuous functions from $\Omega \subset \mathbb{R}^{d_\Omega}$ and that $\mathcal{H}_K$ is a space of continuous functions from $D \subset \mathbb{R}^{d_D}$. Abusing notations we write $Q : \Omega \times \Omega \to \mathbb{R}^{d_\Omega}$ and $K : D \times D \to \mathbb{R}^{d_D}$ for the kernels induced by the operators Q and K. Let $X = (X_1, \ldots, X_n) \subset \Omega$ and $Y = (Y_1, \ldots, Y_m) \subset D$ be distinct collections of points and define their fill-distances

$$h_X := \max_{x' \in \Omega} \min_{x \in X} |x - x'|, \qquad h_Y := \max_{y' \in D} \min_{y \in Y} |y - y'|.$$

This section focuses on operators ϕ and φ that are linear combinations of pointwise measurements in X and Y. The presented results can be extended by using analogs of the sampling inequalities for other linear measurements; see (Houman Owhadi and Clint Scovel, 2019a, Theorem 4.11, Lemma 14.34) for a general framework that allows one to obtain such inequalities.

Let L_Q and L_K be invertible $n \times n$ and $m \times m$ matrices. For $u \in \mathcal{H}_Q$ write $u(X)$ for the n-vector with entries $u(X_i)$ and let $\phi : \mathcal{H}_Q \to \mathbb{R}^n$ be the bounded linear map defined by

$$\phi(u) = L_Q u(X) . \tag{5.29}$$

For $v \in \mathcal{H}_K$ write $v(Y)$ for the m-vector with entries $v(Y_j)$ and let $\varphi : \mathcal{H}_K \to \mathbb{R}^m$ be the bounded linear map defined by

$$\varphi(v) = L_K v(Y) . \tag{5.30}$$

Write $\|\phi\| := \sup_{u \in \mathcal{H}_Q} |\phi(u)|/\|u\|_Q$ and $\|\psi\| := \sup_{U' \in \mathbb{R}^n} \|\psi(U')\|_Q/|U'|$, and similarly $\|\varphi\| := \sup_{v \in \mathcal{H}_K} |\varphi(v)|/\|v\|_K$ and $\|\chi\| := \sup_{V' \in \mathbb{R}^m} \|\chi(V')\|_K/|V'|$. We will also assume the following regularity conditions on the domains Ω, D, the kernels Q, K, and the operator $\mathcal{G}^\dagger$.

Condition 5.3.1. Assume that the following conditions hold.

(5.3.1.1) Ω and D are compact sets with Lipschitz boundary.

(5.3.1.2) There exist indices $s > d_\Omega/2$ and $t > d_D/2$ so that $\mathcal{H}_Q \subset H^s(\Omega)$ and $\mathcal{H}_K \subset H^t(D)$, with inclusions indicating continuous embeddings.

(5.3.1.3) $\mathcal{G}^\dagger$ is a (possibly) nonlinear operator from $H^{s'}(\Omega)$ to $\mathcal{H}_K$ with $s' < s$ that satisfies

$$\|\mathcal{G}^\dagger(u) - \mathcal{G}^\dagger(v)\|_K \le \omega\left(\|u - v\|_{H^{s'}(\Omega)}\right) , \tag{5.31}$$

where $\omega : \mathbb{R} \to \mathbb{R}_+$ is the *modulus of continuity* of $\mathcal{G}^\dagger$.

Note that conditions (5.3.1.2) and (5.3.1.3) imply

$$\|\mathcal{G}^\dagger\|_{B_R(\mathcal{H}_Q) \to \mathcal{H}_K} := \sup_{u \in B_R(\mathcal{H}_Q)} \|\mathcal{G}^\dagger(u)\|_K < +\infty .$$

Proposition 5.3.1. *Suppose that Condition 5.3.1 holds. Let $0 < t' < t$. Then there exist constants $h_\Omega, h_D, C_\Omega, C_D > 0$ such that if $h_X < h_\Omega$ and $h_Y < h_D$, then*

$$\|\mathcal{G}^\dagger(u) - \chi \circ f^\dagger \circ \phi(u)\|_{H^{t'}(D)} \le C_D\, \omega\left(C_\Omega h_X^{s-s'} R\right) + C_D\, h_Y^{t-t'}\left(\|\mathcal{G}^\dagger(0)\|_K + \omega(C_\Omega R)\right),$$

for any $u \in B_R(\mathcal{H}_Q)$, where $f^\dagger$ is defined as in (5.5).

Proof. By the definition of $f^\dagger$ and the triangle inequality we have

$$\begin{aligned}\|\mathcal{G}^\dagger(u) - \chi\circ\varphi\circ\mathcal{G}^\dagger\circ\psi^\dagger\circ\phi(u)\|_{H^{t'}(\Gamma)} \leq &\|\mathcal{G}^\dagger(u) - \mathcal{G}^\dagger\circ\psi\circ\phi(u)\|_{H^{t'}(\Gamma)} \\ &+ \|\mathcal{G}^\dagger\circ\psi\circ\phi(u) - \chi\circ\varphi\circ\mathcal{G}^\dagger\circ\psi\circ\phi(u)\|_{H^{t'}(\Gamma)} \\ &=: T_1 + T_2.\end{aligned}$$

Let us first bound T_1: By conditions (5.3.1.2) and (5.3.1.3), we have

$$T_1 \leq C_D\|\mathcal{G}^\dagger(u) - \mathcal{G}^\dagger\circ\psi\circ\phi(u)\|_K \leq C_D\omega\left(\|u - \psi\circ\phi(u)\|_{H^{s'}(\Omega)}\right).$$

At the same time, since $(u-\psi\circ\phi(u))(X) = 0$, condition (5.3.1.1) and the sampling inequality for interpolation in Sobolev spaces (Arcangéli, López de Silanes, and Torrens, 2007, Thm. 4.1), and condition (5.3.1.2) imply that there exists a constant $h_\Omega > 0$ so that if $h_X < h_\Omega$ then

$$\|u - \psi\circ\phi(u)\|_{H^{s'}(\Omega)} \leq C'_\Omega h_X^{s-s'}\|u - \psi\circ\phi(u)\|_{H^s(\Omega)} \leq C_\Omega h_X^{s-s'}\|u - \psi\circ\phi(u)\|_Q\,, \tag{5.32}$$

where $C'_\Omega, C_\Omega > 0$ are constants that are independent of u. Using $\|u-\psi\circ\phi(u)\|_Q \leq \|u\|_Q$ (Houman Owhadi and Clint Scovel, 2019a, Thm. 12.3) we deduce the desired bound

$$T_1 \leq C_D\omega\left(C_\Omega h_\Omega^{s-s'}\|u\|_Q\right). \tag{5.33}$$

Let us now bound T_2: once again, by the continuous embedding of condition (5.3.1.2) and the sampling inequality for interpolation in Sobolev spaces, we have that there exists $h_D > 0$ so that if $h_Y < h_D$, then for any $v \in H^t(D)$ it holds that

$$\|v-\chi\circ\varphi(v)\|_{H^{t'}(D)} \leq C'_D h_Y^{t-t'}\|v-\chi\circ\varphi(v)\|_{H^t(D)} \leq C_D h_Y^{t-t'}\|v-\chi\circ\varphi(v)\|_K \leq C_D h_Y^{t-t'}\|v\|_K\,.$$

Taking $v \equiv \mathcal{G}^\dagger\circ\psi\circ\phi(u)$, we deduce that

$$\begin{aligned}T_2 &\leq C_D h_Y^{t-t'}\|\mathcal{G}^\dagger\circ\psi\circ\phi(u)\|_K, \\ &\leq C_D h_Y^{t-t'}\left(\|\mathcal{G}^\dagger(0)\|_K + \omega(\|\psi\circ\phi(u)\|_{H^{s'}(\Omega)})\right).\end{aligned}$$

Using $\|\psi\circ\phi(u)\|_{H^{s'}(\Omega)} \leq C_\Omega\|\psi\circ\phi(u)\|_Q \leq C_\Omega\|u\|_Q$ concludes the proof. □

While Proposition 5.3.1 gives an error bound for the distance between the maps $\mathcal{G}^\dagger$ and $\varphi\circ f^\dagger\circ\phi$, we can never compute this map when $N < \infty$ and so we have to approximate this map as well. Given the kernel Γ, our optimal recovery approximant for the map $f^\dagger$ is $\bar{f}$ as in (5.16), which we recall is the minimizer of (5.24).

To proceed, we need to consider another intermediary problem that defines an approximation $\widehat{f}$ to the map $f^\dagger$:

$$\widehat{f} := \begin{cases} \text{Minimize} & \|f\|_\Gamma^2 \\ \text{Over} & f \in \mathcal{H}_\Gamma \text{ such that } f \circ \phi(\mathbf{u}) = f^\dagger \circ \phi(\mathbf{u}) \,. \end{cases} \tag{5.34}$$

We emphasize that the difference between the problems (5.24) and (5.34) is simply in the training data that is injected in the equality constraints, and this difference is quite subtle:

In practical applications, observations may be taken from $\mathcal{G}^\dagger(u_i)$, which is different from $f^\dagger \circ \phi(u_i) \equiv \varphi \circ \mathcal{G}^\dagger \circ \psi \circ \phi(u_i)$. To make our analysis simple, henceforth we assume the following condition on our input data.

Condition 5.3.2. The input data points u_i satisfy

$$u_i = \psi \circ \phi(u_i) \text{ for } i = 1, \ldots, N \,.$$

We observe that this condition implies $\mathcal{G}^\dagger(u_i) = f^\dagger \circ \phi(u_i)$ and $\bar{f} = \widehat{f}$. Removing this assumption requires bounding some norm of the error $f^\dagger - \bar{f}$, and we postpone that analysis to a sequel paper as this step can become very technical.

The next step in our convergence analysis is then to control the error between the maps $\widehat{f}$ and $f^\dagger$ which we will achieve using similar arguments as in the proof of Proposition 5.3.1. For our analysis, we take Γ to be a diagonal, matrix-valued kernel, of the form (5.8) which we recall for reference

$$\Gamma(U, U') = S(U, U')I, \tag{5.35}$$

where I is the $m \times m$ identity matrix and $S : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a real valued kernel.

Proposition 5.3.2. *Suppose that Condition 5.3.2 holds. Let $\Upsilon \subset \mathbb{R}^n$ be a compact set with Lipschitz boundary and consider $U = (U_1, \ldots, U_N) \subset \Upsilon$ with fill distance*

$$h_\Upsilon := \max_{U' \in \Upsilon} \min_{1 \le i \le N} |U_i - U'|.$$

Let Γ be of the form (5.35)*, with S restricted to the set Υ, and suppose $\mathcal{H}_S \subset H^r(\Upsilon)$ for $r > n/2$ and that $f_j^\dagger \in \mathcal{H}_S$ for $j = 1, \ldots, m$. Then there exist constants $h'_\Upsilon, C_\Upsilon > 0$ so that whenever $h_\Upsilon < h'_\Upsilon$ then for any $r' < r$ it holds that*

$$\|f_j^\dagger - \widehat{f}_j\|_{H^{r'}(\Upsilon)} \le C_\Upsilon h_\Upsilon^{r-r'} \|f_j^\dagger\|_S.$$

Proof. The proof is a direct consequence of the fact that the components of $\widehat{f}$ are given by the optimal recovery problems (5.34) and the sampling inequality for interpolation in Sobolev spaces (Arcangéli, López de Silanes, and Torrens, 2007, Thm. 4.1) following the same arguments used in the proof of Theorem 5.3.1. □

We can now combine the above results to obtain the following theorem.

Theorem 5.3.3. *Suppose that Conditions 5.3.1 and 5.3.2 hold in addition to those of Proposition 5.3.2 with a set of inputs $(u_i)_{i=1}^N \subset B_R(\mathcal{H}_Q)$, the set $\Upsilon = \phi\left(B_R(\mathcal{H}_Q)\right)$, and index $n/2 < r' < r$. Then for any $u \in B_R(\mathcal{H}_Q)$, it holds that*

$$\begin{aligned}\|\mathcal{G}^\dagger(u) - \chi \circ \bar{f} \circ \phi(u)\|_{H^{t'}(D)} \leq & C_D\, \omega\left(C_\Omega h_X^{s-s'} R\right) + C_D\, h_Y^{t-t'}\left(\|\mathcal{G}^\dagger(0)\|_K + \omega(C_\Omega R)\right) \\ & + \sqrt{m} C_D C_\Upsilon \|\chi\| h_\Upsilon^{(r-r')} \max_{1\leq j\leq m} \|f_j^\dagger\|_S. \end{aligned} \tag{5.36}$$

Proof. An application of the triangle inequality yields

$$\begin{aligned}\|\mathcal{G}^\dagger(u) - \chi \circ \bar{f} \circ \phi(u)\|_{H^{t'}(D)} \leq & \|\mathcal{G}^\dagger(u) - \chi \circ f^\dagger \circ \phi(u)\|_{H^{t'}(D)} \\ & + \|\chi \circ f^\dagger \circ \phi(u) - \chi \circ \widehat{f} \circ \phi(u)\|_{H^{t'}(D)} \\ & + \|\chi \circ \widehat{f} \circ \phi(u) - \chi \circ \bar{f} \circ \phi(u)\|_{H^{t'}(D)} =: I_1 + I_2 + I_3. \end{aligned}$$

We can bound I_1 immediately using Proposition 5.3.1. Furthermore, by Condition 5.3.2 we have that $I_3 = 0$. So it remains for us to bound I_2: by the continuous embedding of $\mathcal{H}_K$ into $H^{t'}(D)$ we can write

$$\begin{aligned} I_2 \leq C_D \|\chi \circ f^\dagger \circ \phi(u) - \chi \circ \widehat{f} \circ \phi(u)\|_K &\leq C_D \|\chi\| |f^\dagger \circ \phi(u) - \widehat{f} \circ \phi(u)| \\ &\leq C_D \|\chi\| \sqrt{\sum_{j=1}^m \|f_j^\dagger - \widehat{f}_j\|_{H^{r'}(\Upsilon)}^2}, \end{aligned}$$

where the last line follows from the Sobolev embedding theorem and the assumption that $r' > n/2$. Then an application of Proposition 5.3.2 yields

$$I_2 \leq \sqrt{m} C_D C_\Upsilon \|\chi\| h_\Upsilon^{(r-r')} \max_{1\leq j\leq m} \|f_j^\dagger\|_S.$$

□

Convergence theorem

Our next step will be to consider the limits $N, n, m \to \infty$ and show the convergence of $\bar{\mathcal{G}}$ to $\mathcal{G}^\dagger$. To obtain this result we first need to make assumptions on the regularity of the true operator $\mathcal{G}^\dagger$.

For $k \geq 1$ write $D^k\mathcal{G}^\dagger$ for the functional derivative of $\mathcal{G}^\dagger$ of order k. Recall that for $u \in \mathcal{H}_Q$, $D^k\mathcal{G}^\dagger(u)$ is a multilinear operator mapping $\otimes_{i=1}^k \mathcal{H}_Q$ to $\mathcal{H}_K$. For $w_1, \ldots, w_k \in \mathcal{H}_Q$ write $[D^k\mathcal{G}^\dagger(u), \otimes_{i=1}^k w_i]$ for the (multilinear) action of $D^k\mathcal{G}^\dagger(u)$ on $\otimes_{i=1}^k w_i$ and write $\|D^k\mathcal{G}^\dagger(u)\|$ for the smallest constant such that for $w_1, \ldots, w_k \in \mathcal{H}_Q$,

$$\left\|[D^k\mathcal{G}^\dagger(u), \otimes_{i=1}^k w_i]\right\|_{\mathcal{H}_K} \leq \|D^k\mathcal{G}^\dagger(u)\| \prod_{i=1}^k \|w_i\|_{\mathcal{H}_Q}, \tag{5.37}$$

Similarly, for $k \geq 1$ write $D^k f^\dagger$ for the derivation tensor of $f^\dagger$ of order k (the gradient for $k = 1$ and the Hessian for $k = 2$, etc). Recall that for $U \in \mathbb{R}^n$, $D^k f^\dagger(U)$ is a multilinear operator mapping $\otimes_{i=1}^k \mathbb{R}^n$ to $\mathbb{R}^m$. For $W_1, \ldots, W_k \in \mathbb{R}^n$ write $[D^k f^\dagger(U), \otimes_{i=1}^k W_i]$ for the (multilinear) action of $D^k f^\dagger(U)$ on $\otimes_{i=1}^k W_i$ and write $\|D^k f^\dagger(U)\|$ for the smallest constant such that for $W_1, \ldots, W_k \in \mathbb{R}^n$,

$$\left|[D^k f^\dagger(U), \otimes_{i=1}^k W_i]\right| \leq \|D^k f^\dagger(U)\| \prod_{i=1}^k |W_i|. \tag{5.38}$$

where $|\cdot|$ is the Euclidean norm.

Lemma 5.3.3. It holds true that $\|D^k f^\dagger(U)\| \leq \|\varphi\|\|\psi\|^k \|D^k\mathcal{G}^\dagger \circ \psi(U)\|, \ \forall U \in \mathbb{R}^n$.

Proof. The chain rule and the linearity of φ and ψ imply that

$$[D^k f^\dagger(U), \otimes_{i=1}^k W_i] = \varphi[D^k\mathcal{G}^\dagger \circ \psi(U), \otimes_{i=1}^k \psi(W_i)] \,.$$

We then conclude the proof by writing

$$\begin{aligned}\left|[D^k f^\dagger(U), \otimes_{i=1}^k W_i]\right| &\leq \|\varphi\|\|D^k\mathcal{G}^\dagger \circ \psi(U)\| \prod_{i=1}^k \|\psi(W_i)\|_{\mathcal{H}_Q} \\ &\leq \|\varphi\|\|\psi\|^k \|D^k\mathcal{G}^\dagger \circ \psi(U)\| \prod_{i=1}^k |W_i| \,.\end{aligned}$$

□

Let us now consider an infinite and dense sequence of points $X_1, X_2, X_3, \ldots$ of Ω, such that the closure of $\cup_{i=1}^\infty \{X_i\}$ is the closure of Ω. Write X^n for the n-vector formed by the first n points, i.e.,

$$X^n := (X_1, \ldots, X_n) \tag{5.39}$$

and let L_Q^n be an arbitrary invertible $n \times n$ matrix. Further let $\phi^n : \mathcal{H}_Q \to \mathbb{R}^n$ be defined by

$$\phi^n(u) = L_Q^n u(X^n) . \tag{5.40}$$

Write ψ^n for the corresponding optimal recovery ψ-map. Similarly, we assume that we are given an infinite and dense sequence of points $Y_1, Y_2, Y_3, \ldots$ of D, such that the closure of $\cup_{i=1}^{\infty}\{Y_i\}$ is the closure of D. Write Y^m for the m-vector formed by the first m points, i.e.,

$$Y^m := (Y_1, \ldots, Y_m) . \tag{5.41}$$

Let L_K^m be an arbitrary invertible $m \times m$ matrix and let $\varphi^m : \mathcal{H}_K \to \mathbb{R}^m$ be defined by

$$\varphi^m(v) = L_K^m u(Y^m) . \tag{5.42}$$

Write χ^m for the corresponding optimal recovery χ-map. We also assume that we are given a sequence of diagonal matrix-valued kernels $\Gamma^{m,n} : \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{L}(\mathbb{R}^m)$ with scalar-valued kernels $S^n : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ as diagonal entries. Write $\bar{f}_N^{m,n}$ for the corresponding minimizer of (5.24) (also identified by the formula (5.16)) for the above setup.

Theorem 5.3.4. *Let m, n be the dimensionality of the input and output observations $\phi : \mathcal{U} \to \mathbb{R}^n$ and $\varphi : \mathcal{V} \to \mathbb{R}^m$. Suppose that the closure of $\lim_{n\uparrow\infty} \cup_{i=1}^n \{X_i\}$ is equal to the closure of Ω and that the closure of $\lim_{m\uparrow\infty} \cup_{i=1}^m \{Y_i\}$ is equal to the closure of D. Suppose Condition 5.3.1 is satisfied and that*

$$\sup_{u\in B_R(\mathcal{H}_Q)} \left\| [D^k \mathcal{G}^\dagger(u) \right\| < \infty \text{ for all } k \geq 1 , \tag{5.43}$$

for an arbitrary $R > 0$. Assume that for any $n \geq 1$ and any compact set Υ of $\mathbb{R}^n$, the RKHS of S^n restricted to Υ (which we write $\mathcal{H}_{S^n}(\Upsilon)$) is contained in $H^r(\Upsilon)$ for some $r > n/2$ and contains $H^{r'}(\Upsilon)$ for some $r' > 0$ that may depend on n. Let $(u_i)_{i=1}^N$ be a sequence of inputs in $B_R(\mathcal{H}_Q)$. Assume that there exists an integer n_0 such that for $n \geq n_0$, the data points $(u_i)_{i=1}^N$ satisfy Condition 5.3.2, i.e., they satisfy $u_i = \psi^n \circ \phi^n(u_i)$ for all $i \geq 1$. Further assume that the $(\phi^n(u_i))_{1\leq i\leq N}$ are space filling in the sense that for any $n \geq n_0$ we have

$$\lim_{N\to\infty} \sup_{u\in B_R(\mathcal{H}_Q)} \min_{1\leq i\leq N} \left| u_i(X^n) - u(X^n) \right| = 0 . \tag{5.44}$$

Then for any $t' \in (0, t)$, it holds that

$$\lim_{n,m\to\infty} \lim_{N\to\infty} \sup_{u\in B_R(\mathcal{H}_Q)} \| \mathcal{G}^\dagger(u) - \chi^m \circ \bar{f}_N^{m,n} \circ \phi^n(u) \|_{H^{t'}(D)} = 0. \tag{5.45}$$

Proof. Following (Houman Owhadi and Clint Scovel, 2019a, Chap. 12.1) define the projection $P_n^{\mathcal{U}} = \psi^n \circ \phi^n$ onto the range of ψ^n. Since the points X_i and Y_j are dense in Ω and D we have $h_{X^n} \downarrow 0$ as $n \to \infty$ and $h_{Y^m} \downarrow 0$ as $m \to \infty$. Given n, take $\Upsilon = \phi^n\left(B_R(\mathcal{H}_Q)\right)$. Then Lemma 5.3.3 and (5.43) imply that $\bar{f}_j^{m,n} \in H^{r'}(\Upsilon)$ for all $r' \geq 0$. Therefore $\bar{f}_j^{m,n} \in \mathcal{H}_{S^n}(\Upsilon)$. Now (5.44) implies that for any n, the fill distance, in $\phi^n\left(B_R(\mathcal{H}_Q)\right)$, between the points $(\phi^n(u_i))_{1\leq i\leq N}$ goes to zero as $N \to \infty$. Since the conditions of Proposition 5.3.2 are satisfied, we conclude by taking the limit $N \to \infty$ in (5.36) before taking the limit $m, n \to \infty$. □

The effect of the L_Q and L_K preconditioners

We conclude this section and our discussion of convergence results by highlighting the importance of the choice of the matrices L_Q^n and L_K^m in (5.40) and (5.42). It is clear from the bounds (5.36) and (5.38) that our error estimates depend on the norms of the linear operators φ^m, ψ^n and χ^m. To ensure that those norms do not blow up as $n, m \to \infty$ we can select the matrices L_Q^n and L_K^m to be the Cholesky factors of the precision matrices obtained from pointwise measurements of the kernels Q and K, i.e.,

$$L_Q^n (L_Q^n)^T = Q(X^n, X^n)^{-1} \quad \text{and} \quad L_K^m (L_K^m)^T = K(Y^m, Y^m)^{-1}. \tag{5.46}$$

We now obtain the following proposition.

Proposition 5.3.4. If ϕ^n is as in (5.40) and L_Q^n as in (5.46), then $\|\phi^n\| = 1$ and $\|\psi^n\| = 1$. If φ^m is as in (5.30) and L_K^m as in (5.46), then $\|\varphi^m\| = 1$ and $\|\chi^m\| = 1$.

Proof. For $u \in \mathcal{H}_Q$, $|\phi^n(u)|^2 = u(X^n)^T Q(X^n, X^n)^{-1} u(X^n) = \|\psi^n \circ \phi^n(u)\|_Q^2$. Since $\psi^n \circ \phi^n$ is a projection (Houman Owhadi and Clint Scovel, 2019a, Chap. 12.1) we deduce that $\|\phi^n\| = 1$. Using $\psi^n(U') = Q(\cdot, X^n) L_Q^n U'$ leads to $\|\psi^n(U')\|_Q^2 = |U'|^2$ and $\|\psi^n\| = 1$. The proof of $\|\varphi^n\| = 1$ and $\|\chi^n\| = 1$ is similar. □

We note that although useful for obtaining tighter approximation errors, this particular choice for the matrices L_Q^n and L_K^m is not required for convergence if one first takes the limit $N \to \infty$ as in Theorem 5.3.4, which does not put any requirements on the matrices L_Q^n and L_K^m beyond invertibility.

5.4 Numerics

In this section, we present numerical experiments and benchmarks that compare a straightforward implementation of our kernel operator learning framework to state-of-the-art NN-based techniques. We discuss some implementation details of our

method in Section 5.4 followed by the setup of experiments and test problems in Section 5.4 and remark 8. A detailed discussion of our findings is presented in Remark 8.

Implementation considerations

Below we summarize some of the key details in the implementation of our kernel approach for operator learning for benchmark examples. Our code to reproduce the experiments can be found in a public repository[6].

Choice of the kernel Γ

Following our theoretical discussions in Sections 5.2 and 5.3, we primarily take Γ to be a diagonal kernel of the form (5.35). This implies that our estimation of $\bar{f}$ can be split into independent problems for each of its components $\bar{f}_j$ in the RKHS of the scalar kernel S. In our experiments, we investigate different choices of S belonging to the families of the linear kernel, rational quadratic, and Matérn; see Section 5.8 for detailed expressions of these kernels. The rational quadratic kernel has two parameters: the lengthscale l and the exponent α. We tuned these parameters using standard cross validation or log marginal likelihood maximization over the training data (see (Rasmussen and Williams, 2006, p.112) for a detailed description). The Matérn kernel is parameterized by two positive parameters: a smoothness parameter ν and the length scale l. The smoothness parameter ν controls the regularity of the RKHS and we considered $\nu \in \left\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \infty\right\}$. In practice we found that $\nu = \frac{5}{2}$ almost always had the best performance. For a fixed choice of ν we tuned the length scale l similarly to the rational quadratic kernel. We implemented the kernel regressions of the $\bar{f}_j$ and parameter tuning algorithms in scikit-learn for low-dimensional examples and manually in JAX for high-dimensional examples.

Preconditioning and dimensionality reduction

Following (5.29) and (5.30) and the discussion in Section 5.3, we consider two preconditioning strategies for our pointwise measurements, i.e., choices of the matrices L_Q and L_K: (1) we consider the Cholesky factors of the underlying covariance matrices as in (5.46), and (2) we use PCA projection matrices of the input and output functions computed from the training data. We truncated the PCA expansions to preserve $(0.90, 0.95, 0.99)$ of the variance. The use of PCA in learning mappings

[6]https://github.com/MatthieuDarcy/KernelsOperatorLearning/

between infinite dimensional spaces was proposed in (Krischer et al., 1993) and recently revisited in (Bhattacharya et al., 2021; J. Hesthaven and Ubbiali, 2018).

Experimental setup

We compare the test performance of our method with different choices of the kernel S of increasing complexity using the examples in (De Hoop et al., 2022) and (L. Lu, X. Meng, et al., 2022) and their reported test relative L^2 loss (see eq. (5.48) below). We use the data provided by these papers for the training set and the test set[7]. Both articles provide performance comparisons between different variants of Neural Operators (most notably FNO and DeepONet) on a variety of PDE operator learning tasks, where the data is sampled independently from a distribution $(Id, \mathcal{G}^\dagger)^\# \mu$ supported on $\mathcal{U} \times \mathcal{V}$, where μ is a specified (input) distribution on $\mathcal{U}$. The example problems are outlined in detail in Remark 8; a summary of the specific PDEs, problem type, and distribution μ for each test is given in table 5.2. In some instances the train-test split of the data was not clear from the available online repositories in which case we re-sampled them from the assumed distribution μ. The datasets from (L. Lu, X. Meng, et al., 2022) contain 1,000 training data-points per problem (which we will refer to as the "low-data regime"), whereas the datasets from (De Hoop et al., 2022) contain 20000 training data-points (which we will refer to as the "high-data" regime). We make this distinction because the complexity of kernel methods, unlike that of neural networks, may depend on the number of data-points.

Following the suggestion of (De Hoop et al., 2022) we not only compare test errors and training complexity but also the complexity of operator learning at the inference/evaluation stage in Section 5.4. For the examples in (De Hoop et al., 2022), we investigate the accuracy-complexity trade-off of our method against the reported values of that article.

[7] See `https://github.com/Zhengyu-Huang/Operator-Learning` and `https://github.com/lu-group/deeponet-fno`, respectively, for the data

Equation	**Input**	**Output**	**Input Distribution μ**
Burger's	Initial condition	Solution at time T	Gaussian field (GF)
Darcy problem	Coefficient	Solution	Binary function of GF
Advection I	Initial condition	Solution at time T	Random square waves
Advection II	Initial condition	Solution at time T	Binary function of GF
Helmholtz	Coefficient	Solution	Function of Gaussian field
Structural mechanics	Initial force	Stress field	Gaussian field
Navier Stokes	Forcing term	Solution at time T	Gaussian field

Table 5.2: Summary of datasets used for benchmarking. The first three examples were considered in (L. Lu, X. Meng, et al., 2022), and the last four were taken from (De Hoop et al., 2022).

Measures of accuracy

As our first performance metric we measured the accuracy of models by a relative loss on the output space $\mathcal{V}$:

$$\mathcal{R}(\mathcal{G}) = \mathbb{E}_{u\sim\mu}\left[\frac{||\mathcal{G}^\dagger(u) - \mathcal{G}(u)||_{\mathcal{V}}}{||\mathcal{G}^\dagger(u)||_{\mathcal{V}}}\right], \tag{5.47}$$

where $\mathcal{G}^\dagger$ is true operator and $\mathcal{G}$ is a candidate operator. Following previous works, we often took $||u||_{\mathcal{V}} = ||u||_{L^2} := (\int u(x)^2 dx)^{\frac{1}{2}}$, which in turn is discretized using the trapezoidal rule. In practice, we do not have the access to the underlying probability measure μ and we compute the empirical loss on a withheld test set:

$$\mathcal{R}_N(\mathcal{G}) = \frac{1}{N}\sum_{n=1}^{N}\left[\frac{||\mathcal{G}^\dagger(u^n) - \mathcal{G}(u^n)||_{\mathcal{V}}}{||\mathcal{G}^\dagger(u^n)||_{\mathcal{V}}}\right], \qquad u^i \sim \mu. \tag{5.48}$$

Measures of complexity

For our second performance metric we considered the complexity of operator learning algorithms at the inference stage (i.e., evaluating the learned operator). Complexity at inference time is the main metric used in (De Hoop et al., 2022) to compare numerical methods for operator learning. The motivation is that training of the methods can be performed in an offline fashion, and therefore the cost per test example dominates in the limit of many test queries. In particular, they compare the online evaluation costs of the neural networks by computing the requisite floating point operations (FLOPs) per test example. We adopt this metric as well for the methods not based on neural networks that we develop in this work, and we compare, when available, the cost-accuracy tradeoff with the numbers reported in (De Hoop et al.,

2022). We computed the FLOPs with the same assumptions as in the original work: a matrix-vector product where the input vector is in $\mathbb{R}^n$ and the output vector is in $\mathbb{R}^m$ amounts to $m(2-1)$ flops, and non-linear functions with n-dimensional inputs (activation functions for neural networks, kernel computations for kernel methods) are assumed to have cost $O(n)$.

Remark 8 (Training complexity)**.** While the inference complexity of a model eventually dominates the cost of training during applications, the training cost cannot be ignored since the allocated computational resources during this stage may still be limited and the resulting errors will have a profound impact on the quality and performance of the learned operators. Therefore numerical methods in which the offline data assimilation step is cheaper, faster, and more robust will always be preferred. Computing the exact number of FLOPs at training time is difficult to estimate for NN methods, as it depends on the optimization algorithms used, the hyperparameters and the optimization over such hyperparameters, among many other factors. Therefore in this work we limit the training complexity evaluation to the qualitative observation that kernel methods provided in this work are significantly simpler at training time, as they have no NN weights, they do not require the use of stochastic gradient descent, and have few or no hyperparameters which can be tuned using standard methods such as grid search or gradient descent in a low-dimensional space.

Test problems and qualitative results

Below we outline the setup of each of our benchmark problems. In all cases, $\mathcal{U}$ and $\mathcal{V}$ are spaces of real-valued functions with input domains $\Omega, D \subset \mathbb{R}^k$ for $k = 1$ or 2. Whenever $\Omega = D$, we simply write $\mathcal{D}$ for both.

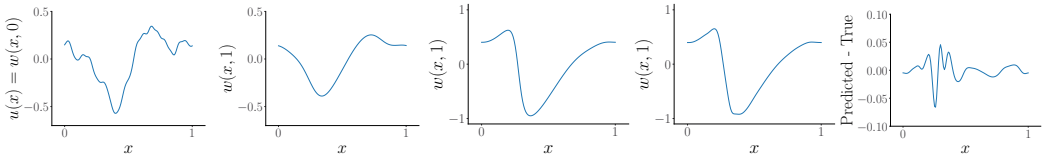
Burger's equation

Consider the one-dimensional Burger's equation:

$$\begin{aligned} \frac{\partial w}{\partial t} + w\frac{\partial w}{\partial x} &= \nu \frac{\partial^2 w}{\partial x^2}, \quad (x,t) \in (0,1) \times (0,1], \\ w(x,0) &= u(x), \qquad x \in (0,1) \end{aligned} \tag{5.49}$$

with $\mathcal{D} = (0,1)$, and periodic boundary conditions. The viscosity parameter ν is set to 0.1. We learn the operator mapping the initial condition u to $v = w(\cdot, 1)$, the solution at time $t = 1$, i.e., $\mathcal{G}^\dagger : w(\cdot, 0) \mapsto w(\cdot, 1)$.

The training data is generated by sampling the initial condition u from a GP with a Riesz kernel, denoted by $\mu = \mathcal{GP}(0, 625(-\Delta + 25I)^{-2})$. As in (L. Lu, X. Meng, et al., 2022), we used a spatial resolution with 128 grid points to represent the input and output functions, and used 1,000 instances for training and 200 instances for testing. Figure 5.4 shows an example of training input and output pairs as well as a test example along with its pointwise error.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

Figure 5.4: Example of training data and test prediction and pointwise errors for the Burger's equation (5.49).

Darcy flow

Consider the two-dimensional Darcy flow problem (5.3). Recall that in this example, we are interested in learning the mapping from the permeability field u to the solution v and the source term w is assumed to be fixed, and hence $\mathcal{D} \equiv \Omega = (0, 1)^2$ and $\mathcal{G}^\dagger : u \mapsto v$. The coefficient u is sampled by setting $u \sim \log \circ h_\sharp \mu$ where $\mu = \mathcal{GP}(0, (-\Delta + 9I)^{-2})$ is a GP and h is binary function mapping positive inputs to 12 and negative inputs to 3. The resulting permeability/diffusion coefficient e^u is therefore piecewise constant. As in (L. Lu, X. Meng, et al., 2022), we use a discretized grid of resolution 29×29, with the data generated by the MATLAB PDE Toolbox. We use 1,000 points for training and 200 points for testing. Figure 5.2 shows an example of training input and output of the map $\mathcal{G}^\dagger$, and an example of predictions along with pointwise error at the test stage.

Advection equations (I and II)

Consider the one-dimensional advection equation:

$$\begin{aligned} \frac{\partial w}{\partial t} + \frac{\partial w}{\partial x} &= 0 \qquad x \in (0, 1), t \in (0, 1] \\ w(x, 0) &= u(x) \quad x \in (0, 1) \end{aligned} \tag{5.50}$$

with $\mathcal{D} = (0, 1)$ and periodic boundary conditions. Similar to the example for Burgers' equation, we learn the mapping from the initial condition u to $v = w(\cdot, 0.5)$, the solution at $t = 0.5$, i.e., $\mathcal{G}^\dagger : w(\cdot, 0) \mapsto w(\cdot, 0.5)$.

This problem was considered in (L. Lu, X. Meng, et al., 2022; De Hoop et al., 2022) with different distributions μ for the initial condition. We will show in the following section how these different distributions lead to different performances. In (L. Lu, X. Meng, et al., 2022), henceforth referred to as Advection I, the initial condition is a square wave centered at $x = c$ of width b and height h:

$$u(x) = h\mathbf{1}_{\{c-\frac{b}{2}, c+\frac{b}{2}\}}, \tag{5.51}$$

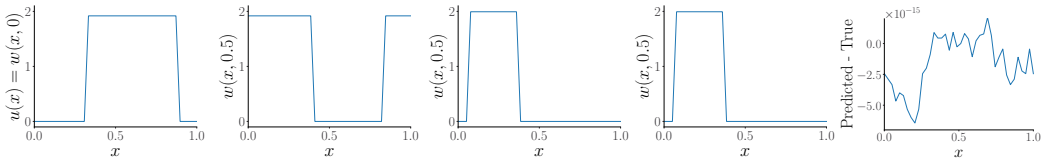
where the parameters $(c, b, h) \sim \mathcal{U}([0.3, 0.7] \times [0.3, 06] \times [1, 2])$. In (De Hoop et al., 2022), henceforth referred to as Advection II, the initial condition is

$$u = -1 + 2\mathbf{1}\{\widetilde{u}_0 \geq 0\}, \tag{5.52}$$

where $\widetilde{u}_0 \sim \mathcal{GP}(0, (-\Delta + 3^2 I)^{-2})$.

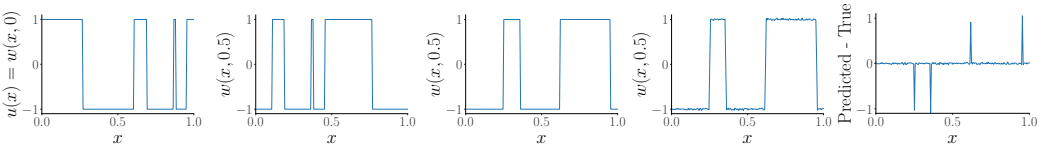
For Advection I, the spatial grid was of resolution 40, and we used 1,000 instances for training and 200 instances for testing. For Advection II, the resolution was of 200 and we used 20,000 training and test instances, following (De Hoop et al., 2022).

Figures 5.5 and 5.6 show an example of training input and output for Advection the I and II problems, respectively. Observe that the functional samples from the distribution in Advection I will have exactly two discontinuities almost surely, but the samples for Advection II can have many more jumps. We observe that prediction is challenging around discontinuities, and hence Advection II is a significantly harder problem (across all benchmarked methods) than Advection I. Figures 5.5 and 5.6 also show an instance of a test sample, along with a prediction and the pointwise errors.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

Figure 5.5: Example of training data and test prediction and pointwise errors for the Advection problem (5.50)-I.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

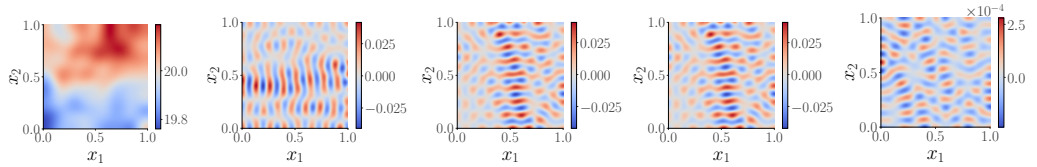
Figure 5.6: Example of training data and test prediction and pointwise errors for the Advection problem (5.50)-II.

Helmholtz's equation

For a given frequency ω and wavespeed field $u : \mathcal{D} \to \mathbb{R}$, with $\mathcal{D} = (0,1)^2$, the excitation field $v : \mathcal{D} \to \mathbb{R}$ solves

$$\left(-\Delta - \frac{\omega^2}{u^2(x)}\right) v = 0, \quad x \in (0,1)^2$$
$$\frac{\partial v}{\partial n} = 0, \quad x \in \{0,1\} \times [0,1] \cup [0,1] \times \{0\} \quad \text{and} \quad \frac{\partial v}{\partial n} = v_N, \quad x \in [0,1] \times \{1\} \tag{5.53}$$

In the results that follow, we take $\omega = 10^3$, $v_N = \mathbf{1}_{\{0.35 \le x \le 0.65\}}$, and we aim to learn the map $\mathcal{G} : u \mapsto v$, i.e., the mapping from the wavespeed field to the excitation field. The distribution μ is specified as the law of $u(x) = 20 + \tanh(\widetilde{u}(x))$, where $\widetilde{u}$ is drawn from the GP, $\mathcal{GP}(0, (-\Delta + 3^2 I)^{-2})$. The training and test data were generated by solving (5.53) with a Finite Element Method on a discretization of size 100×100 of the unit square. Figure 5.7 shows an example of training input and output, a test prediction, and pointwise errors.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

Figure 5.7: Example of training data and test prediction and pointwise errors for the Helmholtz problem (5.53).

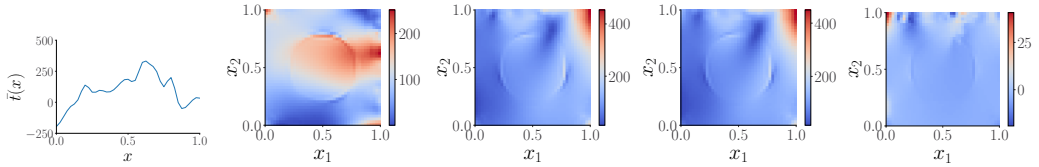
Structural mechanics

We let $\Omega = [0,1]$ $D = [0,1]^2$, the equation that governs the displacement vector w in an elastic solid undergoing infinitesimal deformations is

$$\nabla \cdot \sigma = 0 \quad \text{in } (0,1)^2, \qquad w = \bar{w}, \quad \text{on } \Gamma_w, \qquad \nabla \cdot n = u \quad \text{on } \Gamma_u, \tag{5.54}$$

where the boundary ∂D is split in $[0,1] \times 1 = \Gamma_t$ (the part of the boundary subject to stress) and its complement Γ_u.

The goal is to learn the operator that maps the one-dimensional load u on Γ_u to the two-dimensional von Mises stress field v on Ω, i.e., $\mathcal{G} : u \mapsto v$. Here the distribution μ is $\mathcal{GP}(100, 400^2(-\Delta + 3^2 I)^{-1})$, with Δ being the Laplacian subject to homogeneous Neumann boundary conditions on the space of zero-mean functions. The function v was obtained by a finite element code; see (De Hoop et al., 2022) for implementation details and the constitutive model used. Figure 5.8 shows an example of training input and outputs, a test prediction, and pointwise errors.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

Figure 5.8: Example of training data and test prediction and pointwise errors for the Structural Mechanics problem (5.54).

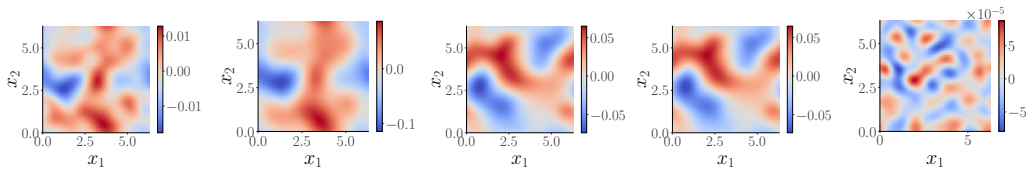
Navier-Stokes equations

Consider the vorticity-stream (ω, ψ) formulation of the incompressible Navier-Stokes equations:

$$\frac{\partial \omega}{\partial t} + (c \cdot \nabla)\omega - \nu \Delta \omega = u, \quad \omega = -\Delta \psi, \quad \int_D \psi = 0, \quad c = \left(\frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1} \right), \tag{5.55}$$

where $\mathcal{D} = [0, 2\pi]^2$, periodic boundary conditions are considered and the initial condition $w(\cdot, 0)$ is fixed. Here we are interested in the mapping from the forcing term u to $v = \omega(\cdot, T)$, the vorticity field at a given time $t = T$, i.e., $\mathcal{G}^\dagger : u \mapsto \omega(\cdot, T)$.

The distribution μ is $\mathcal{GP}(0, (-\Delta + 3^2 I)^{-4})$. The viscosity ν is fixed and equal to 0.025, and the equation is solved on a 64×64 grid with a pseudo-spectral method and Crank-Nicholson time integration; see (De Hoop et al., 2022) for further implementation details. Figure 5.9 shows an example of input and output in the test set, along with an example of test prediction and pointwise errors.



(a) Training input (b) Training output (c) True test (d) Predicted test (e) Pointwise error

Figure 5.9: Example of training data and test prediction and pointwise errors for the Navier-Stokes problem (5.55).

Results and discussion

Below we discuss our main findings in benchmarking our kernel method against state-of-the-art NN based techniques.

Performance against NNs

Table 5.3 summarizes the L^2 relative test error of our vanilla implementation of the kernel method along with those of DeepONet, FNO, PCA-Net, and PARA-Net. We observed that our vanilla kernel method was reliable in terms of accuracy across all examples. In particular, observe that between the Matérn or rational quadratic kernel, we always managed to get close to the other methods, see for example the results for the Burgers' equation or Darcy problem, and even outperform them in several examples such as Navier-Stokes and Helmholtz. Overall we observed that the performance of the kernel method is stable across all examples, suggesting that our method is reliable and provides a good baseline for a large class of problems. Moreover, we did not observe a significant difference in performance in terms of the choice of the particular kernel family once the hyper-parameters were tuned. This indicates that a large class of kernels are effective for these problems. Furthermore, we found the hyper-parameter tuning to be robust, i.e., results were consistent in a reasonable range of parameters such as length scales.

In the high data regime, we found the vanilla kernel method to be the most accurate, although this comes with a greater cost, as seen in Figure 5.10. However, the kernel method appears to provide the highest accuracy for its level of complexity as the accuracy of NNs typically stagnates or even decreases after a certain level of complexity; see the Navier-Stokes and Helmholtz panels of Figure 5.10 where most of the NN methods seem to plateau after a certain complexity level.

We also observed that the linear model did not provide the best accuracy as it quickly saturated in performance. Nonetheless, it provided surprisingly good accuracy at low

levels of complexity: for example, in the case of Navier-Stokes, the linear kernel provided the best accuracy below 10^6 FLOPS of complexity. This indicates that while simple, the linear model can be a valuable low-complexity model. Another notable example is the Advection equation (both I and II), where the operator $\mathcal{G}^\dagger$ is linear. In this case, the linear kernel had the best accuracy and the best complexity-accuracy tradeoff. We note, however, that while the linear model was close to machine precision on Advection I (error on the order $10^{-13}\%$), its performance was significantly worse on Advection II (error on the order of 10%). Moreover, the gap between the linear kernel and all other models was significantly smaller for Advection I; we conjecture this difference in performance is likely due to the setup of these problems.

Finally, we note that the most challenging problem for our kernel method was the Structural Mechanics example. In this case, the vanilla kernel method has higher complexity but did not beat the NNs. In fact, the NNs seem to be able to reduce complexity without loss of accuracy compared to our method.

	Low-data regime			High-data regime			
	Burger's	**Darcy problem**	**Advection I**	**Advection II**	**Hemholtz**	**Structural Mechanics**	**Navier Stokes**
DeepONet	2.15%	2.91%	0.66%	15.24%	5.88%	5.20%	3.63%
POD-DeepONet	1.94%	2.32%	0.04%	n/a	n/a	n/a	n/a
FNO	1.93%	2.41%	0.22%	13.49%	1.86%	4.76%	0.26%
PCA-Net	n/a	n/a	n/a	12.53%	2.13%	4.67%	2.65%
PARA-Net	n/a	n/a	n/a	16.64%	12.54%	4.55%	4.09%
Linear	36.24%	6.74%	$2.15 \times 10^{-13}\%$	11.28%	10.59%	27.11%	5.41%
Best of Matérn/RQ	2.15%	2.75%	$2.75 \times 10^{-3}\%$	11.44%	1.00%	5.18%	0.12%

Table 5.3: Summary of numerical results: we report the L^2 relative test error of our numerical experiments and compare the kernel approach with variations of DeepONet , FNO, PCA-Net, and PARA-Net. We considered two choices of the kernel S, the rational quadratic and the Matérn, but we observed little difference between the two.

Effect of preconditioners

Table 5.4 compares the performance of our method with the Matérn kernel family using various preconditioning steps. Overall we observed that both PCA and Cholesky preconditioning improved the performance of our vanilla kernel method.

The Cholesky preconditioning generally offers the greatest improvement. However, we observed that getting the best results from the Cholesky approach required careful tuning of the parameters of the kernels K and Q which we did using cross-validation. While tuning the parameters does not increase the inference complexity, it does increase the training complexity.

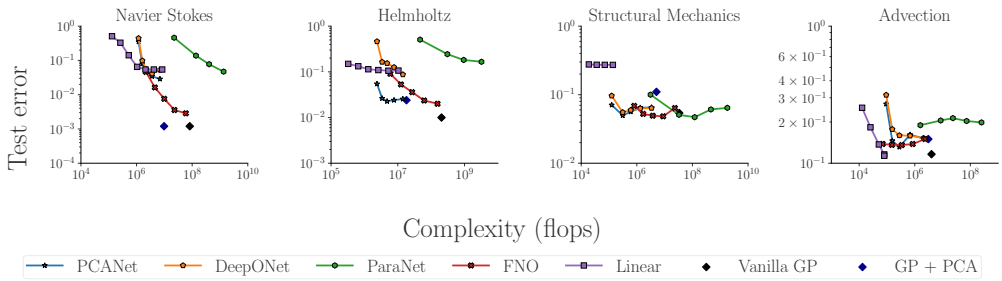


Figure 5.10: Accuracy complexity tradeoff achieved in the problems in (De Hoop et al., 2022). Data for NNs was obtained from the aforementioned article. Linear model refers to the linear kernel, vanilla GP is our implementation with the nonlinear kernels and minimal preprocessing, GP+PCA corresponds to preprocessing through PCA both the input and the output to reduce complexity.

On the other hand, the PCA approach was more robust to changes in hyperparameters, i.e., the number of PCA components following Section 5.4. We observed that applying PCA on the input and output reduces complexity and has varying levels of effectiveness in providing a better cost-accuracy tradeoff. For example, for Navier-Stokes, it greatly reduced the complexity without affecting accuracy. But for the Helmholtz and Advection equations, PCA reduced the accuracy while remaining competitive with NN models. For structural mechanics, however, PCA significantly reduced accuracy and was worse than other models. We hypothesize that the loss in accuracy can be related to the decay of the eigenvalues of the PCA matrix in that example.

	Advection II	**Burger's**	**Darcy problem**
No preprocessing	14.37%	3.04%	4.47%
PCA	14.50%	2.41%	2.89%
Cholesky	11.44%	2.15%	2.75%

Table 5.4: Comparison between Cholesky preconditioning and PCA dimensionality reduction on three examples for our vanilla kernel implementation with the Matérn kernel.

5.5 Conclusions

In this work we presented a kernel/GP framework for the learning of operators between function spaces. We presented an abstract formulation of our kernel framework along with convergence proofs and error bounds in certain asymptotic limits. Numerical experiments and benchmarking against popular NN based algorithms revealed that our vanilla implementation of the kernel approach is competitive and either matches the performance of NN methods or beats them in several benchmarks.

Due to simplicity of implementation, flexibility, and the empirical results, we suggest that the proposed kernel methods are a good benchmark for future, perhaps more sophisticated, algorithms. Furthermore, these methods can be used to guide practitioners in the design of new and challenging benchmarks (e.g, identify problems where vanilla kernel methods do not perform well). Numerous directions of future research exist. In the theoretical direction it is interesting to remove the stringent Condition 5.3.2 and we anticipate this to require a particular selection of the kernel employed to obtain the map $\bar{f}$. Moreover, obtaining error bounds for more general measurement functionals beyond pointwise evaluations would be interesting. One could also adapt our framework to non-vanilla kernel methods such as random features or inducing point methods to provide a low-complexity alternative to NNs in the large-data regime. Finally, since the proposed approach is essentially a generalization of GP Regression to the infinite-dimensional setting, we anticipate that some of the hierarchical techniques of (Houman Owhadi, 2017; Schaäfer, Katzfuss, and Houman Owhadi, 2021; Florian Schäfer, Timothy John Sullivan, and Houman Owhadi, 2021a) could be extended to this setting and provide a better cost-accuracy trade-off than current methods.

5.6 Review of operator valued kernels and GPs

We review the theory of operator valued kernels and GPs (Houman Owhadi, 2023a) as these are utilized throughout the article. Operator-valued kernels were introduced in (Kadri et al., 2016) as a generalization of vector-valued kernels (Alvarez, Rosasco, Lawrence, et al., 2012).

Operator valued kernels

Let $\mathcal{U}$ and $\mathcal{V}$ be separable Hilbert spaces endowed with the inner products $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{V}}$. Write $\mathcal{L}(\mathcal{V})$ for the set of bounded linear operators mapping $\mathcal{V}$ to $\mathcal{V}$.

Definition 5.6.1. We call $G : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V})$ an "operator-valued kernel" if

1. G is Hermitian, i.e. $G(u, u') = G(u', u)^T$ for all $u, u' \in \mathcal{U}$, writing A^T for the adjoint of the operator A with respect to $\langle \cdot, \cdot \rangle_{\mathcal{V}}$.
2. G is non-negative, i.e., for all $m \in \mathbb{N}$ and any set of points $(u_i, v_i)_{i=1}^m \subset \mathcal{U} \times \mathcal{V}$ it holds that $\sum_{i,j=1}^m \langle v_i, G(u_i, u_j) v_j \rangle_{\mathcal{V}} \geq 0$.

We call G non-degenerate if $\sum_{i,j=1}^m \langle v_i, G(u_i, u_j) v_j \rangle_{\mathcal{V}} = 0$ implies $v_i = 0$ for all i whenever $u_i \neq u_j$ for $i \neq j$.

RKHSs

Each non-degenerate, locally bounded and separately continuous operator-valued kernel G is in one to one correspondence with an RKHS $\mathcal{H}$ of continuous operators $\mathcal{G} : \mathcal{U} \to \mathcal{V}$ obtained as the closure of the linear span of the maps $z \mapsto G(z,u)v$ with respect to the inner product identified by the reproducing property

$$\left\langle g, G(\cdot,u)v \right\rangle_{\mathcal{H}} = \left\langle g(u), v \right\rangle_{\mathcal{V}}. \tag{5.56}$$

Feature maps

Let $\mathcal{F}$ be a separable Hilbert space (with inner product $\left\langle \cdot,\cdot \right\rangle_{\mathcal{F}}$ and norm $\|\cdot\|_{\mathcal{F}}$) and let $\psi : \mathcal{U} \to \mathcal{L}(\mathcal{V},\mathcal{F})$ be a continuous function mapping $\mathcal{U}$ to the space of bounded linear operators from $\mathcal{V}$ to $\mathcal{F}$.

Definition 5.6.2. We say that $\mathcal{F}$ and $\psi : \mathcal{U} \to \mathcal{L}(\mathcal{V},\mathcal{F})$ are a *feature space* and a *feature map* for the kernel G if, for all $(u,u',v,v') \in \mathcal{U}^2 \times \mathcal{V}^2$,

$$\left\langle v, G(u,u')v' \right\rangle = \left\langle \psi(u)v, \psi(u')v' \right\rangle_{\mathcal{F}}.$$

Write $\psi^T(u)$, for the adjoint of $\psi(u)$ defined as the linear function mapping $\mathcal{F}$ to $\mathcal{V}$ satisfying

$$\left\langle \psi(u)v, \alpha \right\rangle_{\mathcal{F}} = \left\langle v, \psi^T(u)\alpha \right\rangle_{\mathcal{V}}$$

for $u,v,\alpha \in \mathcal{U}\times\mathcal{V}\times\mathcal{F}$. Note that $\psi^T : \mathcal{U} \to \mathcal{L}(\mathcal{F},\mathcal{V})$ is therefore a function mapping $\mathcal{U}$ to the space of bounded linear functions from $\mathcal{F}$ to $\mathcal{V}$. Writing $\alpha^T\alpha' := \left\langle \alpha,\alpha' \right\rangle_{\mathcal{F}}$ for the inner product in $\mathcal{F}$ we can ease our notations by writing

$$G(u,u') = \psi^T(u)\psi(u') \tag{5.57}$$

which is consistent with the finite-dimensional setting and $v^T G(u,u')v' = (\psi(u)v)^T(\psi(u')v')$ (writing v^Tv' for the inner product in $\mathcal{V}$). For $\alpha \in \mathcal{F}$ write $\psi^T\alpha$ for the function $\mathcal{U} \to \mathcal{V}$ mapping $u \in \mathcal{U}$ to the element $v \in \mathcal{V}$ such that

$$\left\langle v', v \right\rangle_{\mathcal{V}} = \left\langle v', \psi^T(u)\alpha \right\rangle_{\mathcal{V}} = \left\langle \psi(u)v', \alpha \right\rangle_{\mathcal{F}} \text{ for all } v' \in \mathcal{V}.$$

We can, without loss of generality, restrict $\mathcal{F}$ to be the range of $(u,v) \to \psi(u)v$ so that the RKHS $\mathcal{H}$ defined by G is the closure of the pre-Hilbert space spanned by $\psi^T\alpha$ for $\alpha \in \mathcal{F}$. Note that the reproducing property (5.56) implies that for $\alpha \in \mathcal{F}$

$$\left\langle \psi^T(\cdot)\alpha, \psi^T(\cdot)\psi(u)v \right\rangle_{\mathcal{H}} = \left\langle \psi^T(u)\alpha, v \right\rangle_{\mathcal{V}} = \left\langle \alpha, \psi(u)v \right\rangle_{\mathcal{F}}$$

for all $u,v \in \mathcal{U}\times\mathcal{V}$, which leads to the following theorem.

Theorem 5.6.3. The RKHS $\mathcal{H}$ defined by the kernel (5.57) is the linear span of $\psi^T\alpha$ over $\alpha \in \mathcal{F}$ such that $\|\alpha\|_{\mathcal{F}} < \infty$. Furthermore, $\left\langle \psi^T(\cdot)\alpha, \psi^T(\cdot)\alpha' \right\rangle_{\mathcal{H}} = \left\langle \alpha, \alpha' \right\rangle_{\mathcal{F}}$ and

$$\|\psi^T(\cdot)\alpha\|^2_{\mathcal{H}} = \|\alpha\|^2_{\mathcal{F}} \text{ for } \alpha, \alpha' \in \mathcal{F}\,.$$

Interpolation

Let us consider the interpolation problem in operator valued RKHSs.

Problem 3. Let $\mathcal{G}^\dagger$ be an unknown continuous operator mapping $\mathcal{U}$ to $\mathcal{V}$. Given the information[8] $\mathcal{G}^\dagger(\mathbf{u}) = \mathbf{v}$ with the data $(\mathbf{u}, \mathbf{v}) \in \mathcal{U}^N \times \mathcal{V}^N$, approximate $\mathcal{G}^\dagger$.

Using the relative error in $\|\cdot\|_{\mathcal{H}}$-norm as a loss, the minimax optimal recovery solution of Problem 3 is, by (Houman Owhadi and Clint Scovel, 2019a, Thm. 12.4, 12.5), given by

$$\begin{cases} \text{Minimize} & \|\mathcal{G}\|^2_{\mathcal{H}} \\ \text{subject to} & \mathcal{G}(\mathbf{u}) = \mathbf{v}. \end{cases} \tag{5.58}$$

The minimizer is then of the form $\mathcal{G}(\cdot) = \sum_{j=1}^N G(\cdot, u_j)w_j$, where the coefficients $w_j \in \mathcal{V}$ are identified by solving the system of linear equations $\sum_{j=1}^N G(u_i, u_j)w_j = v_i$ for all $i \in \{1, \ldots, N\}$. Using our compressed notation we can rewrite this equation as $G(\mathbf{u}, \mathbf{u})\mathbf{w} = \mathbf{v}$ where $\mathbf{w} = (w_1, \ldots, w_N)$, $\mathbf{v} = (v_1, \ldots, v_N) \in \mathcal{V}^N$ and $G(\mathbf{u}, \mathbf{u})$ is the $N \times N$ block-operator matrix [9] with entries $G(u_i, u_j)$. Therefore, writing $G(\cdot, \mathbf{u})$ for the vector $(G(\cdot, u_1), \ldots, G(\cdot, u_N)) \in \mathcal{H}^N$, the optimal recovery interpolant is given by

$$\bar{\mathcal{G}}(\cdot) = G(\cdot, \mathbf{u})G(\mathbf{u}, \mathbf{u})^{-1}\mathbf{v}\,, \tag{5.59}$$

which implies that the value of (5.58) at the minimum is

$$\|\bar{\mathcal{G}}\|^2_{\mathcal{H}} = \mathbf{v}^T G(\mathbf{u}, \mathbf{u})^{-1}\mathbf{v}\,, \tag{5.60}$$

where $G(\mathbf{u}, \mathbf{u})^{-1}$ is the inverse of $G(\mathbf{u}, \mathbf{u})$, whose existence is implied by the non-degeneracy of G combined with $u_i \neq u_j$ for $i \neq j$.

[8] For a N-vector $\mathbf{u} = (u_1, \ldots, u_N) \in \mathcal{U}^N$ and a function $\mathcal{G} : \mathcal{U} \to \mathcal{V}$, write $\mathcal{G}(\mathbf{u})$ for the N vector with entries $\big(\mathcal{G}(u_1), \ldots, \mathcal{G}(u_N)\big)$.

[9] For $N \geq 1$ let $\mathcal{V}^N$ be the N-fold product space endowed with the inner-product $\left\langle \mathbf{v}, \mathbf{w} \right\rangle_{\mathcal{V}^N} := \sum_{i,j=1}^N \left\langle v_i, w_j \right\rangle_{\mathcal{V}}$ for $\mathbf{v} = (v_1, \ldots, v_N), \mathbf{w} = (w_1, \ldots, w_N) \in \mathcal{V}^N$. $\mathbf{A} \in \mathcal{L}(\mathcal{V}^N)$ given by $\mathbf{A} = \begin{pmatrix} A_{1,1} & \cdots & A_{1,N} \\ \vdots & & \vdots \\ A_{N,1} & \cdots & A_{N,N} \end{pmatrix}$ where $A_{i,j} \in \mathcal{L}(\mathcal{V})$, is called a block-operator matrix. Its adjoint $\mathbf{A^T}$ with respect to $\left\langle \cdot, \cdot \right\rangle_{\mathcal{V}^N}$ is the block-operator matrix with entries $(A^T)_{i,j} = (A_{j,i})^T$.

Ridge regression

Let $\gamma > 0$. A ridge regression (approximate) solution to Problem 3 can be found as the minimizer of

$$\inf_{\mathcal{G}\in\mathcal{H}} \lambda \|\mathcal{G}\|_{\mathcal{H}}^2 + \gamma^{-1} \sum_{i=1}^{N} \|v_i - \mathcal{G}(u_i)\|_{\mathcal{V}}^2 . \tag{5.61}$$

This minimizer is given by the formula

$$\bar{\mathcal{G}}(u) = G(u, \mathbf{u})\big(G(\mathbf{u}, \mathbf{u}) + \gamma I\big)^{-1}\mathbf{v} , \tag{5.62}$$

writing I for the identity matrix. We can further compute directly

$$\|\bar{\mathcal{G}}\|_{\mathcal{H}}^2 = \mathbf{v}^T\big(G(\mathbf{u}, \mathbf{u}) + \gamma I\big)^{-1}\mathbf{v} .$$

Operator-valued GPs

The following definition of operator-valued Gaussian processes is a natural extension of scalar-valued Gaussian fields (Houman Owhadi and Clint Scovel, 2019a).

Definition 5.6.4. (Houman Owhadi, 2023a, Def. 5.1) Let $G : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V})$ be an operator-valued kernel. Let m be a function mapping $\mathcal{U}$ to $\mathcal{V}$. We call $\xi : \mathcal{U} \to \mathcal{L}(\mathcal{V}, \mathbf{H})$ an operator-valued GP if ξ is a function mapping $u \in \mathcal{U}$ to $\xi(u) \in \mathcal{L}(\mathcal{V}, \mathbf{H})$ where $\mathbf{H}$ is a Gaussian space and $\mathcal{L}(\mathcal{V}, \mathbf{H})$ is the space of bounded linear operators from $\mathcal{V}$ to $\mathbf{H}$. Abusing notations we write $\langle \xi(u), v\rangle_{\mathcal{V}}$ for $\xi(u)v$. We say that ξ has mean m and covariance kernel G and write $\xi \sim \mathcal{N}(m, G)$ if $\langle \xi(u), v\rangle_{\mathcal{V}} \sim \mathcal{N}\big(m(u), v^T G(u, u)v\big)$ and

$$\mathrm{Cov}\left(\langle \xi(u), v\rangle_{\mathcal{V}}, \langle \xi(u'), v'\rangle_{\mathcal{V}}\right) = v^T G(u, u')v' . \tag{5.63}$$

We say that ξ is centered if it is of zero mean.

If $G(u, u)$ is trace class ($\mathrm{Tr}[G(u, u)] < \infty$) then $\xi(u)$ defines a measure on $\mathcal{V}$, i.e. a $\mathcal{V}$-valued random variable [10].

Theorem 5.6.5. (Houman Owhadi, 2023a, Thm. 5.2) The law of an operator-valued GP is uniquely determined by its mean m and covariance kernel G. Conversely given m and G there exists an operator-valued GP having mean m and covariance kernel G. In particular if G has feature space $\mathcal{F}$ and map ψ, the e_i form an orthonormal basis of $\mathcal{F}$, and the Z_i are i.i.d. $\mathcal{N}(0, 1)$ random variables, then $\xi = m + \sum_i Z_i \psi^T e_i$ is an operator-valued GP with mean m and covariance kernel G.

[10] Otherwise it only defines a (weak) cylinder-measure in the sense of Gaussian fields.

Theorem 5.6.6. (Houman Owhadi, 2023a, Thm. 5.3) Let ξ be a centered operator-valued GP with covariance kernel $G : \mathcal{U} \times \mathcal{U} \to \mathcal{L}(\mathcal{V})$. Let $\mathbf{u}, \mathbf{v} \in \mathcal{U}^N \times \mathcal{V}^N$. Let $Z = (Z_1, \ldots, Z_N)$ be a random Gaussian vector, independent from ξ, with i.i.d. $\mathcal{N}(0, \gamma I_{\mathcal{V}})$ entries ($\gamma \geq 0$ and $I_{\mathcal{V}}$ is the identity map on $\mathcal{V}$). Then ξ conditioned on $\xi(\mathbf{u}) + Z$ is an operator-valued GP with mean

$$\mathbb{E}\big[\xi(u)\big|\xi(\mathbf{u}) + Z = \mathbf{v}\big] = G(u, \mathbf{u})\big(G(\mathbf{u}, \mathbf{u}) + \gamma I_{\mathcal{V}}\big)^{-1}\mathbf{v} = (5.62) \tag{5.64}$$

and conditional covariance operator

$$G^{\perp}(u, u') := G(u, u') - G(u, \mathbf{u})\big(G(\mathbf{u}, \mathbf{u}) + \gamma I_{\mathcal{V}}\big)^{-1}G(\mathbf{u}, u') \,. \tag{5.65}$$

In particular, if G is trace class, then

$$\sigma^2(u) := \mathbb{E}\Big[\big\|\xi(u) - \mathbb{E}[\xi(u)|\xi(\mathbf{u}) + Z = \mathbf{v}]\big\|_{\mathcal{V}}^2\Big|\xi(\mathbf{u}) + Z = \mathbf{v}\Big] = \mathrm{Tr}\left[G^{\perp}(u, u)\right] . \tag{5.66}$$

Deterministic error estimates for operator-valued regression

The following theorem shows that the standard deviation (5.66) provides deterministic a prior error bounds on the accuracy of the ridge regressor (5.64) to $\mathcal{G}^{\dagger}$ in Problem 3. Local error estimates such as (5.67) below are classical in the Kriging literature (Wu and Schaback, 1993) where $\sigma^2(u)$ is known as the power function/kriging variance; see also (Houman Owhadi, 2015a)[Thm. 5.1] for applications to PDEs.

Theorem 5.6.7. (Houman Owhadi, 2023a, Thm. 5.4) Let $\mathcal{G}^{\dagger}$ be the unknown function of Problem 3 and let $\mathcal{G}(u) = (5.64) = (5.62)$ be its ridge regressor. Let $\mathcal{H}$ be the RKHS associated with G and let $\mathcal{H}_\gamma$ be the RKHS associated with the kernel $G_\gamma := G + \gamma I_{\mathcal{V}}$. It holds true that

$$\big\|\mathcal{G}^{\dagger}(u) - \mathcal{G}(u)\big\|_{\mathcal{V}} \leq \sigma(u)\|\mathcal{G}^{\dagger}\|_{\mathcal{H}} \tag{5.67}$$

and

$$\big\|\mathcal{G}^{\dagger}(u) - \mathcal{G}(u)\big\|_{\mathcal{V}} \leq \sqrt{\sigma^2(u) + \gamma \dim(\mathcal{V})}\|\mathcal{G}^{\dagger}\|_{\mathcal{H}_\gamma} \,, \tag{5.68}$$

where $\sigma(u)$ is the standard deviation (5.66).

5.7 An alternative regularization of operator regression

For $\gamma > 0$, the regularization implied by (5.27) is equivalent to adding noise on the $\varphi(\mathbf{v})$ measurements. If one could observe $\mathbf{v}$ (and not just $\varphi(\mathbf{v})$), then an alternative approach to regularizing the problem is to add noise to $\xi(\mathbf{u})$. To describe this let

$Z' = (Z'_1, \ldots, Z'_N)$ be a random block-vector, independent from ξ, with i.i.d. entries $Z'_j \sim \mathcal{N}(0, \gamma I_{\mathcal{V}})$ for $j = 1, \ldots, N$ (where $I_{\mathcal{V}}$ denotes the identity map on $\mathcal{V}$). Then the GP ξ conditioned on $\xi(\mathbf{u}) = \mathbf{v} + Z'$ is a GP with conditional covariance kernel (5.65) and conditional mean $\widetilde{\mathcal{G}}_\gamma$=(5.62) that is also the minimizer of (5.61). Observing[11] that $\varphi(Z'_i) \sim \mathcal{N}(0, \gamma K(\varphi, \varphi))$, we deduce that $\widetilde{\mathcal{G}}_\gamma = \chi \circ \widetilde{f}_\gamma \circ \phi$ where $\widetilde{f}_\gamma$ minimizes

$$\begin{cases} \text{Minimize} & \|f\|_\Gamma^2 + \gamma^{-1} \sum_{i=1}^N (f(U_i) - V_i)^T K(\varphi, \varphi)^{-1} (f(U_i) - V_i) \,. \\ \text{Over} & f \in \mathcal{H}_\Gamma \,. \end{cases} \tag{5.69}$$

Furthermore, the distribution of ξ conditioned on $\xi(\mathbf{u}) = \mathbf{v} + Z'$ is that of $\chi \circ \widetilde{\zeta}^\perp \circ \phi$ where $\widetilde{\zeta}^\perp \sim \mathcal{N}(\widetilde{f}_\gamma, \widetilde{\Gamma}^\perp)$ is the GP ζ conditioned on $\zeta(\mathbf{U}) = \mathbf{V} + \varphi(Z')$, whose mean is $\widetilde{f}_\gamma$ and conditional covariance kernel is $\widetilde{\Gamma}^\perp(U, U') = \Gamma(U, U') - \Gamma(U, \mathbf{U})(\Gamma(\mathbf{U}, \mathbf{U}) + \gamma A)^{-1}\Gamma(\mathbf{U}, U')$ where A is a $N \times N$ block diagonal matrix with $K(\varphi, \varphi)$ as diagonal entries.

5.8 Expressions for the kernels used in experiments

Below we collect the expressions for the kernels that were referred to in the article or utilized for our numerical experiments. These can be found in many standard textbooks on GPs such as (Rasmussen and Williams, 2006).

The linear kernel

The linear kernel has the simple expression $K_{\text{linear}}(x, x') = \langle x, x' \rangle$ and may be defined on any inner product space. It has no hyper-parameters.

The rational quadratic kernel

The rational quadratic kernel has the expression $K(x, x') = k_{\text{RQ}}(\|x - x'\|)$ where

$$k_{\text{RQ}}(r) = \left(1 + \frac{r^2}{2l^2}\right)^{-\alpha} . \tag{5.70}$$

It has hyper-parameters $\alpha > 0$ and l.

The Matérn parametric family

The Matérn kernel family is of the form $K(x, x') = k(\|x - x'\|$ where

$$k_\nu(r) = \exp\left(-\frac{\sqrt{2\nu}r}{l}\right) \frac{\Gamma(p+1)}{\Gamma(2p+1)} \sum_{i=0}^{p} \frac{(p+1)!}{i!(p-i)!} \left(\frac{\sqrt{8\nu}r}{l}\right)^{p-i} , \tag{5.71}$$

[11] This follows from $\varphi(Z'_i) \sim \mathcal{N}(0, \gamma\varphi\varphi^T)$ where φ^T is the adjoint of φ identified as the linear map from $\mathbb{R}^m$ to $\mathcal{V}$ satisfying $\langle W, \varphi(w) \rangle_{\mathbb{R}^m} = \langle \varphi^\perp W, w \rangle_{\mathcal{V}}$ for $w \in \mathcal{V}$ and $W \in \mathbb{R}^m$ (i.e., $\varphi^T(W) = (K\varphi)W$).

for $\nu = p + \frac{1}{2}$. This kernel has hyper-parameters $p \in \mathbb{Z}_+$ and $l > 0$. In the limiting case where $\nu \to \infty$, the Matérn kernel, we obtain the Gaussian or squared exponential kernel:

$$k_\infty(r) = \exp\left(-\frac{r^2}{2l^2}\right), \tag{5.72}$$

with hyper-parameter $l > 0$.

Chapter 6

ERROR ANALYSIS OF KERNEL/GP METHODS FOR NONLINEAR AND PARAMETRIC PDES

We introduce a priori Sobolev-space error estimates for the solution of arbitrary nonlinear, and possibly parametric, PDEs that are defined in the strong sense, using Gaussian process and kernel based methods. The primary assumptions are (1) a continuous embedding of the reproducing kernel Hilbert space of the kernel into a Sobolev space of sufficient regularity, and (2) the stability of the differential operator and the solution map of the PDE between corresponding Sobolev spaces. The proof is articulated around Sobolev norm error estimates for kernel interpolants and relies on the minimizing norm property of the solution. The error estimates demonstrate dimension-benign convergence rates if the solution space of the PDE is smooth enough. We illustrate these points with applications to high-dimensional nonlinear elliptic PDEs and parametric PDEs. Although some recent machine learning methods have been presented as breaking the curse of dimensionality in solving high-dimensional PDEs, our analysis suggests a more nuanced picture: there is a trade-off between the regularity of the solution and the presence of the curse of dimensionality. Therefore, our results are in line with the understanding that the curse is absent when the solution is regular enough.

6.1 Introduction

In recent years the adoption of machine learning in the natural sciences and engineering has led to the development of new methods for solving PDEs (Raissi, Perdikaris, and George E Karniadakis, 2019; Weinan, Han, and Jentzen, 2017; Weinan and B. Yu, 2018; Li, Kovachki, Azizzadenesheli, Bhattacharya, et al., 2020; L. Lu, Jin, et al., 2021). The majority of these methods rely on the approximation power of artificial neural networks (ANNs) either as a function class to approximate the solution of the PDE or as a high-dimensional function class to approximate the solution map of the PDE. Despite the empirical success of the aforementioned ANN based methods, current theoretical understanding of these PDE solvers is scarce and, beyond particular PDEs (e.g., (Shin, Darbon, and George Em Karniadakis, 2020; Y. Lu et al., 2021; De Ryck and Mishra, 2022)), results are oftentimes limited to existence results rather than convergence guarantees or rates.

Similar to ANNs, kernel methods and Gaussian processes (GPs) have been very effective in scientific computing and machine learning (Scholkopf and Alexander J Smola, 2018; Muandet et al., 2017; Berlinet and Thomas-Agnan, 2011; Williams and Rasmussen, 2006) and at the same time they are supported by rigorous theoretical foundation (Berlinet and Thomas-Agnan, 2011; Wendland, 2004; Houman Owhadi and Clint Scovel, 2019b). Recently in (Yifan Chen, Hosseini, et al., 2021b), the authors introduced a kernel collocation method for solving arbitrary nonlinear PDEs with a rigorous convergence guarantee. The theory presented in that work was based on the assumptions that (1) the solution belongs to the reproducing kernel Hilbert space (RKHS) defined by the underlying kernel which in turn is embedded in the Sobolev space H^s for $s > d/2 +$"order of the PDE" (where d is the dimension of the domain of the PDE) and (2) the fill-distance between collocation points goes to zero. Convergence was proved via a compactness argument but no convergence rates were provided.

The goal of this article is to provide quantitative convergence rates for the PDE solver introduced in (Yifan Chen, Hosseini, et al., 2021b). Our quantitaitve rates also reveal the interplay between the regularity of the solution of the PDE and the dimension d of the problem. At the same time we make improvements to the methodology of (Yifan Chen, Hosseini, et al., 2021b) and extend it to the case of parametric PDEs. In the rest of this section we summarize our main contributions in Section 6.1 followed by a review of the relevant literature in Section 6.1, and an outline of the article in Section 6.1.

Significance of Contributions

The use of meshless, collocation methods with radial basis functions for solving PDEs dates back to the 1990s (Kansa, 1990a; Kansa, 1990b; Franke and Schaback, 1998b; Fasshauer, 1999). Typical approaches include symmetric collocation (Franke and Schaback, 1998b; Fasshauer, 1999) and unsymmetric collocation methods (also known as the Kansa method (Kansa, 1990a; Kansa, 1990b)); see (Schaback and Wendland, 2006; Fornberg and Flyer, 2015) for reviews. It is recognized that the unsymmetric collocation approach may encounter instability issues and require additional techniques (Ling, Opfer, and Schaback, 2006; Schaback, 2007; Ling and Schaback, 2008; Cheung, Ling, and Schaback, 2018), whereas the symmetric collocation approach always yields positive definite symmetric matrices and is stable. The reason is that the symmetric collocation approach includes higher-order derivatives of the kernel as basis functions and leads to solutions that can be

identified as *optimal recovery* (worst case minimax optimal) solutions. This optimal recovery property makes the algorithm generally applicable for *any well-posed linear PDEs* (Hon and Schaback, 2008; Schaback, 2015; Schaback, 2016). Furthermore the analysis becomes straightforward because of this optimality; see (Franke and Schaback, 1998a; Franke and Schaback, 1998b) for linear PDEs and (Böhmer and Schaback, 2013; Böhmer and Schaback, 2020) for some degree of generalization to quasi-linear/nonlinear problems. Note that the optimality has long been recognized, but not extensively acknowledged however, as highlighted in (Schaback, 2015): "This technique has been around since at least 1998, but its optimality properties went unnoticed." The GP-PDE methodology proposed in (Yifan Chen, Hosseini, et al., 2021b) can be seen as a nonlinear generalization of the optimal recovery approach to solving (and learning) *arbitrary classically/strongly defined nonlinear PDEs*. This paper aims to offer a simple and transparent theoretical error analysis of this nonlinear optimal recovery method. This analysis extends the linear setting (Giesl and Wendland, 2007) and shares conceptual steps (in terms of the role of stability and sampling inequalities) with (Böhmer and Schaback, 2013) while explicitly focusing on optimal recovery solutions rather than solutions obtained from finite-dimensional trial spaces and residual minimization. Such theoretical analyses are notably rare within the sphere of machine learning-based PDE solutions (where theoretical guarantees are typically limited to existence results). Furthermore, while there is no general theory for strongly defined arbitrary nonlinear PDEs, the optimal recovery approach provides a way of obtaining general theoretical guarantee for the numerical approximation of such PDEs. Beyond its wide scope, the proposed analysis also lays the groundwork for developing rigorous, efficient, and scalable (near-linear complexity) learning-based methods for arbitrary nonlinear PDEs. This can be achieved by integrating the proposed error estimates with the fast algorithms developed in (Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024) for kernel matrices whose entries contain higher-order derivatives of the kernel, a setting well suited for the optimal recovery approach. We note that the analysis in the paper is focused on the minimizer of a loss function induced by kernel and GP methods. Understanding theoretically how iterative algorithms are able to achieve this minimizer represents another crucial stride towards our ultimate objective; this could be potentially done by combining analysis results for iterative linerization, for example the work in (Becker et al., 2023).

Summary of Contributions

Throughout the article we consider parametric PDEs of the form

$$\begin{cases} \mathcal{P}(u^\star)(\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{x};\boldsymbol{\theta}), & (\boldsymbol{x},\boldsymbol{\theta}) \in \Omega \times \Theta, \\ \mathcal{B}(u^\star)(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{x};\boldsymbol{\theta}), & (\boldsymbol{x},\boldsymbol{\theta}) \in \partial\Omega \times \Theta, \end{cases} \tag{6.1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded connected domain with an appropriately smooth boundary $\partial\Omega$, $\mathcal{P}$ and $\mathcal{B}$ are the interior and boundary differential operators that define the PDE and f, g are the source and boundary data. $\boldsymbol{x}$ denotes the spatial variable with $\boldsymbol{\theta}$ denoting a parameter belonging to a compact set $\Theta \subset \mathbb{R}^p$. The function $u^\star$ denotes the exact, strong solution of this PDE.

We view the solution $u^\star$ as a function on $\overline{\Omega} \times \Theta$ and approximate it in an appropriate RKHS by imposing the PDE as a constraint on a set of collocation points in the product space $\overline{\Omega} \times \Theta$. Our main contributions are four-fold as summarized below:

1. We extend the kernel PDE solver of (Yifan Chen, Hosseini, et al., 2021b) to the case of the parametric PDE (6.1). This extension follows by viewing the solution $u^\star(\boldsymbol{x};\theta)$ as a continuous function defined on $\overline{\Omega}\times\Theta$ and approximating it with a function $u^\dagger$ in an appropriate RKHS $\mathcal{U}$ after imposing the PDE as a constraint on a set of collocation points. At the same time we improve the efficacy and performance of the Gauss-Newton (GN) algorithm of (Yifan Chen, Hosseini, et al., 2021b) through an approach of "linearize first then apply the kernel solver". For many prototypical PDEs this new approach leads to smaller kernel matrices that can be factored or inverted more efficiently. These numerical strategies are outlined in Section 6.2, and our proposed methodology is summarized in Algorithms 9 and 10.

2. We provide explicit a priori convergence rates for the kernel estimator $u^\dagger \in \mathcal{U}$ to the true solution $u^\star$. Our proof relies on three assumptions: (1) the RKHS $\mathcal{U} \subset H^s(\Omega)$ for $s > (d+p)/2 +$ "order of the PDE", (2) the true unique solution $u^\star \in \mathcal{U}$, and (3) the forward PDE operator and the associated solution map of the PDE are Lipschitz stable. Our error estimates are of the general form

$$\|u^\dagger - u^\star\|_{L^2(\Omega)} \lesssim h^s \|u^\star\|_{\mathcal{U}}, \tag{6.2}$$

where h is the fill-distance (mesh-norm) of our collocation points. Indeed, if expressed in terms of N, the number of collocation points, the rate will read as $O(N^{-s/(d+p)})$. The above rate indicates a trade-off between the regularity

of the solution space and the dimension $d + p$ of $\Omega \times \Theta$ stating that the convergence rate is dimension-benign so long as the solution $u^\star$ is sufficiently regular; these results are outlined in Section 6.3.

3. In fact, our method for proving the rate (6.2) is more general than the case of PDEs (see Figure 6.1 for the road map of the proof technique). The proof can be viewed as a recipe for convergence analysis of solutions to nonlinear functional equations of the form $\mathcal{P}(u) = f$ where u, f belong to sufficiently regular function spaces and $\mathcal{P}$ is invertible (at least locally). Then the Lipschitz stability of $\mathcal{P}$ and $\mathcal{P}^{-1}$ plus RKHS interpolation bounds on f yield convergence rates for u. Results at this level of generality are presented in (Schaback, 2016) for linear maps $\mathcal{P}$ which are then extended to nonlinear problems in (Böhmer and Schaback, 2013; Böhmer and Schaback, 2020). In these works, the stability of the discretization method is furthermore assumed. In the GP methodology, this property is guaranteed, due to the minimal RKHS norm and optimal recovery property of the solution. This has been pointed out in (Schaback, 2016, Sec. 10) for linear PDEs. Our theory can be seen as a generalization of the result in (Schaback, 2016) to the nonlinear case.

4. We present a suite of numerical experiments that elucidate and extend our theoretical analysis in item 2. We present an example of a nonlinear elliptic PDE with a prescribed solution of varying regularity in various dimensions. We then explore the interplay between regularity and dimensionality as well as the rate in (6.2). We further verify our result for a one dimensional parametric PDE by varying p, the dimension of the parameter space Θ. Because of this trade-off between regularity and dimensionality, showing that a numerical method remains accurate for a high dimensional PDE may not be an indication that it is breaking the curse of dimensionality but simply an indication that the problem being solved is very regular; see our experiments in Section 6.4 and in particular Section 6.4.

Literature Review

Below we present a brief review of the literature relevant to the current work.

Kernel and Gaussian Process Solvers for PDEs

As mentioned earlier our algorithmic and theoretical developments are focused on the kernel method introduced in (Yifan Chen, Hosseini, et al., 2021b) and extending

$$
\begin{aligned}
\text{Solution operator is stable} &\longrightarrow \|u^\dagger - u_N\|_1 \leq C\|\mathcal{P}(u^\dagger) - \mathcal{P}(u_N)\|_2 \\
\text{Sampling error estimates} &\longrightarrow \|\mathcal{P}(u^\dagger) - \mathcal{P}(u_N)\|_2 \leq Ch^\gamma \|\mathcal{P}(u^\dagger) - \mathcal{P}(u_N)\|_3 \\
\text{Forward operator is stable} &\longrightarrow \|\mathcal{P}(u^\dagger) - \mathcal{P}(u_N)\|_3 \leq C\|u^\dagger - u_N\|_4 \\
\text{Continuous embedding of RKHS} &\longrightarrow \|u^\dagger - u_N\|_4 \leq C\|u^\dagger - u_N\|_{\mathcal{U}} \\
\|u_N\|_{\mathcal{U}} \leq \|u^\dagger\|_{\mathcal{U}} &\longrightarrow \|u^\dagger - u_N\|_{\mathcal{U}} \leq 2\|u^\dagger\|_{\mathcal{U}}
\end{aligned}
$$

$$
\longrightarrow \|u^\dagger - u_N\|_1 \leq Ch^\gamma \|u^\dagger\|_{\mathcal{U}}
$$

Figure 6.1: A summary of the main steps in our proof of convergence rates outlined in Theorems 6.3.1, 6.3.2, 6.3.4 and 6.3.6. The 1–4 norms denote arbitrary norms on appropriate Banach spaces while the $\|\cdot\|_{\mathcal{U}}$-norm can be chosen as an RKHS norm or another desired norm with respect to which the numerical algorithm is stable.

that approach to parametric PDEs. Further extensions and applications of the aforementioned framework can also be found in (Mou, Yang, and Zhou, 2022; R. Meng and Yang, 2022; Long, Mrvaljevic, et al., 2022; Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024). When applied to linear PDEs our kernel method coincides with the so-called symmetric collocation method (Schaback and Wendland, 2006, Sec. 14) and is closely associated with radial basis function (RBF) PDE solvers (Fornberg and Flyer, 2015; Fasshauer, 1999; Franke and Schaback, 1998b). Various error analyses for RBF collocation methods can be found in (Franke and Schaback, 1998a; Franke and Schaback, 1998b). In particular, the article (Giesl and Wendland, 2007) is the closest to our work and their rates coincide with ours in the linear PDE setting. The articles (Ling, Opfer, and Schaback, 2006; Schaback, 2007; Ling and Schaback, 2008; Cheung, Ling, and Schaback, 2018) present similar bounds for the so-called Kansa method (Kansa, 1990a; Kansa, 1990b), a non-symmetric RBF collocation PDE solver. Finally, (Schaback, 2016) presents an abstract set of convergence rates for RBF interpolation of "well-posed" linear maps between regular function spaces that includes RBF PDE solvers as a special case. All of the aforementioned analyses consider linear PDEs and some generalizations to nonlinear problems are studied in (Böhmer and Schaback, 2013; Böhmer and Schaback, 2020).

The deep connection of Kernels and RKHSs to the theory of GPs (Bogachev, 1998; Larkin, 1972; Williams and Rasmussen, 2006; Vaart and Zanten, 2008) suggests that kernel PDE solvers can be viewed from lens of probability theory as a conditioning problem for GPs. While not as extensively developed as the kernel solvers mentioned earlier, this direction has been explored for the solution of linear PDEs as well as

nonlinear ODEs (O. A. Chkrebtii et al., 2016; Cockayne, C. J. Oates, et al., 2019; Houman Owhadi, 2015b; Särkkä, 2011; Swiler et al., 2020) and recent works have extended this idea to some nonlinear and time-dependent PDEs (Cockayne, C. Oates, et al., 2017; Raissi, Perdikaris, and George Em Karniadakis, 2018; J. Wang et al., 2021). The GP interpretation is attractive due to the ability to provide rigorous uncertainty estimates along with the solution to the PDE. The idea here is that the uncertainties can serve as a posterior or a priori error indicators for the PDE solver. Some ideas related to this direction were discussed in (Yifan Chen, Hosseini, et al., 2021b; Cockayne, C. Oates, et al., 2017). A fully probabilistic GP interpretation of our kernel framework for linear PDEs can be found in (Houman Owhadi, 2015b; Cockayne, C. Oates, et al., 2017; Houman Owhadi and Clint Scovel, 2019b) but the case of nonlinear PDEs remains partially investigated (Yifan Chen, Hosseini, et al., 2021b; R. Meng and Yang, 2022; Mou, Yang, and Zhou, 2022; Long, Z. Wang, et al., 2022). Moreover we note that in the GP framework, hierarchical Bayes learning can be used to select kernels to get better convergence rates (Yifan Chen, Houman Owhadi, and A. Stuart, 2021a; Wilson et al., 2016; Houman Owhadi and Yoo, 2019b; Darcy et al., 2023).

Parametric and High-dimensional PDEs

Parametric PDEs are ubiquitous is physical sciences and engineering and in particular in the context of uncertainty quantification (UQ) and solution of stochastic PDEs (SPDEs) (Cialenco, Fasshauer, and Ye, 2012; Ghanem and Spanos, 2003; Le Maitre and Knio, 2010; Ye, 2013). A vast literature exists on the subject, connecting it to reduced basis models (Almroth, Stern, and Brogan, 1978; Noor and J. M. Peters, 1980), emulation of computer codes (Kennedy and O'Hagan, 2001), reduced order models (Lucia, Beran, and Silva, 2004), and numerical homogenization (Houman Owhadi and Clint Scovel, 2019b); for settings that most closely resemble our problems we refer the reader to (Ghanem and Spanos, 2003; Xiu, 2010; Cohen and DeVore, 2015) for a general overview. Broadly speaking, the dominant approaches for approximation of high-dimensional and parametric solution maps include polynomial/Taylor approximation methods (Beck et al., 2012; Chkifa, Cohen, DeVore, et al., 2012; Chkifa, Cohen, and Schwab, 2014; Nobile, Raúl Tempone, and Webster, 2008; Nobile, Raul Tempone, and Webster, 2008); Galerkin methods (Gunzburger, Webster, and G. Zhang, 2014; Cohen, DeVore, and Schwab, 2010); reduced basis methods (J. S. Hesthaven, Rozza, Stamm, et al., 2016); and more recently ANN operator learning techniques such as (Li, Kovachki, Azizzadenesheli, Bhattacharya,

et al., 2020; L. Lu, Jin, et al., 2021). In comparison to the aforementioned works we propose to directly approximate the solution of the parametric PDE as a function on the tensor product space of the physical and parameter domains in a similar spirit as (Kempf, Wendland, and Rieger, 2019). The recent article (Batlle, Darcy, et al., 2023) also presents a kernel based operator learning approach to various PDE problems including parametric PDEs.

The Curse of Dimensionality

Although the trade-off between regularity and accuracy is well understood in numerical approximation/integration, where it has led to the development of the Kolmogorov N-width and stress tests for finite-element methods (Pinkus, 2012; Melenk, 2000; Babuška and Osborn, 2000), its impact is oftentimes overlooked when communicating the convergence of Machine Learning and Deep Learning methods for high-dimensional PDEs. In particular, since artificial neural networks (ANNs) can be interpreted as kernel methods (Neal, 1996; J. Lee et al., 2017; Houman Owhadi, 2023b) with data-dependent parameterized kernels, our results raise the further question of understanding whether the (empirically observed) convergence of ANN-based methods for high-dimensional PDEs is an indication of the absence of the curse (i.e., the regularity of the solution in selected numerical experiments is high) or the breaking of that curse. In particular, empirically observing numerical accuracy for an algorithm and particular solutions is insufficient to prove that the curse of dimensionality is broken, and one must also show that the underlying problem and those solutions are not too regular. We emphasize that the curse of dimensionality referred to here is the one associated with the worsening of the accuracy of a numerical approximation algorithm as a function of the dimension of the domain of the PDE as opposed to the impact of the curse on the number of degrees of freedom in the implementation level (e.g., finite difference methods suffer from that second curse but ANN/kernel based methods do not).

The Potential Value of Kernel/GP Methods

The proposed work aims to further develop Gaussian Process (GP) and kernel methods for solving PDEs. We are motivated to do so because GP methods have the potential to offer the best of both worlds by combining the profound theory underlying traditional methods (and in particular finite element methods) with the ease of implementation of emerging Deep Learning (DL) methods. They also come equipped with automatic uncertainty quantification (UQ) capabilities, not readily

available in either traditional or deep-learning based methods. And finally they provide easily implementable meshless methods that can be used to benchmark other machine learning based algorithms such as PINNs (Raissi, Perdikaris, and George E Karniadakis, 2019).

Compared to traditional methods (such as finite element methods (FEM), finite volume methods (FVM), finite difference methods (FDM), spectral methods, etc), GP methods generalize meshless, RBF, optimal recovery methods and are flexible and applicable in high dimensions. Compared to DL methods that use an expressive neural network representation, GPs offer transparent methods that are easy to reproduce and analyze. Furthermore the natural probabilistic interpretation of GPs enables convenient UQ and also facilitates the process of scientific discovery itself (Houman Owhadi and Clint Scovel, 2019b); FEM and DL methods do not interface so cleanly with UQ. Moreover, with hierarchical kernel learning (Yifan Chen, Houman Owhadi, and A. Stuart, 2021a; Wilson et al., 2016; Houman Owhadi and Yoo, 2019b; Darcy et al., 2023), GP methods can also be made highly expressive. In fact, as shown in the table below, GP methods can offer many advantages over traditional and DL methods. In the context of PDEs these advantages include greater flexibility, applicability in high dimensions, provable guarantees, near-linear complexity computation, Occam's razor principle in the design of statistical models, mathematical transparency and interpretability, and ease of reproducibility; see Table 6.1. Although software support for GPs is currently not as advanced as that for DL and traditional methods, GPs are still easy to program and can be seamlessly integrated into an engineering pipeline. Table 6.1 should be interpreted in this light: as an argument for further deployment and development of software infrastructure for GP-PDE based methods.

Method	Ease of implementation in high-dimensions	Provable guarantees	Near linear complexity	Occam's razor	Transparent	Ease of reproducibility	Built-in UQ	Software support
Trad.	✗	✓	✓	✓	✓	✓	✗	✓
Kernel	✓	✓	✓	✓	✓	✓	✓	Limited
ANN	✓	Limited	✗	✗	✗	Limited	✗	✓

Table 6.1: A qualitative comparison of the properties of traditional PDE solvers (such as FEM, FVM, FDM, spectral methods, etc.) against kernel methods and ANNs.

Given their long training times, ANN-based methods may not be competitive with FEM in low dimensions (Grossmann et al., 2023). In contrast, GP-based methods

can achieve near-linear complexity when combined with fast algorithms for kernel methods such as the sparse Cholesky factorization (Florian Schäfer, Katzfuss, and Houman Owhadi, 2021; Florian Schäfer, Timothy John Sullivan, and Houman Owhadi, 2021b; Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024). In some applications, these algorithms can be competitive (both in terms of complexity and accuracy) even when compared to highly optimized algebraic multigrid solvers such as AMGCL and Trilinos (J. Chen et al., 2021). GP methods are naturally amenable to analysis and come with simple provable guarantees, while ANN-based methods involve complicated optimizations and many heuristics, which can make them hard to understand. GP methods fit Occam's razor, offering a clarity of purpose in their structure. We can understand why and when they work, which is of scientific importance (Feynman, 1998). Therefore, it is of potential value to benchmark deep learning methods against kernel-based methods to ensure that the deep part of a DL method serves a significant purpose beyond adding complexity.

Outline of the Article

The rest of the article is organized as follows: Ww present a brief overview of our GP and kernel approach for solving nonlinear and parametric PDEs in Section 6.2; our error analysis is outlined in Section 6.3, followed by numerical experiments in Section 6.4 and conclusions in Section 6.5. Auxiliary results are collected in Sections 6.6 to 6.8.

6.2 Kernel Methods for Parametric PDEs

In this section we extend the kernel methodology of (Yifan Chen, Hosseini, et al., 2021b) to the case of parametric PDEs as outlined in Section 6.2. Some numerical strategies and ideas for improving the efficiency of the solver are discussed in Section 6.2.

Solving Parametric PDEs

Let us consider bounded connected domains $\Omega \in \mathbb{R}^d$ with a Lipschitz boundary for $d \geq 1$ and $\Theta \subset \mathbb{R}^p$ for $p \geq 1$. We consider nonlinear and parametric PDEs of the form

$$\begin{cases} \mathcal{P}(u^\star)(\boldsymbol{x};\boldsymbol{\theta}) = f(\boldsymbol{x};\boldsymbol{\theta}), & (\boldsymbol{x},\boldsymbol{\theta}) \in \Omega \times \Theta, \\ \mathcal{B}(u^\star)(\boldsymbol{x};\boldsymbol{\theta}) = g(\boldsymbol{x};\boldsymbol{\theta}), & (\boldsymbol{x},\boldsymbol{\theta}) \in \partial\Omega \times \Theta, \end{cases} \tag{6.3}$$

where $\mathcal{P}, \mathcal{B}$ are nonlinear differential operators in the interior and boundary of Ω, $\boldsymbol{\theta}$ is a parameter, and f, g are the PDE source and boundary data. For now we

assume that the above PDE is well-posed and has a unique solution $u^\star(\boldsymbol{x};\boldsymbol{\theta})$ which is assumed to exist in the strong sense over $\overline{\Omega}$ and for all values of $\boldsymbol{\theta} \in \Theta$. In (Yifan Chen, Hosseini, et al., 2021b) the authors introduced a GP/kernel method for solving nonlinear PDEs of the form (6.3) without parametric dependence. Here we extend that approach to the parametric case.

Let $\Upsilon := \Omega \times \Theta$ and $\partial\Upsilon := \partial\Omega \times \Theta$ and write $\mathbf{s} := (\boldsymbol{x}, \boldsymbol{\theta})$. Choose $M \geq 1$ collocation points $\{\mathbf{s}_m\}_{m=1}^{M} \in \overline{\Upsilon}$ such that $\{\mathbf{s}_m\}_{m=1}^{M_\Omega} \in \Upsilon$ and $\{\mathbf{s}_m\}_{m=M_\Omega+1}^{M} \in \partial\Upsilon$[1] and consider a kernel $K : \overline{\Upsilon} \times \overline{\Upsilon} \to \mathbb{R}$ with its corresponding RKHS denoted by $\mathcal{U}$ and norm $\|\cdot\|_{\mathcal{U}}$. We then propose to approximate $u^\star(\mathbf{s})$ by solving the optimization problem

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}} & \|u\|_{\mathcal{U}} \\ \text{st} & \mathcal{P}(u)(\mathbf{s}_m) = f(\mathbf{s}_m), \quad m = 1, \dots, M_\Omega, \\ & \mathcal{B}(u)(\mathbf{s}_m) = g(\mathbf{s}_m), \quad m = M_\Omega + 1, \dots, M. \end{cases} \tag{6.4}$$

Observe that our approach above approximates the solution u as a function defined on the set product set Υ which is different from previous works (Beck et al., 2012; Chkifa, Cohen, DeVore, et al., 2012; Chkifa, Cohen, and Schwab, 2014; Nobile, Raúl Tempone, and Webster, 2008; Nobile, Raul Tempone, and Webster, 2008) where the solution map $\boldsymbol{\theta} \to u^\star(\cdot;\boldsymbol{\theta})$, as a mapping from Θ to an appropriate function space, is characterized and approximated. The latter approach requires different discretization methods for the $\boldsymbol{\theta}$ parameter and the functions $u^\star(\cdot;\boldsymbol{\theta})$ while our approach leads to a meshless collocation method on the product space which is desirable and convenient at the level of implementation, following (Yifan Chen, Hosseini, et al., 2021b, Sec. 3.1) (see also (Giesl and Wendland, 2007)).

We make the following assumption on the differential operators $\mathcal{P}, \mathcal{B}$.

Assumption 1. There exist bounded and linear operators $L_1, \dots, L_{Q_\Omega} \in \mathcal{L}\big(\mathcal{U}; C(\Upsilon)\big)$ and $L_{Q_\Omega+1}, \dots, L_Q \in \mathcal{L}\big(\mathcal{U}; C(\partial\Upsilon)\big)$ for some $1 \leq Q_\Omega < Q$ together with maps $P : \mathbb{R}^{Q_\Omega} \to \mathbb{R}$ and $B : \mathbb{R}^{Q-Q_\Omega} \to \mathbb{R}$, which may be nonlinear, so that $\mathcal{P}, \mathcal{B}$ can be written as

$$\begin{aligned} \mathcal{P}(u)(\mathbf{s}) &= P\Big(L_1(u)(\mathbf{s}), \dots, L_{Q_\Omega}(u)(\mathbf{s})\Big) \qquad \forall \mathbf{s} \in \Upsilon, \\ \mathcal{B}(u)(\mathbf{s}) &= B\Big(L_{Q_\Omega+1}(u)(\mathbf{s}), \dots, L_Q(u)(\mathbf{s})\Big) \quad \forall \mathbf{s} \in \partial\Upsilon. \end{aligned} \tag{6.5}$$

We briefly introduce a running example of a parametric PDE for which the above assumptions can be verified easily.

[1] Note that we do not specifically ask for collocation points on $\partial\Omega \times \partial\Theta$ since we may not have boundary data on the $\boldsymbol{\theta}$ parameter.

Example 6.2.1 (Nonlinear Darcy flow)**.** *Consider the nonlinear Darcy flow PDE*

$$\begin{cases} -div_{\boldsymbol{x}}\big(\exp(a(\boldsymbol{x},\boldsymbol{\theta}))\nabla u\big)(\boldsymbol{x}) + \tau(u(\boldsymbol{x})) = 1, & \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) = 0, & \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{6.6}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded domain with a Lipschitz boundary and $\tau : \mathbb{R} \to \mathbb{R}$ is a continuous and nonlinear map. We assume that the permeability field is parameterized as

$$a(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{j=1}^{p} \theta_j \psi_j(\boldsymbol{x}), \tag{6.7}$$

where $\theta_j \in (0,1)$ so that $\Theta = (0,1)^p$ and $\psi_j \in C(\overline{\Omega})$. Substituting into the PDE and expanding the differential operator we can rewrite our nonlinear PDE as

$$\begin{cases} -\exp\left(\sum_{j=1}^{p} \theta_j \psi_j(\boldsymbol{x})\right) \sum_{j=1}^{p} \theta_j \nabla_{\boldsymbol{x}} \psi_j(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} u(\boldsymbol{x};\boldsymbol{\theta}) \\ \qquad - \exp\left(\sum_{j=1}^{p} \theta_j \psi_j(\boldsymbol{x})\right) \Delta_{\boldsymbol{x}} u(\boldsymbol{x};\boldsymbol{\theta}) + \tau(u(\boldsymbol{x};\boldsymbol{\theta})) = 1, \quad (\boldsymbol{x},\boldsymbol{\theta}) \in \Omega \times \Theta, \\ u(\boldsymbol{x};\boldsymbol{\theta}) = 0, \quad (\boldsymbol{x},\boldsymbol{\theta}) \in \partial\Omega \times \Theta, \end{cases}$$

where we used subscripts on the differential operators to highlight that derivatives are computed for the $\boldsymbol{x}$ variable only and not $\boldsymbol{\theta}$. We also did not use the compact notation $\mathbf{s} \equiv (\boldsymbol{x},\boldsymbol{\theta})$ since it is more helpful to be able to distinguish between the $\boldsymbol{x}$ and $\boldsymbol{\theta}$ variables in this example. We can directly verify Assumption 1 with the bounded and linear operators

$$\begin{aligned} &L_1 : u(\boldsymbol{x};\boldsymbol{\theta}) \mapsto u(\boldsymbol{x};\boldsymbol{\theta}), \\ &L_2 : u(\boldsymbol{x};\boldsymbol{\theta}) \mapsto \exp\left(\sum_{j=1}^{p} \theta_j \psi_j(\boldsymbol{x})\right) \sum_{j=1}^{p} \theta_j \nabla_{\boldsymbol{x}} \psi_j(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} u(\boldsymbol{x};\boldsymbol{\theta}), \\ &L_3 : u(\boldsymbol{x};\boldsymbol{\theta}) \mapsto \exp\left(\sum_{j=1}^{p} \theta_j \psi_j(\boldsymbol{x})\right) \Delta_{\boldsymbol{x}} u(\boldsymbol{x};\boldsymbol{\theta}) \\ &L_4 : u(\boldsymbol{x};\boldsymbol{\theta}) \mapsto u(\boldsymbol{x};\boldsymbol{\theta}). \end{aligned}$$

Note that operators L_1, L_4 are the same here since the point values of u appear in both the interior and boundary conditions. Thus we have $Q_\Omega = 3$ and $Q = 4$ and the maps

$$P(t_1,t_2,t_3) = -t_2 - t_3 + \tau(t_1), \qquad B(t_1) = t_1.$$

If $\mathcal{U}$ is sufficiently regular and Assumption 1 holds, then we can define the functionals $\phi_m^q \in \mathcal{U}^*$ for $1 \le q \le Q$ as

$$\phi_m^q := \delta_{(\mathbf{s}_m)} \circ L_q, \quad \text{where} \quad \begin{cases} 1 \le m \le M_\Omega & \text{if} \quad 1 \le q \le Q_\Omega \\ M_\Omega + 1 \le m \le M & \text{if} \quad Q_{\Omega+1} \le q \le Q. \end{cases} \tag{6.8}$$

In what follows we write $[\phi, u]$ to denote the duality pairing between $\mathcal{U}$ and $\mathcal{U}^*$ and further use the shorthand notation $\boldsymbol{\phi}^{(q)}$ to denote the vector of dual elements ϕ_m^q for a fixed index q. Note that $\boldsymbol{\phi}^{(q)} \in (\mathcal{U}^*)^{\otimes M_\Omega}$ if $q \le Q_\Omega$ but $\boldsymbol{\phi}^{(q)} \in (\mathcal{U}^*)^{\otimes (M - M_\Omega)}$ if $q > Q_\Omega$ in order to accommodate different differential operators defining the PDE and the boundary conditions. We further write $N = M_\Omega Q_\Omega + (M - M_\Omega)(Q - Q_\Omega)$ and define

$$\boldsymbol{\phi} = \left(\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(Q)}\right) \in (\mathcal{U}^*)^{\otimes N}.$$

Henceforth we write ϕ_n for $n = 1, \ldots, N$ to denote the entries of the vector $\boldsymbol{\phi}$ and write $[\boldsymbol{\phi}, u] = ([\phi_1, u], \ldots, [\phi_N, u]) \in \mathbb{R}^N$. With this notation we rewrite problem (6.4) as

$$\begin{cases} \text{minimize} & \|u\|_{\mathcal{U}} \\ \text{st} & F([\boldsymbol{\phi}, u]) = \boldsymbol{y}, \end{cases}$$

where the data vector $\boldsymbol{y} \in \mathbb{R}^M$ has entries

$$y_m := \begin{cases} f(\mathbf{s}_m), & \text{if} \quad 1 \le m \le M_\Omega, \\ g(\mathbf{s}_m), & \text{if} \quad M_\Omega + 1 \le m \le M, \end{cases}$$

and $F : \mathbb{R}^N \to \mathbb{R}^M$ is a nonlinear map whose output components are defined as

$$\left(F([\boldsymbol{\phi}, u])\right)_m := \begin{cases} P\left([\phi_m^1, u], \ldots, [\phi_m^{Q_\Omega}, u]\right) & \text{if } 1 \le m \le M_\Omega, \\ B\left([\phi_m^{Q_\Omega+1}, u], \ldots, [\phi_m^{Q}, u]\right) & \text{if } M_\Omega + 1 \le m \le M. \end{cases} \tag{6.9}$$

Further define the kernel vector field

$$K(\cdot, \boldsymbol{\phi}) : \overline{\Upsilon} \to \mathcal{U}^N, \qquad K(\mathbf{s}, \boldsymbol{\phi})_k := [\phi_k, K(\mathbf{s}, \cdot)] \tag{6.10}$$

and the kernel matrix

$$K(\boldsymbol{\phi}, \boldsymbol{\phi}) \in \mathbb{R}^{N \times N}, \qquad K(\boldsymbol{\phi}, \boldsymbol{\phi})_{nk} := [\phi_n, K(\cdot, \boldsymbol{\phi})_k]. \tag{6.11}$$

We can then characterize the minimizers of (6.4) via the following representer theorem which is a direct consequence of (Yifan Chen, Hosseini, et al., 2021b, Prop. 2.3):

Proposition 6.2.2. *Suppose Assumption 1 holds and $K(\boldsymbol{\phi}, \boldsymbol{\phi})$ is invertible. Then a function $u^\dagger : \overline{\Upsilon} \to \mathbb{R}$ is a minimizer of* (6.4) *if and only if*

$$u^\dagger(\mathbf{s}) = K(\mathbf{s}, \boldsymbol{\phi}) K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} z^\dagger, \tag{6.12}$$

where $z^\dagger \in \mathbb{R}^N$ solves

$$\begin{cases} \underset{z \in \mathbb{R}^N}{\text{minimize}} & z^T K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} z, \\ \text{st} & F(z) = \boldsymbol{y}, \end{cases} \tag{6.13}$$

with the nonlinear map F defined in (6.9).

This result allows us to reduce the infinite-dimensional optimization problem (6.4) to a finite-dimensional optimization problem without incurring any approximation errors; it is an instance of the well-known family of representer theorems (Scholkopf and Alexander J Smola, 2018, Sec. 4.2). Thus, to find an approximation to $u^\star$ we simply need to solve (6.13) and apply the formula (6.12); algorithms for this task are discussed next.

Numerical Strategies

We now summarize various numerical strategies for solution of (6.13). These strategies are naturally applicable to non-parametric PDEs as they can be viewed as a special case of (6.3) with a fixed parameter. In Section 6.2 we summarize a Gauss-Newton algorithm that was introduced in (Yifan Chen, Hosseini, et al., 2021b) followed by a new and, often, more efficient strategy that linearizes the PDE first before formulating the optimization problem in Section 6.2.

Gauss-Newton

To solve the optimization problem (6.13), a Gauss-Newton algorithm was proposed in (Yifan Chen, Hosseini, et al., 2021b) which we recall briefly. The equality constraints can be dealt with either by elimination or relaxation. Suppose that there exists a map $\overline{F} : \mathbb{R}^{N-M} \times \mathbb{R}^M \to \mathbb{R}^N$ so that

$$F(z) = \boldsymbol{y} \quad \text{if and only if} \quad z = \overline{F}(\boldsymbol{w}, \boldsymbol{y}), \quad \text{for a unique } \boldsymbol{w} \in \mathbb{R}^{N-M}.$$

Then, we rewrite (6.13) as the unconstrained optimization problem

$$\underset{\boldsymbol{w} \in \mathbb{R}^{N-M}}{\text{minimize}} \quad \overline{F}(\boldsymbol{w}, \boldsymbol{y})^T K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} \overline{F}(\boldsymbol{w}, \boldsymbol{y}). \tag{6.14}$$

Then a minimizer $\boldsymbol{w}^\dagger$ of (6.14) can be approximated with a sequence of elements $\boldsymbol{w}^\ell$ defined iteratively via $\boldsymbol{w}^{\ell+1} = \boldsymbol{w}^\ell + \alpha^\ell \delta \boldsymbol{w}^\ell$, where $\alpha^\ell > 0$ is an appropriate step size while $\delta \boldsymbol{w}^\ell$ is the minimizer of the optimization problem

$$\underset{\delta \boldsymbol{w} \in \mathbb{R}^{N-M}}{\text{minimize}} \quad \left(\overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) + \nabla_{\boldsymbol{w}} \overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) \delta \boldsymbol{w}\right)^T K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} \left(\overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) + \nabla_{\boldsymbol{w}} \overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) \delta \boldsymbol{w}\right).$$

Alternatively, if the map $\overline{F}$ does not exist or is hard to compute, i.e., eliminating the constraints is not feasible, then we consider the relaxed problem

$$\underset{z \in \mathbb{R}^N}{\text{minimize}} \quad \frac{1}{2} z^T K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} z + \frac{1}{2\beta^2} |F(z) - \boldsymbol{y}|^2,$$

for a sufficiently small parameter $\beta > 0$. Here $\|\cdot\|$ is the L^2 norm of the vector. A minimizer $z_\beta^\dagger$ of the above problem can be approximated with a sequence z^ℓ where $z^{\ell+1} = z^\ell + \alpha^\ell \delta z^\ell$ where δz^ℓ is the minimizer of

$$\underset{\delta z \in \mathbb{R}^N}{\text{minimize}} \quad \delta z^T K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} z^\ell + \frac{1}{2\beta^2} |F(z^\ell) + \nabla F(z^\ell) \delta z - \boldsymbol{y}|^2.$$

We summarize the proposed Gauss-Newton algorithm for solution of parametric PDEs in Algorithm 9.

Algorithm 9 Kernel Methods for Parametric PDEs using Gauss-Newton (Section 6.2)

Input: PDE of the form Equation (6.3) defined on $\Upsilon = \Omega \times \Theta$ with boundary condition on $\partial\Upsilon = \partial\Omega \times \Theta$, $M \geq 1$ collocation points in $\overline{\Upsilon}$, and kernel $K : \overline{\Upsilon} \times \overline{\Upsilon} \to \mathbb{R}$

Output: Approximation $u^\dagger(s)$ to exact solution $u^*(s)$

$N \leftarrow M_\Omega Q_\Omega + (M - M_\Omega)(Q - Q_\Omega)$
for $i = 1$ to N {Build kernel matrix $K(\boldsymbol{\phi}, \boldsymbol{\phi})$} **do**
  for $j = 1$ to N **do**
    $K(\boldsymbol{\phi}, \boldsymbol{\phi})_{i,j} \leftarrow [\phi_i, K(\cdot, \boldsymbol{\phi})_j]$
  end for
end for
while not converged **do**
  $\delta \boldsymbol{w}^\ell \leftarrow \arg\min_{\delta \boldsymbol{w}} \left(\overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) + \nabla_{\boldsymbol{w}} \overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) \delta \boldsymbol{w}\right)^T$
    $\cdot K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} \left(\overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) + \nabla_{\boldsymbol{w}} \overline{F}(\boldsymbol{w}^\ell, \boldsymbol{y}) \delta \boldsymbol{w}\right)$
  $\boldsymbol{w}^{\ell+1} \leftarrow \boldsymbol{w}^\ell + \alpha^\ell \delta \boldsymbol{w}^\ell$
end while
$z^\dagger \leftarrow \overline{F}(\boldsymbol{w}, \boldsymbol{y})$
$u^\dagger(\mathbf{s}) \leftarrow K(\mathbf{s}, \boldsymbol{\phi}) K(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1} z^\dagger$ {Apply representer formula}

Linearize then Optimize

The Gauss-Newton approach of Section 6.2 is applicable to wide families of nonlinear PDEs. The primary computational bottleneck of that approach is the construction and factorization of the kernel matrix $K(\boldsymbol{\phi}, \boldsymbol{\phi})$ which for some PDEs can be prohibitively large. To get around this difficulty we propose an alternative approach to approximating the solution of (6.4) by first linearizing the PDE operators before applying Proposition 6.2.2. The resulting approach is more intrusive in comparison to the Gauss-Newton method as it requires explicit calculations involving the PDE but often leads to smaller kernel matrices and better performance. This method can also be viewed as applying the methodology of (Yifan Chen, Hosseini, et al., 2021b; Giesl and Wendland, 2007) to discretize successive linearizations of the PDE.

Let $u^\dagger$ denote the minimizer of (6.4) as before. Assuming that the operators $\mathcal{P}$ and $\mathcal{B}$ are Fréchet differentiable we then approximate $u^\dagger$ with a sequence of elements u^ℓ obtained by solving the problem

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}} & \|u\|_{\mathcal{U}} \\ \quad \text{st} & \left(\mathcal{P}(u^{\ell-1}) + \mathcal{P}'(u^{\ell-1})(u-u^{\ell-1})\right)|_{\mathbf{s}_m} = f(\mathbf{s}_m), \quad m = 1,\dots,M_\Omega, \\ & \left(\mathcal{B}(u^{\ell-1}) + \mathcal{B}'(u^{\ell-1})(u-u^{\ell-1})\right)|_{\mathbf{s}_m} = g(\mathbf{s}_m), \quad m = M_\Omega+1,\dots,M, \end{cases} \tag{6.15}$$

where $\mathcal{P}'$ and $\mathcal{B}'$ are the Fréchet derivatives of $\mathcal{P}$ and $\mathcal{B}$.

Let us further suppose that Assumption 1 holds. Observing that the constraints in (6.15) are linear in u we obtain an explicit formula for u^ℓ by (Yifan Chen, Hosseini, et al., 2021b, Prop. 2.2):

$$u^\ell(\mathbf{s}) = K(\mathbf{s}, \widetilde{\boldsymbol{\phi}}^{\ell-1}) K(\widetilde{\boldsymbol{\phi}}^{\ell-1}, \widetilde{\boldsymbol{\phi}}^{\ell-1})^{-1} \boldsymbol{z}^{\ell-1}, \tag{6.16}$$

where $\boldsymbol{z}^{\ell-1} = (z_1^{\ell-1}, \dots, z_M^{\ell-1})^T$ has entries

$$z_m^{\ell-1} = \begin{cases} \left(f - \mathcal{P}(u^{\ell-1}) + \mathcal{P}'(u^{\ell-1})u^{\ell-1}\right)|_{\mathbf{s}_m}, & \text{if } 1 \le m \le M_\Omega, \\ \left(f - \mathcal{P}(u^{\ell-1}) + \mathcal{B}'(u^{\ell-1})u^{\ell-1}\right)|_{\mathbf{s}_m}, & \text{if } M_\Omega + 1 \le m \le M. \end{cases} \tag{6.17}$$

The vectors $\widetilde{\boldsymbol{\phi}}^{\ell-1} \in (\mathcal{U}^*)^{\otimes M}$ are obtained by concatenating the dual elements

$$\widetilde{\phi}_m^{\ell-1} := \begin{cases} \delta_{(\mathbf{s}_m)} \circ \mathcal{P}'(u^{\ell-1}), & \text{if } 1 \le m \le M_\Omega, \\ \delta_{(\mathbf{s}_m)} \circ \mathcal{B}'(u^{\ell-1}), & \text{if } M_\Omega + 1 \le m \le M. \end{cases} \tag{6.18}$$

We note that the above scheme implicitly assumes that $\mathcal{U}$ is sufficiently regular so that the derivatives $\mathcal{P}'(u^{\ell-1})$ and $\mathcal{B}'(u^{\ell-1})$ can be regarded as linear operators mapping

$\mathcal{U}$ to $C(\overline{\Upsilon})$ where pointwise evaluation is well-defined. The most important feature of the linearize-then-optimize approach is that the kernel matrices $K(\widetilde{\boldsymbol{\phi}}^{\ell-1}, \widetilde{\boldsymbol{\phi}}^{\ell-1})$ are of size $M \times M$ while the kernel matrix $K(\boldsymbol{\phi}, \boldsymbol{\phi})$, used in the Gauss-Newton approach of Section 6.2, is of size $N \times N$. Note that $N/M \approx Q_\Omega$ and for instance, in Example 6.2.1, $Q_\Omega = 3$ so there is an approximately $3\times$ reduction in the kernel matrix size. Thus, the linearize-then-optimize approach requires the inversion of much smaller kernel matrices at each iteration but these matrices need to be updated successively since the $\widetilde{\boldsymbol{\phi}}^{\ell}$ depend on the previous solution $u^{\ell-1}$. In the case where Assumption 1 holds and P, B are differentiable, then the $\mathcal{P}'(u)$ and $\mathcal{B}'(u)$ operators can be written explicitly as

$$\mathcal{P}'(u^{\ell-1}) : u \mapsto \nabla P\left(L_1(u^{\ell-1}), \ldots, L_{Q_\Omega}(u^{\ell-1})\right)^T \begin{bmatrix} L_1(u) \\ \vdots \\ L_{Q_\Omega}(u) \end{bmatrix},$$

$$\mathcal{B}'(u^{\ell-1}) : u \mapsto \nabla B\left(L_{Q_\Omega+1}(u^{\ell-1}), \ldots, L_{Q}(u^{\ell-1})\right)^T \begin{bmatrix} L_{Q_\Omega+1}(u) \\ \vdots \\ L_{Q}(u) \end{bmatrix}.$$

We note that while the "linearize then optimize" approach can reduce the size of kernel matrices by a constant factor, which is significant in practice, the size still scales with the number of collocation points. For very large-scale problems requiring many collocation points, we can further employ fast algorithms for these kernel matrices; see for example (Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024).

Remark 9. The "linearize then optimize" approach performs linearization first at the continuous level, while the Gauss-Newton iteration linearizes at the discrete level after applying the representor theorem and transforming the optimization problem into an unconstrained form, either through elimination or relaxation. The "linearize then optimize" approach and the Gauss-Newton iteration are mathematically equivalent if the latter is implemented using elimination with a specific choice of $\overline{F}$. This equivalence is demonstrated in (Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024, Sec. 5.1) for a nonlinear elliptic example, where the algorithm is also shown to be equivalent to a sequential quadratic programming approach for solving (6.13). In general, these approaches may differ in how the nonlinear operators $\mathcal{P}, \mathcal{B}$, or the nonlinear map F are represented

Finally, we summarize our linearize-then-optimize approach as a pseudo-algorithm within Algorithm 10.

Algorithm 10 The linearize-then-optimize approach to parametric PDEs

Input: PDE of the form Equation (6.3) defined on $\Upsilon = \Omega \times \Theta$ with boundary condition on $\partial\Upsilon = \partial\Omega \times \Theta$, $M \geq 1$ collocation points in $\overline{\Upsilon}$, and kernel $K : \overline{\Upsilon} \times \overline{\Upsilon} \to \mathbb{R}$

Output: Approximation $u^\dagger(s)$ to exact solution $u^*(s)$

1: **while** not converged {Iteratively approximate $u^\dagger$} **do**

2: For $m = 1, \ldots, M$:

3: $$z_m^{\ell-1} \leftarrow \begin{cases} \left(f - \mathcal{P}(u^{\ell-1}) + \mathcal{P}'(u^{\ell-1})u^{\ell-1}\right)(\mathbf{s}_m), & \text{if } m \leq M_\Omega \\ \left(f - \mathcal{P}(u^{\ell-1}) + \mathcal{B}'(u^{\ell-1})u^{\ell-1}\right)(\mathbf{s}_m), & \text{if } m > M_\Omega \end{cases} \quad \{\text{Build } \boldsymbol{z}^{\ell-1}\}$$

4: For $m = 1, \ldots, M$:

5: $$\widetilde{\boldsymbol{\phi}}_m^{\ell-1} \leftarrow \begin{cases} \delta_{(\mathbf{s}_m)} \circ \mathcal{P}'(u^{\ell-1}), & \text{if } m \leq M_\Omega \\ \delta_{(\mathbf{s}_m)} \circ \mathcal{B}'(u^{\ell-1}), & \text{if } m > M_\Omega \end{cases} \{\text{Build } \widetilde{\boldsymbol{\phi}}^{\ell-1}\}$$

6: $u^\ell(\mathbf{s}) \leftarrow K(\mathbf{s}, \widetilde{\boldsymbol{\phi}}^{\ell-1})K(\widetilde{\boldsymbol{\phi}}^{\ell-1}, \widetilde{\boldsymbol{\phi}}^{\ell-1})^{-1}\boldsymbol{z}^{\ell-1}$

7: **end while**

6.3 Error Analyses

We now present our main theoretical results concerning convergence rates for the minimizers $u^\dagger$ of (6.4) to the respective true solutions $u^\star$. We start in Section 6.3 by articulating the abstract framework, main theorem, and proof. We then consider the simple setting of a nonlinear PDE in Section 6.3 where the RKHS $\mathcal{U}$ already satisfies the boundary conditions of the PDE to convey the main ideas of the proof in a simple setting. Non-trivial boundary conditions are then considered in Section 6.3 followed by the case of parametric PDEs in Section 6.3. Our proof technique is a generalization of the results of (Giesl and Wendland, 2007; Schaback, 2016) to the case of nonlinear and parametric PDEs that are Lipschitz stable and well-posed.

An Abstract Framework for Obtaining Convergence Rates

We present here an abstract theoretical result that allows us to obtain convergence rates for nonlinear operator equations. Our error analyses concerning the numerical solutions $u^\dagger$ and the true solution to the PDE $u^\star$ then follow as applications of this abstract result. Our main result here can also be viewed as a generalization of the results of (Schaback, 2016, Sec. 10), which focused on linear operators, to the nonlinear case.

Let us consider operator equations of the form

$$\mathcal{T}(v^\star) = w^\star, \tag{6.19}$$

where $v^\star$, $w^\star$ are elements of appropriate Banach spaces and $\mathcal{T}$ is a nonlinear map. In the setting of PDEs the map $\mathcal{T}$ is defined by the differential operator of the PDE, $v^\star$ coincides with the solution and $w^\star$ is the source/boundary data. Broadly speaking our goal is to approximate the solution $v^\star$ under assumptions on its regularity and the stability properties of the map $\mathcal{T}$. To this end, we present a general result that allows us to control the error of approximating $v^\star$ given an appropriate candidate $v^\dagger$. Henceforth we write $B_r(V)$ to denote the ball of radius r centered at zero in a Banach space V.

Theorem 6.3.1. *Consider abstract Banach spaces $(V_i, \|\cdot\|_i)_{i=1}^4$ as well as $(\mathcal{U}, \|\cdot\|_{\mathcal{U}})$. Suppose the following conditions are satisfied for any choice of $r > 0$ (all the appeared constants $C(r)$ are non-decreasing regarding r):*

(A1) For any pair $v, v' \in B_r(V_1)$ there exists a constant $C = C(r) > 0$ so that

$$\|v - v'\|_1 \leq C\|\mathcal{T}(v) - \mathcal{T}(v')\|_2. \tag{6.20}$$

(A2) For any pair $v, v' \in B_r(V_4)$ there exists a constant $C = C(r) > 0$ so that

$$\|\mathcal{T}(v) - \mathcal{T}(v')\|_3 \leq C\|v - v'\|_4. \tag{6.21}$$

(A3) For any $v \in V_4$, there exists a constant $C > 0$ so that

$$\|v\|_4 \leq C\|v\|_{\mathcal{U}}.$$

(A4) There exists a set $\widetilde{V} \subset V_2 \cap V_3$ and a constant $\varepsilon > 0$, so that for all $w, w' \in \widetilde{V}$ it holds that

$$\|w - w'\|_2 \leq \varepsilon\|w - w'\|_3. \tag{6.22}$$

Suppose problem (6.19) *is uniquely solvable with $v^\star \in \mathcal{U}$ and let $v^\dagger \in \mathcal{U}$ be any other function such that*

(A5) $\mathcal{T}(v^\star), \mathcal{T}(v^\dagger) \in \widetilde{V}$.

(A6) There exists a constant $C > 0$, independent of $v^\star$ and $v^\dagger$, so that

$$\|v^\dagger\|_{\mathcal{U}} \leq C\|v^\star\|_{\mathcal{U}}. \tag{6.23}$$

Then there exists a constant $C > 0$, depending only on $\|v^\star\|_{\mathcal{U}}$, such that

$$\|v^\dagger - v^\star\|_1 \leq C\varepsilon\|v^\star\|_{\mathcal{U}}.$$

Proof. By (A1) we have that

$$\|v^\dagger - v^\star\|_1 \leq C\|\mathcal{T}(v^\dagger) - \mathcal{T}(v^\star)\|_2. \tag{6.24}$$

Then (A4) and (A5) imply that $\|\mathcal{T}(v^\dagger) - \mathcal{T}(v^\star)\|_2 \leq C\varepsilon\|\mathcal{T}(v^\dagger) - \mathcal{T}(v^\star)\|_3$. By the triangle inequality we have $\|\mathcal{T}(v^\dagger) - \mathcal{T}(v^\star)\|_3 \leq \|\mathcal{T}(v^\dagger) - \mathcal{T}(0)\|_3 + \|\mathcal{T}(v^\star) - \mathcal{T}(0)\|_3$. Using (A2), (A3), and (A6) in that order, we get $\|\mathcal{T}(v^\dagger) - \mathcal{T}(0)\|_3 \leq C\|v^\dagger\|_4 \leq C\|v^\dagger\|_{\mathcal{U}} \leq C\|v^\star\|_{\mathcal{U}}$. Similarly, we have $\|\mathcal{T}(v^\star) - \mathcal{T}(0)\|_3 \leq C\|v^\star\|_{\mathcal{U}}$. Combining these bounds we obtain $\|\mathcal{T}(v^\dagger) - \mathcal{T}(v^\star)\|_3 \leq C\varepsilon\|v^\star\|_{\mathcal{U}}$ which yields the desired result due to (6.24). □

Let us provide some remarks regarding the assumptions of the theorem. In our PDE examples we often take the V_i spaces to be Sobolev spaces of appropriate smoothness while $\mathcal{U}$ is taken as an RKHS that is sufficiently smooth and so $v^\star \in \mathcal{U}$ amounts to an assumption on the regularity of the true solution to the problem. Conditions (A1) and (A2) amount to forward and inverse Lipschitz stability of the operator $\mathcal{T}$ while (A4) is often given by a sampling/Poincaré-type inequality for our numerical method. We treat the constant ε separately from the other constants in the theorem since in practice ε often coincides with some power of the resolution (fill-distance/meshnorm) of our numerical scheme, constituting the rate of convergence of the method. Assumption (A3) also concerns the regularity of the RKHS and the choice of the space V_4 (we simply ask for $\mathcal{U}$ to be continuously embedded in V_4) and is a matter of the setup of the problem. Condition (A6) is less natural as it requires the norm of the approximate solution $v^\dagger$ to be controlled by the norm of $v^\star$. While this condition does not hold for many numerical approximation schemes, we will see that it follows easily from the setup of our collocation/optimal recovery scheme.

In plain words, the most important message of Theorem 6.3.1 is that: *given Condition* (6.23) *and the Lipschitz-continuity of $\mathcal{T}$ and its inverse, it follows that the*

approximation error between $v^\dagger$ and $v^\star$ is bounded by the approximation error between $\mathcal{T}(v^\dagger)$ and $\mathcal{T}(v^\star)$. This result can be applied to both GP/kernel and ANN based collocation methods, since both seek to minimize the error between $\mathcal{T}(v^\dagger)$ and $\mathcal{T}(v^\star)$ at collocation points. This Condition (6.23) is automatically satisfied for our GP/Kernel based methods that solve problems of the form

$$v^\dagger = \operatorname{argmin}\|v\|_{\mathcal{U}} \qquad \text{s.t.} \qquad [\phi_i, \mathcal{T}(v)] = [\phi_i, \mathcal{T}(v^\star)], \quad i = 1, \dots, M,$$

with $\mathcal{T}$ denoting the differential operator of a PDE and ϕ_i denoting a set of dual elements (e.g. pointwise evaluations at collocation points). Then since the true solution $v^\star$ satisfies the PDE for an infinite collection of dual elements (e.g. pointwise within a set, or in a weak sense) then we immediately have that $\|v^\dagger\|_{\mathcal{U}} \leq \|v^\star\|_{\mathcal{U}}$. One can also take $\mathcal{U}$ to be a Barron space (indeed the V_i norms could be arbitrary) to obtain an analogous result for ANNs, but it is unclear if this setup coincides with (or leads to) any practical algorithms.

The Case of Second Order Nonlinear PDEs

We begin our error analysis in the case where (6.3) does not depend on the parameter $\boldsymbol{\theta}$ and homogeneous Dirichlet boundary conditions are imposed, i.e., nonlinear second order PDEs of the form

$$\begin{cases} \mathcal{P}(u^\star)(\boldsymbol{x}) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ u^\star(\boldsymbol{x}) = 0, & \boldsymbol{x} \in \partial\Omega. \end{cases} \tag{6.25}$$

The choice of Dirichlet boundary conditions is only made for simplicity here and can be replaced with other conditions of interest. We will also consider approximate boundary conditions in Section 6.3. We further assume that the kernel K is chosen so that the elements of $\mathcal{U}$ readily satisfy the boundary conditions of the PDE and consider optimization problems of the form

$$\begin{cases} \underset{u \in \mathcal{U}}{\text{minimize}} & \|u\|_{\mathcal{U}} \\ \text{st} & \mathcal{P}(u)(\boldsymbol{x}_m) = f(\boldsymbol{x}_m), \quad m = 1, \dots, M_\Omega, \\ & u(\boldsymbol{x}) = 0, \qquad\qquad \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{6.26}$$

where $X_\Omega := \{\boldsymbol{x}_m\}_{m=1}^{M_\Omega} \subset \Omega$ are a set of collocation points. We need to impose appropriate assumptions on the RKHS $\mathcal{U}$, the domain Ω and the PDE operator $\mathcal{P}$.

Assumption 2. The following conditions hold:

(B1) *(Regularity of the domain)* $\Omega \subset \mathbb{R}^d$ is a compact set with a Lipschitz boundary.

(B2) *(Stability of $\mathcal{P}$)* There exist indices $\gamma > 0$ and $k \in \mathbb{N}$ satisfying $d/2 < k + \gamma$ and $s \geq 1, \ell \in \mathbb{R}$, so that for any $r > 0$ it holds that

$$\|u_1 - u_2\|_{H^\ell(\Omega)} \leq C\|\mathcal{P}(u_1) - \mathcal{P}(u_2)\|_{H^k(\Omega)}, \quad \forall u_1, u_2 \in B_r(H^\ell(\Omega) \cap H_0^1(\Omega)) \tag{6.27}$$

$$\|\mathcal{P}(u_1) - \mathcal{P}(u_2)\|_{H^{k+\gamma}(\Omega)} \leq C\|u_1 - u_2\|_{H^s(\Omega)}, \quad \forall u_1, u_2 \in B_r(H^s(\Omega) \cap H_0^1(\Omega)) \tag{6.28}$$

where $C = C(r) > 0$ is independent of the u_i's. The space $H^s(\Omega) \cap H_0^1(\Omega)$ can be equipped with the norm $\|\cdot\|_{H^s(\Omega)}$, which is used to define the balls above.

(B3) $\mathcal{U}$ is continuously embedded in $H^s(\Omega) \cap H_0^1(\Omega)$.

Item (B1) is standard while (B3) dictates the choice of the RKHS $\mathcal{U}$, and in turn the kernel, which should be made based on a priori knowledge about regularity of the strong solution $u^\star$. We highlight that, asking elements of $\mathcal{U}$ to satisfy the boundary conditions is only practical for simple domains and boundary conditions such as periodic, Dirichlet, or Neumann conditions on hypercubes or spheres. Assumption (B2) on the other hand is a question in the analysis of nonlinear PDEs and is independent of our numerical scheme; simply put we require the PDE to be Lipschitz well-posed with respect to the right hand side/source term.

We are now ready to present our first theoretical result characterizing the convergence of the minimizer $u^\dagger$ of (6.26) to $u^\star$ the strong solution of (6.25).

Theorem 6.3.2. *Suppose Assumption 2 is satisfied and let $u^\star \in \mathcal{U}$ denote the unique strong solution of* (6.25)*. Let $u^\dagger$ be a minimizer of* (6.26) *with a set of collocation points $X_\Omega \subset \Omega$ and define their fill-distance*

$$h_\Omega := \sup_{\boldsymbol{x}' \in \Omega} \inf_{\boldsymbol{x} \in X_\Omega} |\boldsymbol{x}' - \boldsymbol{x}|.$$

Then there exists a constant $h_0 > 0$ so that if $h_\Omega < h_0$ then

$$\|u^\dagger - u^\star\|_{H^\ell(\Omega)} \leq C h_\Omega^\gamma \|u^\star\|_{\mathcal{U}},$$

where the constant $C > 0$ is independent of $u^\dagger$, and h_Ω.

Proof. We will obtain the result by applying Theorem 6.3.1 with the map $\mathcal{T} \equiv \mathcal{P}$ and the spaces $V_1 \equiv H^\ell(\Omega)$, $V_2 \equiv H^k(\Omega)$, $V_3 \equiv H^{k+\gamma}(\Omega)$, and $V_4 \equiv H^s(\Omega)$. With this setup we proceed to verify the conditions of Theorem 6.3.1: Condition (A1) follows from (6.27), (A2) follows from (6.28), and (A3) follows from (B3).

Condition (A6) holds since $u^\dagger$ is a minimizer of (6.26) and so $\|u^\dagger\|_{\mathcal{U}} \leq \|u^\star\|_{\mathcal{U}}$, since $u^\star$ is feasible but satisfies additional constraints compared with $u^\dagger$, i.e., it solves the PDE over the entire set Ω. Thus, (6.23) is verified with constant $C = 1$.

It remains to verify (A4): Let $\bar{f} = \mathcal{P}(u^\dagger) - \mathcal{P}(u^\star)$ and observe that $\bar{f}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in X_\Omega$. Thus $\bar{f} \in H^{k+\gamma}(\Omega)$ is zero on X_Ω and an application of Proposition 6.6.1 yields the existence of a constant $h_0 > 0$ so that whenever $h_\Omega < h_0$ then $\|\bar{f}\|_{H^k(\Omega)} \leq Ch_\Omega^\gamma \|\bar{f}\|_{H^{k+\gamma}(\Omega)}$. This verifies (6.22) with $\varepsilon \equiv Ch_\Omega^\gamma$. □

Remark 10. We note that Item (B2) and in turn Theorem 6.3.2 can easily be modified to a local version where the stability estimates (6.28) and (6.28) are stated for u_1, u_2 belonging to a ball of radius $r > 0$ around the true solution $u^\star$. Then one can obtain an asymptotic rate for $\|u^\dagger - u^\star\|_{H^\ell(\Omega)}$ under the additional assumption that $u^\dagger$ is sufficiently close to $u^\star$.

Remark 11. The assumptions and results of Assumption 2 are analogous to the one used to obtain error estimates in numerical homogenization for elliptic PDEs (Houman Owhadi and Clint Scovel, 2019b). In particular Theorem 6.3.2 can be extended to the setting where measurements on the PDE are not pointwise but involve integral operators and where the coefficients may be rough.

We now present a brief example where Assumption 2 can be verified and so Theorem 6.3.2 is applicable to obtain convergence rates for our GP/kernel collocation solver.

Example 6.3.3 (Nonlinear Darcy flow continued)**.** *Let us consider the nonlinear Darcy flow PDE* (6.6) *and assume that* Ω *has a smooth boundary and* $a(\boldsymbol{x}) \in C^\infty(\Omega)$ *is fixed and satisfies* $a(\boldsymbol{x}) \geq 1$*. Further suppose* $\tau(z) = 1 + \tanh(\beta z)$ *for a fixed constant* $\beta > 0$ *to be determined. Now pick* $k = \lceil d/2 + \alpha \rceil$ *from which it follows that* $H^k(\Omega)$ *is continuously embedded in* $C^\alpha(\bar{\Omega})$ *(Gilbarg and Trudinger, 1977, Thm. 7.26) and fix an integer* $\gamma > 0$[2]*. It is then straightforward to verify* (6.28) *with* $s = k + \gamma + 2$ *as well as Assumption 2(iii) by choosing the kernel K to be the*

[2]We only assume the exponents are integers for simplicity but our arguments can be generalized to the case of non-integer indices

Green's function of the operator $(-\Delta)^s$ on the domain Ω, subject to homogeneous Dirichlet boundary conditions. We note that approximating this Green's function can be expensive in practice and in Section 6.3 we propose a way around this step by using collocation points on the boundary of Ω to impose the boundary conditions.

Furthermore our assumptions on a and τ imply that $\mathcal{P}$ is uniformly elliptic in $\bar{\Omega}$ (see (Gilbarg and Trudinger, 1977, Part II) for definition of ellipticity for nonlinear elliptic PDEs). Since τ is smooth it follows from (Gilbarg and Trudinger, 1977, Thm. 13.8) that, for any $\alpha \in (0,1)$ and $f \in C^\alpha(\bar{\Omega})$, the PDE

$$\begin{cases} -div\big(\exp(a)\nabla u\big) + \tau(u) = f, & \boldsymbol{x} \in \Omega, \\ u = 0, & \boldsymbol{x} \in \partial\Omega, \end{cases} \tag{6.29}$$

has a solution $u \in C^2(\bar{\Omega})$. Now pick $f_1, f_2 \in H^k(\Omega)$ which, by the aforementioned Sobolev embedding result, belong to $C^\alpha(\bar{\Omega})$. Write $u_1, u_2 \in C^2(\bar{\Omega})$ for the solution of the PDE with both right hand sides and observe that the difference $w := u_1 - u_2$ solves the PDE

$$-div\big(\exp(a)\nabla w\big) = f_1 - f_2 + \tau(u_2) - \tau(u_1).$$

Standard stability results for linear elliptic PDEs then imply the bound

$$\|w\|_{H^2(\Omega)} \leq B\left(\|f_1 - f_2\|_{L^2(\Omega)} + \|\tau(u_2) - \tau(u_1)\|_{L^2(\Omega)}\right),$$

for a constant $B > 0$ independent of w, f_1, f_2, u_1, u_2. Since τ is globally β-Lipschitz we infer that $\|\tau(u_1) - \tau(u_2)\|_{L^2(\Omega)} \leq \beta\|w\|_{L^2(\Omega)}$ which, together with the subsequent bound, yields

$$\|w\|_{H^2(\Omega)} \leq \frac{B}{1 - B\beta}\|f_1 - f_2\|_{L^2(\Omega)}.$$

Thus, assumption (6.27) *is satisfied with $\ell = 2$ as long as $\beta B < 1$.*

Handling Boundary Conditions

We now turn our attention to the case where (6.3) is still independent of the $\boldsymbol{\theta}$ parameter but involves non-trivial boundary conditions, i.e.,

$$\begin{cases} \mathcal{P}(u^\star)(\boldsymbol{x}) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ \mathcal{B}(u^\star)(\boldsymbol{x}) = g(\boldsymbol{x}), & \boldsymbol{x} \in \partial\Omega. \end{cases} \tag{6.30}$$

We will further assume that the elements of $\mathcal{U}$ do not satisfy the boundary conditions exactly and so boundary collocation points are utilized to approximate those

conditions leading to the problem

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}} & \|u\|_{\mathcal{U}} \\ \quad\text{st} & \mathcal{P}(u)(\boldsymbol{x}_m) = f(\boldsymbol{x}_m), \quad m = 1,\dots,M_\Omega, \\ & \mathcal{B}(u)(\boldsymbol{x}_m) = g(\boldsymbol{x}_m), \quad m = M_\Omega+1,\dots,M, \end{cases} \tag{6.31}$$

where $X_\Omega := \{\boldsymbol{x}_m\}_{m=1}^{M_\Omega} \subset \Omega$ are the interior collocation points as before and $X_{\partial\Omega} := \{\boldsymbol{x}_m\}_{m=M_\Omega+1}^{M} \subset \partial\Omega$ are the boundary collocation points. We will state our assumptions and results for PDEs in $d > 1$ dimensions since in the 1D case we can, in principle, impose the boundary conditions exactly by placing some collocation points on boundary. The main difference, in comparison to Theorem 6.3.2, is that here we need to impose new assumptions on the PDE operators $\mathcal{P}$ and $\mathcal{B}$ and the boundary of Ω to be able to use Proposition 6.6.1 (sampling inequality on manifolds) in the final step of the proof to obtain approximation rates for the boundary data.

Assumption 3. The following conditions hold:

(C1) *(Regularity of the domain and its boundary)* $\Omega \subset \mathbb{R}^d$ with $d > 1$ is a compact set and $\partial\Omega$ is a smooth connected Riemannian manifold of dimension $d-1$ endowed with a geodesic distance $\rho_{\partial\Omega}$.

(C2) *(Stability of the PDE)*

There exist $\gamma > 0$ and $k, t \in \mathbb{N}$ satisfying $d/2 < k + \gamma$ and $(d-1)/2 < t + \gamma$, and $s, \ell \in \mathbb{R}$, so that for any $r > 0$ it holds that

$$\begin{aligned} \|u_1 - u_2\|_{H^\ell(\Omega)} \leq & C\big(\|\mathcal{P}(u_1) - \mathcal{P}(u_2)\|_{H^k(\Omega)} \\ & + \|\mathcal{B}(u_1) - \mathcal{B}(u_2)\|_{H^t(\partial\Omega)}\big) \quad \forall u_1, u_2 \in B_r(H^\ell(\Omega)), \end{aligned} \tag{6.32}$$

$$\begin{aligned} \|\mathcal{P}(u_1) - \mathcal{P}(u_2)\|_{H^{k+\gamma}(\Omega)} + \|\mathcal{B}(u_1) - \mathcal{B}(u_2)\|_{H^{t+\gamma}(\partial\Omega)} & \\ \leq C\|u_1 - u_2\|_{H^s(\Omega)}, \qquad \forall u_1, u_2 \in B_r(H^s(\Omega)), & \end{aligned} \tag{6.33}$$

where $C = C(r) > 0$ is a constant independent of the u_i.

(C3) $\mathcal{U}$ is continuously embedded in $H^s(\Omega)$.

Observe that the above assumptions are analogous to Assumption 2 with the exception that we no longer work with the restricted Sobolev spaces H_0^k since we do not need to impose the boundary conditions. However, we need to state our stability results for both $\mathcal{P}$ and $\mathcal{B}$. We emphasize that the verification of condition (C2) remains

a question in the analysis of PDEs. We are now ready to extend Theorem 6.3.2 to the case of non-trivial boundary conditions.

Theorem 6.3.4. *Suppose Assumption 3 is satisfied and let $u^\star \in \mathcal{U}$ denote the unique strong solution of* (6.25)*. Let $u^\dagger$ be a minimizer of* (6.26) *with a set of collocation points $X \subset \overline{\Omega}$ where $X_\Omega \subset X$ denotes the collocation points in the interior of Ω and $X_{\partial\Omega}$ denotes the collocation points on the boundary $\partial\Omega$. Define the fill-distances*

$$h_\Omega := \sup_{\boldsymbol{x}' \in \Omega} \inf_{\boldsymbol{x} \in X_\Omega} |\boldsymbol{x}' - \boldsymbol{x}|, \qquad h_{\partial\Omega} := \sup_{\boldsymbol{x}' \in \partial\Omega} \inf_{\boldsymbol{x} \in X_{\partial\Omega}} \rho_{\partial\Omega}(\boldsymbol{x}', \boldsymbol{x}),$$

where $\rho_{\partial\Omega} : \partial\Omega \times \partial\Omega \to \mathbb{R}_+$ is the geodesic distance defined on $\partial\Omega$ (see Section 6.6), and set $\bar{h} := \max\{h_\Omega, h_{\partial\Omega}\}$. Then there exists a constant $h_0 > 0$ so that if $\bar{h} < h_0$ then

$$\|u^\dagger - u^\star\|_{H^s(\Omega)} \leq C\bar{h}^\gamma \|u^\star\|_{\mathcal{U}},$$

where $C > 0$ is independent of $u^\dagger$ and $\bar{h}$.

Proof. The proof follows an identical approach to Theorem 6.3.2 and applies Theorem 6.3.1 with the appropriate setup. We take the operator $\mathcal{T} : u \mapsto \big(\mathcal{P}(u), \mathcal{B}(u)\big)$. We then choose the spaces $V_1 \equiv H^\ell(\Omega)$, $V_2 \equiv H^k(\Omega) \times H^t(\partial\Omega)$, $V_3 \equiv H^{k+\gamma}(\Omega) \times H^{t+\gamma}(\Omega)$, and $V_4 \equiv H^s(\Omega)$ where we equip V_2 with the norm $\|(f, g)\|_2 := \|f\|_{H^k(\Omega)} + \|g\|_{H^t(\partial\Omega)}$ and similarly for V_3 with the $H^k(\Omega)$ and $H^t(\partial\Omega)$ norms replaced by $H^{k+\gamma}(\Omega)$ and $H^{t+\gamma}(\partial\Omega)$ norms.

Analogously to the proof of Theorem 6.3.2, we can verify Conditions (A1), (A2), and (A3) by the hypothesis of the theorem. Condition (A6) is also satisfied since $u^\dagger$ is a minimizer of (6.31) and so $\|u^\dagger\|_{\mathcal{U}} \leq \|u^\star\|_{\mathcal{U}}$ as $u^\dagger$ satisfies more relaxed constraints.

It remains for us to verify (A4). Repeating the same argument as in the proof of Theorem 6.3.2, in the interior of Ω, yields the bound

$$\|\mathcal{P}(u^\dagger) - \mathcal{P}(u^\star)\|_{H^k(\Omega)} \leq C h_\Omega^\gamma \|\mathcal{P}(u^\dagger) - \mathcal{P}(u^\star)\|_{H^{k+\gamma}(\Omega)}, \tag{6.34}$$

whenever $h_\Omega < h_1$ and h_1 is a sufficiently small constant that is independent of $u^\dagger$ and $u^\star$.

Let $\bar{g} = \mathcal{B}(u^\dagger) - \mathcal{B}(u^\star)$ which satisfies $\bar{g}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in X_{\partial\Omega}$ and so $\bar{g} \in H^{t+\gamma}(\partial\Omega)$ is zero on the set $X_{\partial\Omega}$. Then Proposition 6.6.1 implies the existence of a constant $h_2 > 0$ so that whenever $h_{\partial\Omega} < h_2$ we have

$$\|\bar{g}\|_{H^t(\partial\Omega)} \leq C h_{\partial\Omega}^\gamma \|\bar{g}\|_{H^{t+\gamma}(\partial\Omega)}.$$

Now take $h_0 = \min\{h_1, h_2\}$ and combine the above bound with (6.34), and substitute the definition of $\bar{g}$ to get

$$\|\mathcal{P}(u^\dagger) - \mathcal{P}(u^\star)\|_{H^k(\Omega)} + \|\mathcal{B}(u^\dagger) - \mathcal{B}(u^\star)\|_{H^t(\partial\Omega)} \\ \leq C\bar{h}^\gamma \big(\|\mathcal{P}(u^\dagger) - \mathcal{P}(u^\star)\|_{H^{k+\gamma}(\Omega)} + \|\mathcal{B}(u^\dagger) - \mathcal{B}(u^\star)\|_{H^{t+\gamma}(\partial\Omega)}\big),$$

whenever $\bar{h} < h_0$. This verifies (A4) with $\varepsilon \equiv C\bar{h}^\gamma$. □

Remark 12. We highlight that our statement of Theorem 6.3.4 can easily be extended to PDEs with mixed boundary conditions simply by modifying the norm that is chosen on the boundary, i.e., the spaces V_2 and V_4, so long as we can prove the requisite stability estimates in condition (C2). In particular, this idea will allow us to obtain errors for time-dependent PDEs, cast as a static PDE in a space-time domain Ω with the initial and boundary conditions imposed as mixed conditions on $\partial\Omega$. In fact, in the case of time-dependent PDEs we do not need to impose the boundary conditions on all of $\partial\Omega$ but only on a subset.

We now return to our running example to verify Assumption 3 for the Darcy flow PDE.

Example 6.3.5 (Nonlinear Darcy flow continued)**.** *Consider the PDE* (6.29) *but this time with the boundary condition $u = g$ on $\partial\Omega$ for a function $g \in H^{t+\gamma}(\partial\Omega)$ with $t > \min\{3/2, (d-1)/2\}$ and $\gamma > 0$. Now fix a function $\varphi \in H^{t+\gamma+1/2}(\Omega)$ so that its trace coincides with g and define $v = u - \varphi$ and observe that u solves the above PDE if v solves*

$$\begin{cases} -div\big(\exp(a)\nabla v\big) + \tau'(v) = f', & \boldsymbol{x} \in \Omega, \\ v = 0, & \boldsymbol{x} \in \partial\Omega, \end{cases}$$

where we defined $\tau'(v) := \tau(v+\varphi)$ and $f' := f + div\big(\exp(a)\nabla\varphi\big)$. Now observe that the functions τ' and f' still satisfy the same conditions as τ, f in Example 6.3.3 and so we obtain existence and uniqueness of the solutions v and in turn u.

Now consider two solutions u_1, u_2 arising from source terms f_1, f_2 and boundary data g_1, g_2. Then the error $w = u_1 - u_2$ solves the PDE

$$\begin{cases} -div\big(\exp(a)\nabla w\big) = f + \tau(u_2) - \tau(u_1), & \boldsymbol{x} \in \Omega, \\ w = g_1 - g_2, & \boldsymbol{x} \in \partial\Omega. \end{cases}$$

By standard stability results for linear elliptic PDEs (McLean, 2000, Thm. 4.18) we have

$$\|w\|_{H^2(\Omega)} \leq B(\|f_1 - f_2\|_{L^2(\Omega)} + \|\tau(u_1) - \tau(u_2)\|_{L^2(\Omega)} + \|g_1 - g_2\|_{H^{3/2}(\Omega)}).$$

We can now repeat the same argument as in the final steps of Example 6.3.3 to get the bound

$$\|w\|_{H^2(\Omega)} \leq \frac{B}{1-B\beta}\left(\|f_1 - f_2\|_{H^2(\Omega)} + \|g_1 - g_2\|_{H^{3/2}(\Omega)}\right),$$

which verifies Assumption 3(ii) with $s = 2$ provided that $\beta B < 1$.

The Case of Parametric PDEs

We now consider the setting of the parametric PDE (6.3). Our error estimates can be viewed as further extending Theorem 6.3.4 with additional assumptions due to the fact that we will need to approximate the solutions on the set $\Upsilon = \Omega \times \Theta$ as well as its relevant boundary which needs to be sufficiently regular for us to apply Proposition 6.6.1. Beyond this technical point, the statement and proof of the result for parametric PDEs is identically to PDEs with boundary conditions and so we state our results succinctly, starting with the requisite assumptions on the parametric PDE.

Assumption 4. The following conditions hold:

(D1) $\Omega \subset \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^p$ are compact sets such that $\partial\Omega$ and $\partial\Theta$ are smooth Riemannian manifolds of dimensions $d-1$ and $p-1$ respectively.

(D2) *(Stability of the parametric PDE)* There exist $\gamma > 0$ and $k, t \in \mathbb{N}$ satisfying $(d+p)/2 < k + \gamma$ and $(d+p-1)/2 < t + \gamma$, and Banach spaces V_1 and V_4 so that for any $r > 0$ it holds that

$$\begin{aligned}&\|u_1 - u_2\|_1 \\ &\quad \leq C\left(\|\mathcal{P}(u_1) - \mathcal{P}(u_2)\|_{H^k(\Upsilon)} + \|\mathcal{B}(u_1) - \mathcal{B}(u_2)\|_{H^t(\partial\Upsilon)}\right) \quad \forall u_1, u_2 \in B_r(V_1),\end{aligned} \tag{6.35}$$

$$\begin{aligned}&\|\mathcal{P}(u_1) - \mathcal{P}(u_2)\|_{H^{k+\gamma}(\Upsilon)} + \|\mathcal{B}(u_1) - \mathcal{B}(u_2)\|_{H^{t+\gamma}(\partial\Upsilon)} \\ &\qquad\qquad \leq C\|u_1 - u_2\|_4, \quad \forall u_1, u_2 \in B_r(V_4),\end{aligned} \tag{6.36}$$

where $C = C(r) > 0$ is a constant independent of the u_i.

(D3) $\mathcal{U}$ is continuously embedded in V_4.

Unlike Assumptions 2 and 3 here we left the function spaces V_1 and V_4 as generic Banach spaces of functions $u : \Upsilon \mapsto \mathbb{R}$ since, for parametric PDEs, we can often obtain the desired stability results in non-standard norms, such as the mixed norm in Example 6.3.7 below, as opposed to the Sobolev norms used for the non-parametric

PDE setting. More generally, one may also impose V_2, V_3 to be generic Banach spaces rather than the standard Sobolev spaces. The Sobolev space setting suffices for applications in this paper.

With the above assumptions we can now present our main result for the parametric PDE setting. The proof is omitted since it is identical to that of Theorem 6.3.4 except that (1) the argument on $\partial\Omega$ is now repeated for $\partial\Upsilon = \partial\Omega \times \Theta$ which is in general a smooth manifold with boundary but this modification does not affect any of the steps in the proof, and (2) the results are stated in terms of the norm on the space V_1.

Theorem 6.3.6. *Suppose Assumption 4 is satisfied and let $u^\star \in \mathcal{U}$ denote the unique strong solution of* (6.3)*. Let $u^\dagger$ be a minimizer of* (6.4) *with a set of collocation points $S \subset \Upsilon \cup \partial\Upsilon$ where $S_\Upsilon \subset S$ denotes the collocation points in the interior of Υ and $S_{\partial\Upsilon}$ denotes the collocation points on the boundary $\partial\Upsilon$. Define the fill-distances*

$$h_\Upsilon := \sup_{\mathbf{s}' \in \Upsilon} \inf_{\mathbf{s} \in S_\Upsilon} |\mathbf{s}' - \mathbf{s}|, \qquad h_{\partial\Upsilon} := \sup_{\mathbf{s}' \in \partial\Upsilon} \inf_{\mathbf{s} \in S_{\partial\Upsilon}} \rho_{\partial\Upsilon}(\mathbf{s}', \mathbf{s}),$$

where $\rho_{\partial\Upsilon} : \partial\Upsilon \times \partial\Upsilon \to \mathbb{R}_+$ is the geodesic distance defined on $\partial\Upsilon$ (see Section 6.6), and set $\bar{h} := \max\{h_\Upsilon, h_{\partial\Upsilon}\}$. Then there exists a constant $h_0 > 0$ so that if $\bar{h} < h_0$ then

$$\|u^\dagger - u^\star\|_1 \leq C\bar{h}^\gamma \|u^\star\|_{\mathcal{U}},$$

where $C > 0$ is independent of $u^\dagger$ and $\bar{h}$.

We end this section by returning to our example of the Darcy flow PDE but this time in the setting where the coefficient a and the source f are dependent on a finite dimensional parameter $\boldsymbol{\theta}$. We will show that Assumption 4 can be verified in this case, with V_1 and V_4 taken as Banach spaces with mixed Sobolev and L^2 norms, and so Theorem 6.3.6 is applicable.

Example 6.3.7 (1D Parametric Darcy flow PDE)**.** *Consider the parametric elliptic PDE*

$$\begin{cases} -div\big(A(\boldsymbol{x}, \boldsymbol{\theta})\nabla u\big) = f(\boldsymbol{x}, \boldsymbol{\theta}) & \boldsymbol{x} \in \Omega, \\ u = 0, & \boldsymbol{x} \in \partial\Omega, \end{cases}$$

over a compact domain Ω and $\boldsymbol{\theta} \in \Theta$ where both Ω and Θ are assumed to satisfy condition (D1); e.g., take Ω and Θ to be unit balls. For simplicity we are ignoring the boundary operator in this case and imposing homogeneous Dirichlet boundary conditions. In this example we assume a is smooth in both $\boldsymbol{x}$ and $\boldsymbol{\theta}$, and there exists $m, M > 0$ such that $m \leq A(\boldsymbol{x}, \boldsymbol{\theta}) \leq M$. As a concrete example we may take

$A(\boldsymbol{x},\boldsymbol{\theta}) = \sum_{j=1}^{p} \theta_j \psi_j(\boldsymbol{x})$ *where the* ψ_j *are a set of smooth functions on* $\overline{\Omega}$ *that are uniformly bounded from below.*

First, the boundness of the operator $\mathcal{P}$ *is straightforward to obtain, since* a *is smooth and its derivatives will be bounded in the bounded domain* $\Omega \times \Theta$*. More precisely, for any* $\gamma > 0$*, since* $f = -div\big(A(\boldsymbol{x},\boldsymbol{\theta})\nabla u\big) = -\nabla_{\boldsymbol{x}} A \cdot \nabla_{\boldsymbol{x}} u - A\Delta_{\boldsymbol{x}} u$*, there exists some constant* C *independent of* u *and* f *such that*

$$\|f\|_{H^{\gamma}(\Omega\times\Theta)} \leq C\|u\|_{H^{\gamma+2}(\Omega\times\Theta)}\,.$$

Due to the linearity of the equation, by replacing u *by* $u_1 - u_2$ *and noting that* $f = \mathcal{P}u = \mathcal{P}u_1 - \mathcal{P}u_2$*, we obtain the forward stability*

$$\|\mathcal{P}u_1 - \mathcal{P}u_2\|_{H^{\gamma}(\Omega\times\Theta)} \leq C\|u_1 - u_2\|_{H^{\gamma+2}(\Omega\times\Theta)}\,. \tag{6.37}$$

For the backward stability estimate, via intergation by parts, we have

$$\begin{aligned}
\int_{\Upsilon} A(\boldsymbol{x},\boldsymbol{\theta})|\nabla_{\boldsymbol{x}} u(\boldsymbol{x},\boldsymbol{\theta})|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\theta} &= \int_{\Upsilon} u(\boldsymbol{x},\boldsymbol{\theta}) f(\boldsymbol{x},\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\theta} \\
&\leq \int_{\Theta} \|u(\cdot,\boldsymbol{\theta})\|_{L^2(\Omega)} \|f(\cdot,\boldsymbol{\theta})\|_{L^2(\Omega)} \,\mathrm{d}\boldsymbol{\theta} \\
&\leq C_0 \int_{\Theta} \|\nabla_{\boldsymbol{x}} u(\cdot,\boldsymbol{\theta})\|_{L^2(\Omega)} \|f(\cdot,\boldsymbol{\theta})\|_{L^2(\Omega)} \,\mathrm{d}\boldsymbol{\theta} \\
&\leq C_1 \|u\|_{L^2(\Theta,H_0^1(\Omega))} \|f\|_{L^2(\Theta,L^2(\Omega))}\,,
\end{aligned}$$

where in the first and third inequalities, we used the Cauchy-Schwarz inequality; in the second inequality, we used the Poincaré inequality as $u(\cdot,\boldsymbol{\theta})$ *is zero on* $\partial\Omega$*. Here we used the notation:*

$$\|u\|^2_{L^2(\Theta,H_0^1(\Omega))} := \int_{\Theta} \|u(\cdot,\boldsymbol{\theta})\|^2_{H_0^1(\Omega)} \,\mathrm{d}\boldsymbol{\theta} \quad \textit{and} \quad \|f\|^2_{L^2(\Theta,L^2(\Omega))} := \int_{\Theta} \|f(\cdot,\boldsymbol{\theta})\|^2_{L^2(\Omega)} \mathrm{d}\boldsymbol{\theta}.$$

Note that the $L^2(\Theta, L^2(\Omega))$ *norm is also equivalent to the* $L^2(\Omega \times \Theta)$ *norm. Now, using the bound on* A*, we obtain that there exists a constant* C *such that*

$$\|u\|_{L^2(\Theta,H^1(\Omega))} \leq C\|f\|_{L^2(\Omega\times\Theta)}\,.$$

Similar to the proof for the forward stability, the backward stability follows by the linearity of the equation. We have

$$\|u_1 - u_2\|_{L^2(\Theta,H^1(\Omega))} \leq C\|\mathcal{P}u_1 - \mathcal{P}u_2\|_{L^2(\Omega\times\Theta)}. \tag{6.38}$$

Thus it follows that we can verify condition (6.35) *with the norm* $\|\cdot\|_1 \equiv \|\cdot\|_{L^2(\Theta,H_0^1(\Omega))}$, $\|\cdot\|_4 = \|\cdot\|_{H^{\gamma+2}(\Omega\times\Theta)}$*, and* $k = 0$*.*

Bounding Fill-distances

Our bounds in Theorems 6.3.2 and 6.3.6 are given in terms of the fill distances h of our collocation points. In this section, we provide an upper bound of these fill-in distances in terms of the number of collocation points, under the assumptions that the points are randomly drawn according to uniform distributions both in the interior of the domains and their pertinent boundaries. Throughout this section we only consider the case of non-parametric PDEs, and hence we work with Ω, assumed to be a compact subset of $\mathbb{R}^d$ with boundary $\partial\Omega$ which is a compact smooth manifold of dimension $d-1$. We focus on the non-parametric setting for simplicity and our results can easily be extended to the parametric PDE setting by simply replacing Ω with Υ as a compact subset of $\mathbb{R}^{d+p}$.

Proposition 6.3.8. *Suppose we sample M_Ω points in Ω and $M_{\partial\Omega}$ points on $\partial\Omega$, uniformly with respect to the canonical volume and surface measures. Let $\delta > 0$. Then, with probability at least $1-\delta$, the fill-in distances h_Ω and $h_{\partial\Omega}$ satisfy*

$$h_\Omega \leq C\left(\frac{\log(M_\Omega/\delta)}{M_\Omega}\right)^{1/d}, \quad h_{\partial\Omega} \leq C\left(\frac{\log(M_{\partial\Omega}/\delta)}{M_{\partial\Omega}}\right)^{1/(d-1)},$$

where C is a constant independent of $M_\Omega, M_{\partial\Omega}$, and δ.

The proof of Proposition 6.3.8 can be found in Section 6.7.

Let's combine Proposition 6.3.8 and previous error estimates to get error bounds regarding the number of collocation points. In the case of Theorem 6.3.2 where there is no boundary, we get

$$\|u^\dagger - u^\star\|_{H^s(\Omega)} \leq C\left(\frac{\log(M_\Omega/\delta)}{M_\Omega}\right)^{\gamma/d}\|u^\star\|_{\mathcal{U}},$$

while in the case of Theorem 6.3.4 where boundary is considered, we have

$$\|u^\dagger - u^\star\|_{H^s(\Omega)} \leq C\left(\left(\frac{\log(M_\Omega/\delta)}{M_\Omega}\right)^{\gamma/d} + \left(\frac{\log(M_{\partial\Omega}/\delta)}{M_{\partial\Omega}}\right)^{\gamma/(d-1)}\right)\|u^\star\|_{\mathcal{U}}.$$

More generally, we note that the bounds in Proposition 6.3.8 can be applied to the abstract setting in Theorem 6.3.1 when ϵ depends on the fill-in distance.

If s and γ are appropriately chosen such that the required assumptions hold, and $\gamma \geq d/2$, then the convergence rate is at least as fast as the Monte Carlo rate, for uniformly sampled collocation points. There is no curse of dimensionality in this case.

6.4 Numerical Experiments

In this section, we study several numerical examples to demonstrate the interplay between the dimensionality of the problem and the regularity of the solution. Our theory demonstrates that this interplay is central to determining the convergence rate, and hence accuracy, of the methodology studied in this paper.

In Section 6.4, we consider a high dimensional elliptic PDE with smooth solutions. By varying the dimension of the problem and the frequency of the solution, we demonstrate dimension-benign convergence rates, and in particular the accuracy is better when the frequency of the solution is lower. In Section 6.4, we consider a high dimensional parametric PDE problem to illustrate the importance of choosing kernels that adapted to the regularity of the solution. In Section 6.4, we present a high dimensional Hamilton-Jacobi-Bellman (HJB) equation, which goes beyond our theory and demonstrates the interplay between dimensionality and regularity.

High Dimensional PDEs

Consider the variable coefficient nonlinear elliptic PDEs

$$\begin{cases} -\nabla \cdot (A\nabla u) + u^3 = f, & \text{in } \Omega, \\ u = g, & \text{on } \partial\Omega. \end{cases} \tag{6.39}$$

We set $A(\boldsymbol{x}) = \exp\left(\sin\left(\sum_{j=1}^{d} \cos(x_j)\right)\right)$, and the ground truth solution

$$u^\star(\boldsymbol{x};\beta) = \exp(\sin(\beta \sum_{j=1}^{d} \cos(x_j))),$$

where we have a parameter β to control the frequency of u. The right hand side and boundary data are obtained using A and $u^\star$.

In the experiment, we choose the domain Ω to be the unit ball in $\mathbb{R}^d$ for $d = 2, 3, \ldots, 6$. We sample $M_\Omega = 1000, 2000, 4000, 8000$ points uniformly in the interior, and respectively $M_{\partial\Omega} = 200, 400, 800, 1600$ points uniformly on the boundary.

After selecting the kernel function, the number of iteration steps in our algorithm is set to be 3 with initial solution 0. We sample another set of M_Ω test points and evaluate the L^2 error of the solution on these points. The results are averaged over 10 independent draws of the uniform collocation points.

In the first experiment, we choose the Matérn kernel with $\nu = 7/2$ and with length-scale $\sigma = 0.25\sqrt{d}$. We choose $\beta = 1, 4$, to compare the convergence given ground

truth with different frequencies. The results are shown in fig. 6.2. It is clear that when β is small, the accuracy is better. The slopes of convergence curves also have a tendency to improve for $d \geq 3$ if we increase β.

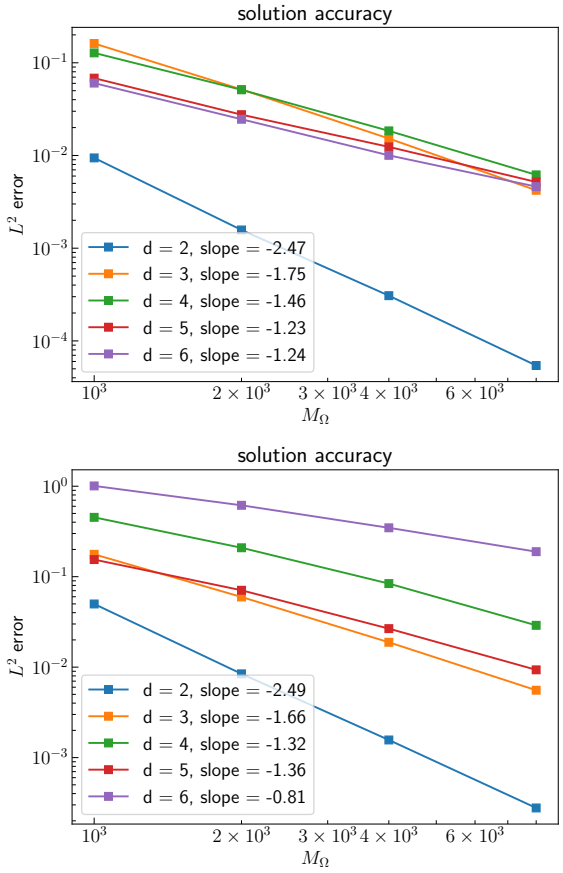


Figure 6.2: L^2 test errors of solutions to Problem (6.39) as a function of the number of collocation points. Left: $\beta = 1$; right: $\beta = 4$. In both cases, we choose Matérn kernel with $\nu = 7/2$. Reported slopes in the legend denote empirical convergence rates.

In the second experiment, we fix $\beta = 4$, and choose the Matérn kernel with $\nu = 5/2, 9/2$ and with lengthscale $\sigma = 0.25\sqrt{d}$. Results are shown in Figure 6.3. Comparing $\nu = 5/2, 9/2$ and $\nu = 7/2$ in the last example, we observe that increasing ν leads to faster convergence. This is due to the fact that the true solution is smooth. In dimension $d = 2$, we can identify the exact convergence rate as $\nu - 1$. In all dimensions, the rate is faster than the Monte Carlo rate. We observe that the regularity of the solution softens the effect of the curse of dimensionality, i.e., convergence rates are better in higher dimensions when β is smaller.

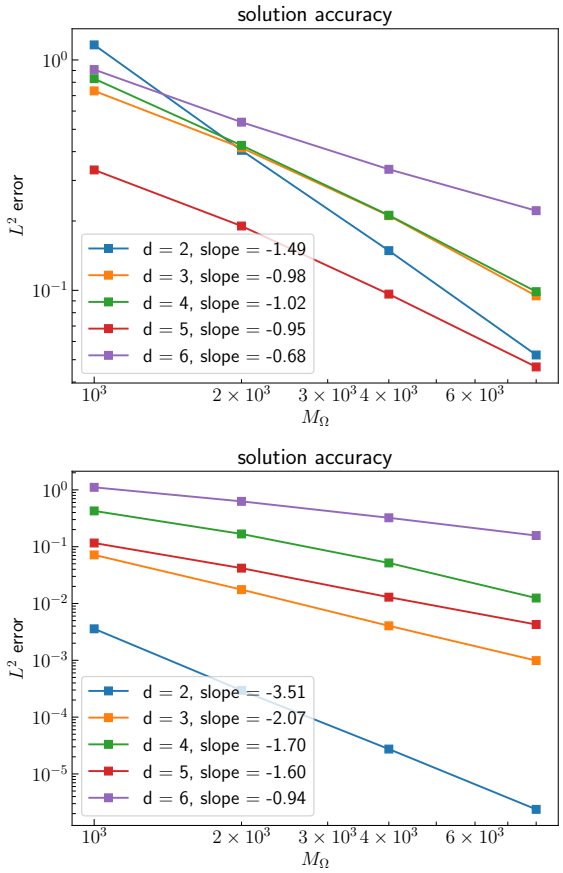


Figure 6.3: L^2 test errors of solutions to Problem (6.39) as a function of the number of collocation points with $\beta = 4$. Left: Matérn kernel with $\nu = 5/2$; right: Matérn kernel with $\nu = 9/2$. Reported slopes in the legend denote empirical convergence rates.

Parametric PDEs

We consider a parametric version of the linear ($\tau = 0$) darcy flow problem in example 6.2.1:

$$\begin{cases} -\mathrm{div}\big(\exp(a(\boldsymbol{x},\boldsymbol{\theta}))\nabla u\big)(\boldsymbol{x}) = f(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) = g(\boldsymbol{x}), & \boldsymbol{x} \in \partial\Omega. \end{cases} \tag{6.40}$$

Following the general form eq. (6.3), we aim to obtain the solution as a function taking values in the product space Υ. eq. (6.40) can be rewritten in terms of $\mathbf{s} = (\boldsymbol{x}, \boldsymbol{\theta})$ with new forcing terms $\widehat{f}$ and $\widehat{g}$ depending only on the first coordinate of $\mathbf{s}$

$$\begin{cases} -\mathrm{div}_{\boldsymbol{x}}\big(A(\mathbf{s})\nabla_{\boldsymbol{x}} u(\mathbf{s})\big)(\mathbf{s}) = \widehat{f}(\mathbf{s}) = f(\boldsymbol{x}), & \mathbf{s} \in \Upsilon, \\ u(\mathbf{s}) = \widehat{g}(\mathbf{s}) = g(\boldsymbol{x}), & \mathbf{s} \in \partial\Upsilon. \end{cases} \tag{6.41}$$

Recall that we defined $\partial\Upsilon = \partial\Omega \times \Theta$. For our numerical example, we let $d = 1$ and vary p. We set $A(x, \theta) = 2 + \theta_0 + \sum_{j=1}^{p} \frac{\theta_j}{j^k} \sin(\pi x + j)$, $f(x) = x$ and $g(x) = 0$, a similar setting as in (Chkifa, Cohen, DeVore, et al., 2012). We choose $\Omega = [0, 1]$, and $\Theta = [0, 1]^p$, for $p = 2, 3, \ldots, 6$. Note $1 \leq A(\mathbf{s}) \leq 4$ since the sum is in $[-1, 1]$ for all p and $\theta \in \Theta$, matching the setting of example 6.3.7.

We sample different M_Ω points uniformly in the interior, and $M_{\partial\Omega} = M_\Omega/10$ points uniformly on the boundary of x. We do two experiments with different choices of kernel, in the first (fig. 6.4, left), a vanilla Gaussian kernel with different length scales for the x and θ dimension, and with a scaling of the length scale in θ proportional to $\sqrt{p}$. In the second one (fig. 6.4, right), we adapt the Gaussian kernel to the decay in $A(x, \theta)$, by including the decay of $1/j^k$ in the norm in θ space used by the kernel. We see significant improvement in test error using this adaptation in high dimensions, which suggests future research directions of kernel adaptation to the specific form of the PDE. In all cases, we use a cross-validation procedure for hyperparameter tuning and we observe the average L^2 test error on an independent set of test points for different values of p and M_Ω. Since $d = 1$ we computed our ground truth solution by numerically integrating Equation eq. (6.41) using quadrature.

As mentioned, this problem was also explored by (Chkifa, Cohen, DeVore, et al., 2012), in which sparse multivariate polynomials are used to estimate the solution with a rate independent of the number of parameters, provided the decay of the coefficient functions is large enough (in ℓ^p for some $0 < p < 1$). While this assumption is satisfied in this example, our method's convergence rate greatly depends on the dimension of θ when the kernel is not adapted to the particular equations and coefficients $A(x, \theta)$. Our results indicate improvement in the dependence of convergence rates on dimension when the kernel is adapted to the regularity of A. It remains open whether our kernel based approach (which is not specific to parametric equations) can achieve the same dimension independent convergence rates as the ones in (Chkifa, Cohen, DeVore, et al., 2012) (which apply even in the countably infinite dimensional case and which they refer to as breaking the curse of dimensionality) for parametric elliptic PDEs with rapidly decreasing parametric dependence as specified above (this assumption implies a finite number of effective parameters).

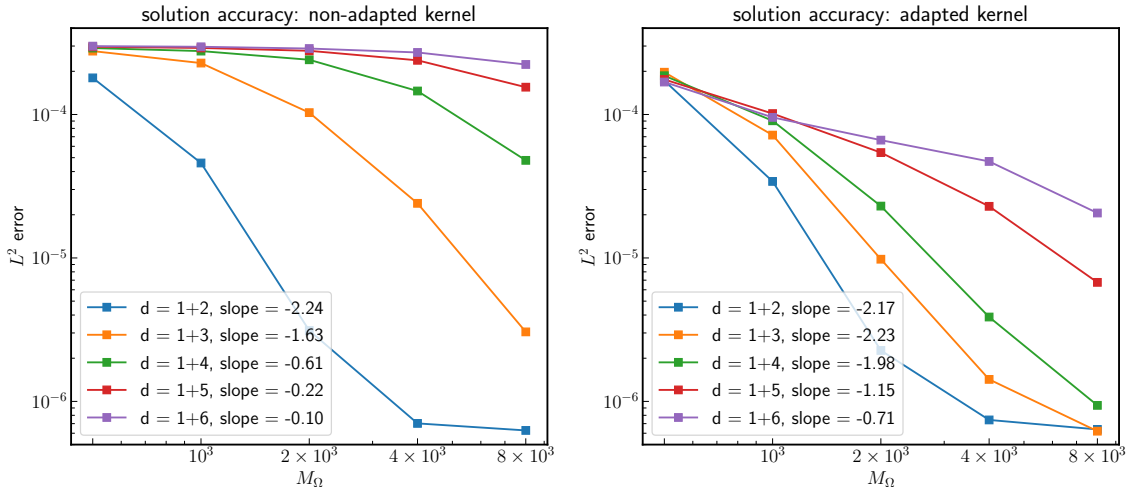


Figure 6.4: L^2 test error of solutions to Problem (6.41) as a function of the number of collocation points. Left: vanilla gaussian kernel; Right: Gaussian kernel adapted to the regularity of A. Reported slopes in the legend denote empirical convergence rates.

High Dimensional HJB Equation

Consider a prototypical HJB equation:

$$\begin{aligned}(\partial_t + \Delta)V(\boldsymbol{x}, t) - |\nabla V(\boldsymbol{x}, t)|^2 &= 0 \\ V(\boldsymbol{x}, T) &= g(\boldsymbol{x})\,,\end{aligned} \tag{6.42}$$

where $g(\boldsymbol{x}) = \log(\frac{1}{2} + \frac{1}{2}|\boldsymbol{x}|^2), \boldsymbol{x} \in \mathbb{R}^d, t \in [0, T]$. We are interested in solving $V(\boldsymbol{x}_0, 0)$ for some $\boldsymbol{x}_0 \in \mathbb{R}^d$. We adopt the stochastic differential equation (SDE) formula for representing the solution of the PDEs, following (Weinan, Han, and Jentzen, 2017; Richter, Sallandt, and Nüsken, 2021). More specifically, consider the SDE

$$dX_s = \sqrt{2}dW_s,\ X_0 = \boldsymbol{x}_0\,. \tag{6.43}$$

We define $Y_s = V(X_s, s), Z_s = \sqrt{2}\nabla V(X_s, s)$. By Ito's formula, one obtains

$$dY_s = \frac{1}{2}|Z_s|^2 ds + Z_s \cdot dW_s\,. \tag{6.44}$$

The strategy is to integrate the above SDE backward to Y_0. An implicit[3] Euler discretization from time t_{n+1} to t_n ($\Delta t = t_{n+1} - t_n$) leads to the following equation:

$$V(X_{t_{n+1}}, t_{n+1}) = V(X_{t_n}, t_n) + |\nabla V(X_{t_n}, t_n)|^2 \Delta t + \sqrt{2}\nabla V(X_{t_n}, t_n) \cdot \xi_{n+1}\sqrt{\Delta t}\,. \tag{6.45}$$

[3]Implicit because are integrating backwards in time.

Algorithmically, we sample J different paths of the forward SDE in (6.43), namely $X_{t_n}^{(j)}, 1 \le j \le J$, using the Euler–Maruyama scheme. Then, backward in time, we apply our kernel method, namely to solve the following optimization problem

$$\begin{cases} \underset{u\in\mathcal{U}}{\text{minimize}} & \|u\|_{\mathcal{U}} \\ \text{st} & u(X_{t_n}^{(j)}, t_n) + |\nabla u(X_{t_n}^{(j)}, t_n)|^2 \Delta t + \sqrt{2}\nabla u(X_{t_n}^{(j)}, t_n) \cdot \xi_{n+1}\sqrt{\Delta t} = V(X_{t_{n+1}}^{(j)}, t_{n+1}) \end{cases} \tag{6.46}$$

to get the solution $V(\cdot, t_n)$, assuming $V(\cdot, t_{n+1})$ has been solved. Iterating this process, we end up with the solution $V(\boldsymbol{x}_0, 0)$. We can understand the algorithm as applying our kernel method iteratively with the sample path as the collocation points.

Experimentally, we consider $d = 100$ as in (Weinan, Han, and Jentzen, 2017; Richter, Sallandt, and Nüsken, 2021). We aim to solve $V(\boldsymbol{x}_0, 0)$ for $\boldsymbol{x}_0 = 0$. The ground truth is $V(\boldsymbol{x}_0, 0) = 4.589992$ provided in (Weinan, Han, and Jentzen, 2017). We sample $J = 2000$ paths from $\boldsymbol{x}_0$ and choose the inverse quadratic kernel $k(\boldsymbol{x}, \boldsymbol{y}; \sigma) = \left(\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2d\sigma^2} + 1\right)^{-1}$. We use the "linearize-then-optimize" approach to compute an approximate solution to (6.46). The nugget term is set to be $\eta = 10^{-3}$. The result is shown in Table 6.2.

σ	10	25	50	100	200
Computed solution $V(\boldsymbol{x}_0, 0)$	5.6042	4.6366	4.6039	4.6021	4.6021
Relative accuracy	22.10%	1.0154%	0.303%	0.2638%	0.2638%

Table 6.2: Numerical results for the HJB equation (6.42), computing the quantity $V(\boldsymbol{x}_0, 0)$.

We observe that a suitable choice of the lengthscale of the kernel is crucial to obtain an accurate solution. Compared to the relative accuracy of 0.171% (reported in (Richter, Sallandt, and Nüsken, 2021)) using neural networks (DenseNet like architecture with four hidden layers) to solve (6.45), the accuracy of using kernel methods with a simple quadratic kernel is comparable. Moreover, the lengthscale of the kernel is very large, indicating that the solution behavior of this HJB equation is very smooth; similar "blessings of dimensionality" have been reported and discussed in (Richter, Sallandt, and Nüsken, 2021), where they used a constant function (and the terminal function g) as ansatz to solve (6.45) and obtained very high accuracy[4]. Thus, this HJB example in dimension 100 demonstrates again the trade-off between the smoothness of the solution and the curse of dimensionality.

[4]We anticipate that using the feature map perspective of kernel methods with constants and g as features will achieve a similar accuracy as in (Richter, Sallandt, and Nüsken, 2021). We did not pursue this here to avoid using strong prior information on the solution beyond regularity.

6.5 Conclusions

In this paper, we conducted an error analysis of GP and kernel based methods for solving PDEs. We provided convergence rates under the assumptions that (1) the solution belongs to the RKHS which is embedding to some Sobolev space of sufficient regularity, and (2) the underlying forward and inverse PDE operator is stable in corresponding Sobolev spaces.

Our analysis relies on the crucial minimizing norm property of the numerical solution in the kernel/GP methodology. The analysis could be seamlessly generalized to the function class of NNs and other norms such as non-quadratic norms if we can formulate the training process as a minimization problem over the related norm.

We emphasize that our convergence rates hold for the exact minimizer of the minimization problem. In practice, finding such a minimizer algorithmically can be a separate and challenging problem. Our numerical experience suggests that Gauss-Newton iterations usually perform well, and typically, 2-5 iterations are sufficient for convergence. Therefore, we can combine the error analysis in this paper and the fast implementation of the algorithm in (Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024) to obtain a near-linear complexity solver for nonlinear PDEs with rigorous accuracy guarantee.

It is worth mentioning that this paper focuses only on analyzing the MAP estimator within the GP interpretation. Exploring the posterior distribution of the GP can provide a means for quantifying uncertainty in the solution. In particular, analyzing the posterior contraction is an interesting direction for future research.

Acknowledgments

The authors gratefully acknowledge support by the Air Force Office of Scientific Research under MURI award number FA9550-20-1-0358 (Machine Learning and Physics-Based Modeling and Simulation). BH acknowledges support by the National Science Foundation grant number NSF-DMS-2208535 (Machine Learning for Bayesian Inverse Problems). HO also acknowedges support by the Department of Energy under award number DE-SC0023163 (SEA-CROGS: Scalable, Efficient and Accelerated Causal Reasoning Operators, Graphs and Spikes for Earth and Embedded Systems) and a Department of Defense Vannevar Bush Faculty Fellowship.

6.6 Sobolev Sampling Inequalities on Manifolds

Below we collect useful sampling inequalities for Sobolev functions defined on smooth manifolds with corners. Following (J. M. Lee, 2012, Chs. 1, 16) we consider a smooth, compact Riemannian manifold $\mathcal{M} \subset \mathbb{R}^d$ of dimension $k \leq d$ with corners, i.e., a Riemannian manifold with a smooth structure with corners; see (J. M. Lee, 2012, Ch. 16). On such a manifold we define the natural geodesic distance

$$\rho_{\mathcal{M}} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}, \qquad \rho_{\mathcal{M}}(x, y) := \inf \int_0^1 \|\dot{\ell}(t)\| dt,$$

where the infimum is taken over all piecewise smooth paths $\ell : [0, 1] \mapsto \mathcal{M}$ satisfying the boundary conditions $\ell(0) = x$ and $\ell(1) = y$, and $\|\dot{\ell}(t)\|$ is the length of the tangent vector $\dot{\ell}(T)$ under the Riemannian metric.

Following (Fuselier and G. B. Wright, 2012) (see also (Taylor, 2013, Sec. 4.3)) we further consider the Sobolev spaces $H^k(\mathcal{M})$ of functions defined on $\mathcal{M}$ as follows: Let $\mathcal{A} = \{M_j, \Psi_j\}_{j=1}^N$ be an atlas for $\mathcal{M}$ and let $\{\kappa_j\}$ be a partition of unity of $\mathcal{M}$, subordinate to M_j. Then given functions $u : \mathcal{M} \to \mathbb{R}$ we define the Sobolev norms and the associated Sobolev spaces $H^s(\mathcal{M})$ as

$$H^s(\mathcal{M}) := \{u : \mathcal{M} \to \mathbb{R} \mid \|u\|_{H^s(\mathcal{M})} < +\infty\}, \qquad \|u\|_{H^s(\mathcal{M})} := \left(\sum_{j=1}^N \|\pi_j(u)\|^2_{H^s(\Xi_j)} \right)^{1/2},$$

where the maps π_j are defined as

$$\pi_j(f) := \begin{cases} \kappa_j(f(\Psi_j^{-1}(y))), & \text{if } y \in \Psi_j(M_j), \\ 0 & \text{otherwise.} \end{cases}$$

and the sets Ξ_j are given by

$$\Xi_j := \begin{cases} \mathbb{R}^k & \text{if } \Psi_j \text{ is an interior chart,} \\ \{(x_1, \ldots, x_k) \in \mathbb{R}^k | x_1 \geq 0\} & \text{if } \Psi_j \text{ is a boundary chart,} \\ \{(x_1, \ldots, x_k) \in \mathbb{R}^k | x_1 \geq 0, \ldots, x_k \geq 0\} & \text{if } \Psi_j \text{ is a corner chart.} \end{cases}$$

Put simply, the Sobolev spaces $H^s(\mathcal{M})$ are functions on $\mathcal{M}$ that, locally after the flattening of the manifold belong to the standard Sobolev spaces H^s. With these notions at hand we then recall the following result of (Fuselier and G. B. Wright, 2012), which was proven by those authors for smooth embedded manifolds without boundary or corners. However, a brief investiation of the proof of that result reveals that it can immediately be generalized to our setting with manifolds with corners. In

fact, the idea of the proof is to use the atlas to locally flatten the manifold and apply classic sampling theorems such as (Arcangéli, López de Silanes, and Torrens, 2007, Thm. 4.1) on each patch. The only difference in the case of manifolds with corners is that the patches do not only map to $\mathbb{R}^k$ but rather to the subspaces Ξ_j depending on whether the corresponding chart is an interior, boundary, or corner chart.

Proposition 6.6.1 ((Fuselier and G. B. Wright, 2012, Lem. 10))**.** *Suppose $\mathcal{M} \subset \mathbb{R}^d$ is a smooth, compact, Riemannian manifold with corners, of dimension k and let $s > k/2$ and $r \in \mathbb{N}$ satisfy $0 \leq r \leq \lceil s \rceil - 1$. Let $X \subset \mathcal{M}$ be a discrete set with mesh norm $h_{\mathcal{M}}$ defined as*

$$h_{\mathcal{M}} := \sup_{x' \in \mathcal{M}} \inf_{x \in X} \rho_{\mathcal{M}}(x, x').$$

Then there is a constant $h_0 > 0$ depending only on $\mathcal{M}$ such that if $h_{\mathcal{M}} < h_0$ and if $u \in H^s(\mathcal{M})$ satisfies $u|_X = 0$ then

$$\|u\|_{H^r(\mathcal{M})} \leq C h_{\mathcal{M}}^{s-r} \|u\|_{H^s(\mathcal{M})}.$$

Here $C > 0$ is a constant independent of $h_{\mathcal{M}}$ and u.

6.7 Bounds on Fill Distances

This section collects a result from (Reznikov and Saff, 2016) for bounding the fill-in distance for randomly distributed points on a manifold.

Assume $(\mathcal{M}, \rho)$ is a metric space, and μ is a finite positive Borel measure supported on $\mathcal{M}$. Let $X = \{x_1, ..., x_N\}$ be a set of N points, independently and randomly drawn from μ. Define the fill-in distance

$$h_{\mathcal{M}} = \sup_{x' \in \mathcal{M}} \inf_{x \in X} \rho(x, x') \,. \tag{6.47}$$

Then, (Reznikov and Saff, 2016, Thm. 2.1) implies the following,

Proposition 6.7.1. *Suppose Φ is a continuous non-negative strictly increasing function on $(0, \infty)$ satisfying $\Phi(r) \to 0$ as $r \to 0^+$. If there exists a positive number r_0 such that $\mu(B(x, r)) \geq \Phi(r)$ holds for all $x \in \mathcal{M}$ and every $r < r_0$, then there exist positive constants c_1, c_2, c_3, and α_0 such that for any $\alpha > \alpha_0$, we have*

$$\mathbb{P}\left[h_{\mathcal{M}} \geq c_1 \Phi^{-1}\left(\frac{\alpha \log N}{N} \right) \right] \leq c_2 N^{1 - c_3 \alpha}. \tag{6.48}$$

We use this proposition to prove Proposition 6.3.8.

Proof of Proposition 6.3.8. We apply Proposition 6.7.1. For the bounded domain $\Omega \subset \mathbb{R}^d$, we know that there exists a constant C such that $\Phi(r) = Cr^d$ will satisfy the assumption in Proposition 6.7.1. Moreover, we choose α such that $c_2 M_\Omega^{1-c_3\alpha} \leq \delta$. This implies that $\alpha \geq \frac{1}{c_3 \log M_\Omega} \log(c_2 M_\Omega/\delta)$. Pick $\alpha = \frac{C'}{c_3 \log M_\Omega} \log(c_2 M_\Omega/\delta)$ for some $C' \geq 1$ such that $\alpha \geq \alpha_0$. Then Proposition 6.7.1 shows that with probability at least $1 - \delta$,

$$h_\Omega \leq c_1 \Phi^{-1}\left(\frac{\alpha \log M_\Omega}{M_\Omega}\right) \leq C''\left(\frac{\log(M_\Omega/\delta)}{M_\Omega}\right)^{1/d},$$

where C'' is a constant independent of M_Ω and δ. The bound on $h_{\partial\Omega}$ can be proved similarly by choosing $\Phi(r) = Cr^{d-1}$. □

6.8 The Choice of Nugget Terms

For numerical stability, we add a diagonal adaptive nugget term to the kernel matrix in our computation, such that

$$u^{\ell+1}(x) = K(x, \boldsymbol{\phi}^l)[K(\boldsymbol{\phi}^l, \boldsymbol{\phi}^l) + \eta \mathrm{diag}(K(\boldsymbol{\phi}^l, \boldsymbol{\phi}^l))]^{-1} \begin{pmatrix} \left(f - u^\ell + \mathcal{P}'(u^\ell)u^\ell\right)|_{\mathbf{s}_\Omega} \\ \left(g - u^\ell + \mathcal{B}'(u^\ell)u^\ell\right)|_{\mathbf{s}_{\partial\Omega}}. \end{pmatrix}$$

Typically $\eta = 10^{-10}$. This nugget term is similar to the adaptive nugget term proposed in (Yifan Chen, Hosseini, et al., 2021b). It is much more effective than the naive choice of $K(\boldsymbol{\phi}^l, \boldsymbol{\phi}^l)+\eta I$, since the conditioning of the interior block and the boundary block in the kernel matrix differs dramatically.

Chapter 7

DISCOVERING GRAPHICAL STRUCTURE AND FUNCTIONAL RELATIONSHIPS WITHIN DATA

Most problems within and beyond the scientific domain can be framed into one of the following three levels of complexity of function approximation. **Type 1:** Approximate an unknown function given input/output data. **Type 2:** Consider a collection of variables and functions, some of which are unknown, indexed by the nodes and hyperedges of a hypergraph (a generalized graph where edges can connect more than two vertices). Given partial observations of the variables of the hypergraph (satisfying the functional dependencies imposed by its structure), approximate all the unobserved variables and unknown functions. **Type 3:** Expanding on Type 2, if the hypergraph structure itself is unknown, use partial observations of the variables of the hypergraph to discover its structure and approximate its unknown functions. These hypergraphs offer a natural platform for organizing, communicating, and processing computational knowledge. While most scientific problems can be framed as the data-driven discovery of unknown functions in a computational hypergraph whose structure is known (Type 2), many require the data-driven discovery of the structure (connectivity) of the hypergraph itself (Type 3). We introduce an interpretable Gaussian Process (GP) framework for such (Type 3) problems that does not require randomization of the data, nor access to or control over its sampling, nor sparsity of the unknown functions in a known or learned basis. Its polynomial complexity, which contrasts sharply with the super-exponential complexity of causal inference methods, is enabled by the nonlinear analysis of variance capabilities of GPs used as a sensing mechanism.

Introduction

The three levels of complexity of function approximation.

As illustrated in Fig. 7.1.(a-c), Type 1, Type 2 and Type 3 problems can be formulated as completing or discovering hypergraphs where nodes represent variables and edges represent functional dependencies. The graph in Type 1 has only two variables and one unknown function. The graph in Type 2 has multiple variables and (some possibly unknown) functions, and the connectivity of the graph is known. The graph in Type 3 has an unknown connectivity (functional dependencies between variables

may be unknown) and this is the focus of this work. Current methods for solving Type 1 and 2 problems include Deep Learning (DL) methods, which benefit from extensive hardware and software support but have limited guarantees. Despite their prevalence, Type 3 challenges have been largely overlooked due to their inherent complexity. Causal inference methods (Morgan and Winship, 2015; M. Glymour, Pearl, and Jewell, 2016) and probabilistic graphs (Stegle et al., 2010; Lopez-Paz et al., 2015) and sparse regression methods (Doostan and Houman Owhadi, 2011; Brunton, Proctor, and Kutz, 2016), offer potential avenues for addressing Type 3 problems. However, it is important to note that their application to these problems necessitates additional assumptions. Causal inference models, for instance, typically assume randomized data and some level of access to the data generation process or its underlying distributions. Sparse regression methods, on the other hand, rely on the assumption that functional dependencies have a sparse representation within a known basis. In this paper, we do not impose these assumptions, and thus, these particular techniques may not be applicable. Furthermore while the complexity of Bayesian causal inference methods may grow super-exponentially with the number d of variables, the complexity of our method is that of d parallel computations of polynomial complexities bounded between $\mathcal{O}(d)$ (best case) and $\mathcal{O}(d^4)$ (worst case).

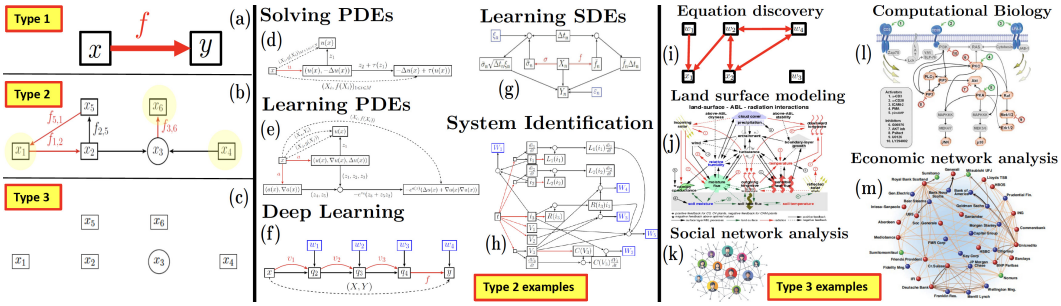


Figure 7.1: The three levels of complexity of function approximation.

Generalizing Gaussian Process methods.

Although Gaussian Process (GP) methods are sometimes perceived as a well-founded but old technology limited to curve fitting (Type 1 problems), they have recently been generalized, beyond Type 1 problems, to an interpretable framework (Computational Graph Completion or CGC (Houman Owhadi, 2022)) for solving Type 2 problems (Houman Owhadi, 2023a; Yifan Chen, Hosseini, et al., 2021c; Batlle, Yifan Chen, et al., 2023; Yifan Chen, Houman Owhadi, and Florian Schäfer, 2024; Darcy et al., 2023; Hamzi, Houman Owhadi, and Kevrekidis, 2023), all while maintaining the simple and transparent theoretical and computational guarantees

of kernel/optimal recovery methods (Micchelli and Rivlin, 1977; H. Owhadi and C. Scovel, 2019). This paper introduces a comprehensive GP framework for solving Type 3 problems, which is interpretable and amenable to analysis. This framework leverages the Uncertainty Quantification (UQ) properties of GP methods, which do not have an immediate natural counterpart in DL methods. It is based on a kernel generalization (Wahba, 2003; Houman Owhadi, Clint Scovel, and Yoo, 2021) of variance-based sensitivity analysis guiding the discovery of the structure of the hypergraph. Here, variables are linked via GPs, and those contributing to the highest data variance unveil the hypergraph's structure. This GP variance decomposition of the data leads to signal-to-noise and a Z-score that can be employed to determine whether a given variable can be approximated as a nonlinear function of a subset of other variables.

The scope of Type 1, 2 and 3 problems.

The scope of Type 1, 2 and 3 problems is immense. Numerical approximation (H. Owhadi and C. Scovel, 2019; H. Owhadi, C. Scovel, and F. Schäfer, 2019; Florian Schäfer, Timothy John Sullivan, and Houman Owhadi, 2021c; Florian Schäfer and Houman Owhadi, 2021), Supervised Learning, and Operator Learning (Hamzi, Maulik, and Houman Owhadi, 2021b; Hamzi and Houman Owhadi, 2021; Florian Schäfer and Houman Owhadi, 2023; Batlle, Darcy, et al., 2023) can all be formulated as **Type 1 problems**, i.e., as approximating unknown functions given (possibly with noisy and infinite/high-dimensional) inputs/output data. The common GP based solution to these problems is to replace the underlying unknown function by a GP and compute its MAP estimator given available data. **Type 2 problems** include (Fig. 7.1.(d-h)) solving and learning (possibly stochastic) ordinary or partial differential equations (Yifan Chen, Hosseini, et al., 2021c; Darcy et al., 2023), Deep Learning (Houman Owhadi, 2023a), dimension reduction, reduced-ordered modeling, system identification (Houman Owhadi, 2022), closure modeling, etc. Indeed, all these problems can be formulated as completing a computational graph (Houman Owhadi, 2022). In this formulation, variables and functions are represented by the nodes and the edges of the graph whose structure corresponds to the functional dependencies between variables. Some of the functions and variables may be unknown, and by completing, we mean approximating the unknown functions (colored in red in Fig. 7.1) given samples from the observed variables. The common GP-based solution to Type 2 problems is to simply replace unknown functions by GPs and compute their MAP/MLE estimators given available data and constraints

imposed by the structure of the graph (Houman Owhadi, 2022). While most problems in Computational Sciences and Engineering (CSE) and Scientific Machine Learning (SciML) can be framed as Type 1 and Type 2 challenges, many problems in science can only be categorized as **Type 3 problems**, i.e., discovering the structure/connectivity of the graph itself from data prior to its completion. Indeed the scope of Type 3 problems extends well beyond Type 2 problems and includes equation discovery (Fig. 7.1.(i)); the modeling of land surface interactions in weather prediction (Fig. 7.1.(j) from (Dirmeyer et al., 2019), discovering possibly hidden functional dependencies between state variables for a finite number of snapshots of those variables); social network analysis (Fig. 7.1.(k) from (Gittell and Ali, 2021), discovering functional dependencies between quantitative markers associated with each individual in situations where the connectivity of the network may be hidden); economic network analysis (Fig. 7.1.(m) from (Schweitzer et al., 2009), discovering functional dependencies between the economic markers of different agents or companies, which is significant to systemic risk analysis); and computational biology (Fig. 7.1.(l) from (Sachs et al., 2005), identifying pathways and interactions between genes from their expression levels).

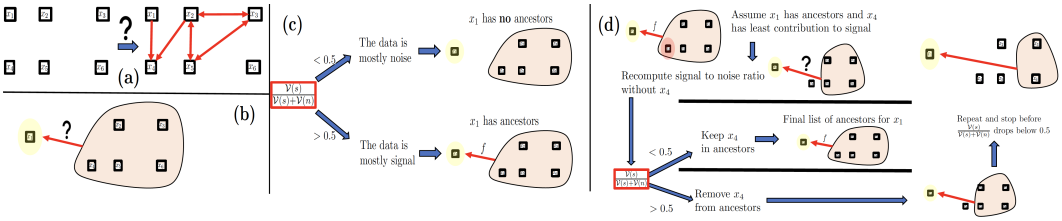


Figure 7.2: Ancestors identification in Type 3 problem.

Overview of the proposed approach for Type 3 problems.

We first present an algorithmic overview of the proposed GP-based approach for Type 3 problems. For ease of presentation, we consider the simple setting of Fig. 7.2.(a) where we are given N samples on the variables $x_1, \ldots, x_6$. After measurements/collection, these variables are normalized to have zero mean and unit variance. Our objective is to uncover the underlying dependencies between them.

A signal-to-noise ratio to decide whether or not a node has ancestors.

Our algorithm's core concept is the identification of ancestors for each node in the graph. Let's explore this idea in the context of a specific node, say x_1, as depicted in Fig. 7.2(b). Determining whether x_1 has ancestors is akin to asking if x_1 can be expressed as a function of $x_2, x_3, \ldots, x_6$. In other words, can we find a function f

(living in a pre-specified space of functions that could be of controlled regularity) such that

$$x_1 \approx f(x_2, \dots, x_6) \,? \tag{7.1}$$

To answer this question we regress f with a centered GP $\xi \sim \mathcal{N}(0, \Gamma)$ whose covariance function Γ is an additive kernel of the form $\Gamma = K_s + \gamma\delta(x - y)$, where K_s is a smoothing kernel, $\gamma > 0$ is a regularization parameter and $\delta(x - y)$ is the white noise covariance operator. This is equivalent to assuming the GP ξ to be the sum of two independent GPs, i.e., $\xi = \xi_s + \xi_n$ where $\xi_s \sim \mathcal{N}(0, K_s)$ is a smoothing/signal GP and $\xi_n \sim \mathcal{N}\big(0, \gamma\delta(x - y)\big)$ is a noise GP. Writing $\mathcal{H}_{K_s}$ for the Reproducing Kernel Hilbert Space (RKHS) induced by the kernel K_s, this is also equivalent to approximating f with a minimizer of

$$\inf_{f \in \mathcal{H}_{K_s}} \|f\|^2_{K_s} + \frac{1}{\gamma}\big\|f(X) - Y\big\|^2_{\mathbb{R}^N}, \tag{7.2}$$

where $\|\cdot\|^2_{\mathbb{R}^N}$ is the Euclidean norm on $\mathbb{R}^N$, X is the input data on f obtained as an $N \times 5$-matrix whose rows X_i are the samples on $x_2, \dots, x_6$, Y is the output data on f obtained as an N-vector whose entries are obtained from the samples on x_1, and $f(X)$ is a N-vector whose entries are the evaluations $f(X_i)$. At the minimum

$$\mathcal{V}(s) := \|f\|^2_{K_s} \tag{7.3}$$

quantifies the data variance explained by the signal GP ξ_s and

$$\mathcal{V}(n) := \frac{1}{\gamma}\big\|f(X) - Y\big\|^2_{\mathbb{R}^N} \tag{7.4}$$

quantifies the data variance explained by the noise GP ξ_n (Houman Owhadi, Clint Scovel, and Yoo, 2021). This allows us to define the signal-to-noise ratio

$$\frac{\mathcal{V}(s)}{\mathcal{V}(s) + \mathcal{V}(n)} \in [0, 1]\,. \tag{7.5}$$

If $\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)} < 0.5$[1], then, as illustrated in Fig. 7.2.(c), we deduce that x_1 has no ancestors, i.e., x_1 cannot be approximated as function of $x_2, \dots, x_6$. Conversely if $\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)} > 0.5$, then, we deduce that x_1 has ancestors, i.e., x_1 can be approximated as function of $x_2, \dots, x_6$.

[1] We will later present a version with a more sophisticated method for pruning, but we keep the 0.5 threshold in this example for simplicity.

Selecting the signal kernel K_s

This process is repeated by selecting the kernel K_s to be linear ($K_s(x, x') = 1 + \beta_1 \sum_i x_i x_i'$), quadratic ($K_s(x, x') = 1+\beta_1 \sum_i x_i x_i' + \beta_2 \sum_{i \leq j} x_i x_j x_i' x_j'$) or fully nonlinear to identify f as linear, quadratic, or nonlinear. In the case of a nonlinear kernel, we employ

$$K_s(x, x') = 1 + \beta_1 \sum_i x_i x_i' + \beta_2 \sum_{i \leq j} x_i x_j x_i' x_j' + \beta_3 \prod_i (1 + k(x_i, x_i')). \tag{7.6}$$

where k is a universal kernel, such as a Gaussian or a Matérn kernel, with all parameters set to 1, and β_i assigned the default value 0.1. We select K_s as the first kernel that surpasses a signal-to-noise ratio of 0.5. If no kernel reaches this threshold, we conclude that x_1 lacks ancestors.

Pruning ancestors based on signal-to-noise ratio.

Once we establish that x_1 has ancestors, the next step is to prune its set of ancestors iteratively. We remove nodes with the least contribution to the signal-to-noise ratio and stop before that ratio drops below 0.5 as illustrated in Fig. 7.2.(d). To describe this, assume that K_s is as in (7.6). Then K_s is an additive kernel that can be decomposed into two parts:

$$K_s = K_1 + K_2\,, \tag{7.7}$$

where $K_1 = 1+\beta_1 \sum_{i \neq 1,2} x_i x_i' + \beta_2 \sum_{i \leq j, i, j \neq 1,2} x_i x_j x_i' x_j' + \beta_3 \prod_{i \neq 1,2} (1+k(x_i, x_i'))$ does not depend on x_2 and $K_2 = K_s - K_1$ depends on x_2. This decomposition allows us to express f as the sum of two components:

$$f = f_1 + f_2\,, \tag{7.8}$$

where f_1 does not depend on x_2, f_2 depends on x_2 and $(f_1, f_2) = \operatorname{argmin}_{(g_1, g_2) \in \mathcal{H}_{K_1} \times \mathcal{H}_{K_2} \text{ s.t. } g_1 + g_2 = f} \|g_1\|^2_{K_1} + \|g_2\|^2_{K_2}$. Furthermore, $\|f\|^2_{K_s} = \|f_1\|^2_{K_1} + \|f_2\|^2_{K_2}$, and $\frac{\|f_2\|^2_{K_1}}{\|f\|^2_{K_s}} \in [0, 1]$ quantifies the contribution of x_2 to the signal data variance. Following the procedure illustrated in Fig. 7.2.(d), if, for example, x_4 is found to have the least contribution to the signal data variance, we recompute the signal-to-noise ratio without x_4 in the set of ancestors for x_1. If that ratio is below 0.5, we do not remove x_4 from the list of ancestors, and x_2, x_3, x_4, x_5, x_6 is the final set of ancestors of x_1. If this ratio remains above 0.5, we proceed with the removal. This iterative process continues, and we stop before the signal-to-noise ratio drops below 0.5 to identify the final list of ancestors of x_1. The most efficient version of our proposed algorithm does not use a threshold of 0.5 on the signal-to-noise ratio to prune ancestors, but it rather employs

an inflection point in the noise-to-signal ratio $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$ as a function of the number q of ancestors (Fig. 7.3.(d)). To put it simply, after ordering the ancestors in decreasing contribution to the signal, the final number q of ancestors is determined as the maximizer of $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q+1) - \frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$.

Computational complexity

We will now present a detailed analysis of the computational demands of the proposed method as a function of the number of variables, denoted as d, and the number of samples, N, pertaining to these variables. In the worst case, the proposed approach necessitates, for each of the d variables: for $i = 1, \ldots, d-1$, regressing a function mapping $d-i$ variables to the variable of interest and performing a mode decomposition, as exemplified in (7.8), to identify the variable with the minimal contribution to the signal. Since these two steps have the same cost, it follows that, in the worst case, the total computational complexity of the proposed method is $O(\mathbf{d^2N^3})$ which corresponds to product of the number of double-looping operations, d^2, and the cost of kernel regression from N samples which, without acceleration, is N^3 (i.e., the cost of inverting a $N \times N$ dense kernel matrix). However, if kernel scalability techniques are utilized, such as when the kernel has a rank k (for example, $k = d$ if the kernel is linear) or is approximated by a kernel of rank k (e.g., via a random feature map), then this worst-case bound can be reduced to $O(\mathbf{d^2Nk^2})$ by reducing the complexity of each regression step from $O(N^3)$ to $O(Nk^2)$. Note that the statistical accuracy of the proposed approach requires that $N > d$ if the dependence of the unknown functions on their inputs is not sparse. Moreover, in the absence of kernel scalability techniques, the worst-case memory footprint of the method is $O(N^2)$ due to the necessity of handling dense kernel matrices. However, once the functional ancestors of each variable are determined, these matrices can be discarded. Consequently, only one such matrix needs to be retained in memory at any given time.

Results

The following examples and experiments illustrate the proposed approach.

The Fermi-Pasta-Ulam-Tsingou system

The Fermi-Pasta-Ulam-Tsingou (FPUT) system (Palais, 1997) is a prototypical chaotic dynamical system. It is composed of M masses indexed by $j \in \{0, \ldots, M-1\}$ with equilibrium position jh with $h = 1/M$. Each mass is tethered to its two adjacent masses by a nonlinear spring, and the displacement of the mass x_j adheres to the

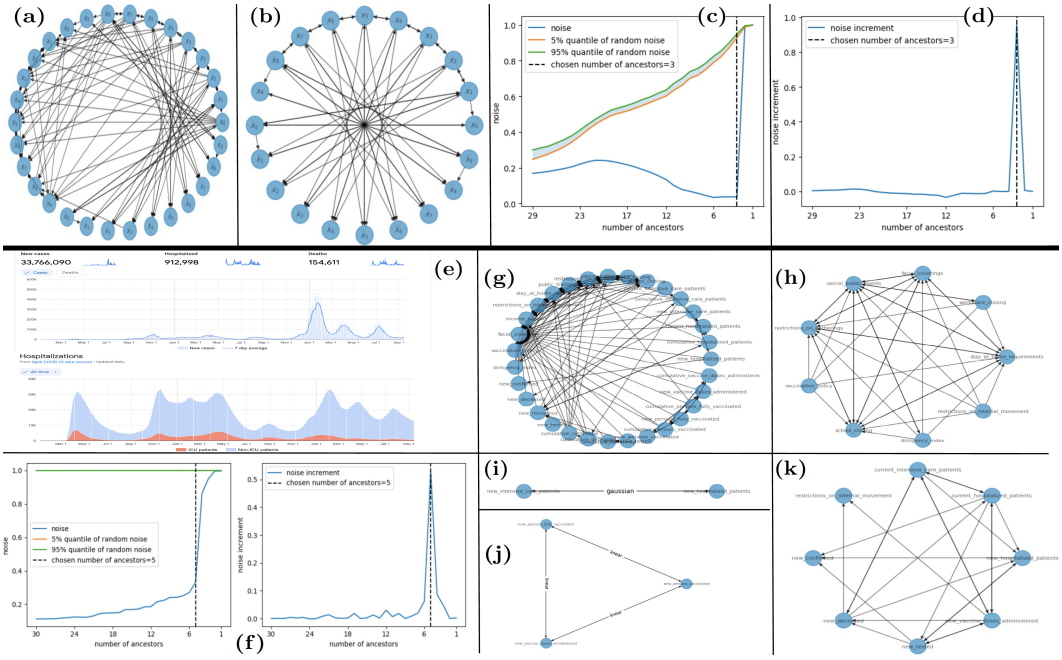


Figure 7.3: (a-d) The Fermi-Pasta-Ulam-Tsingou system. (e-k) The Google Covid 19 open data.

equation:

$$\ddot{x}_j = \frac{c^2}{h^2}(x_{j+1} + x_{j-1} - 2x_j)\left(1 + \alpha(x_{j+1} - x_{j-1})\right), \tag{7.9}$$

where $\alpha(x) = x^2$, $c = 1$ and $M = 10$. We use fixed boundary conditions by adding two more masses, with $x_{-1} = x_M = 0$. We take a total of 1000 snapshots from multiple trajectories and the observed variables are the positions, velocities, and accelerations of all the underlying masses. In the graph discovery phase, every other node is initially deemed a potential ancestor for a specified node of interest. We then proceed to iteratively remove the node with the least signal contribution. The step resulting in the largest surge in the noise-to-signal ratio is inferred as one eliminating a crucial ancestor, thereby pinpointing the final ancestor set. Fig. 7.3.(c) shows a plot of the noise-to-signal ratio $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$ as a function of the number q of proposed ancestors for the variable $\ddot{x}_7$ and with Z-test quantiles (in the absence of signal, the noise-to-signal ratio should fall within the shaded area with probability 0.9). Removing a node essential to the equation of interest causes the noise-to-signal ratio to markedly jump from approximately 25% to 99%. Fig. 7.3.(d) shows a plot of the noise-to-signal ratio increments $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q) - \frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q-1)$ as a function of the number q of ancestors for the variable $\ddot{x}_7$. Note that the increase in the noise-to-signal ratio is significantly higher compared to previous removals when an essential node was removed. Therefore, while solely relying on a fixed threshold to decide

when to cease the removals might prove challenging, evaluating the increments in noise-to-signal ratios offers a clear guideline for efficiently and reliably pruning ancestors. The recovered full graph, depicted in Fig. 7.3.(a), is remarkably accurate despite the nonlinear nature of the model and the fact that our prior only encodes that the nonlinearity is smooth. Therefore, our algorithm does not require a dictionary or extensive knowledge of the structure of the unknown functions. Notably, velocity variables are accurately identified as non-essential and omitted from the ancestors of position and acceleration variables. Fig. 7.3.(b), which omits velocity variables for clarity, further elucidates the accurate recovery of dependencies. The dependencies are the simplest and clearest possible. They match exactly those of the original equations except for the boundary particles for which we recover valid equivalent equations.

The Google Covid 19 open data.

Consider the COVID-19 data from Google[2]. We focus on a single country, France, to ensure consistency in the data and avoid considering cross-border variations that are not directly reflected in the data. We select 31 variables that describe the state of the country during the pandemic, spanning over 500 data points, with each data point corresponding to a single day. These variables are categorized as the following datasets: (1) Epidemiology dataset: Includes quantities such as new infections, cumulative deaths, etc. (2) Hospital dataset: Provides information on the number of admitted patients, patients in intensive care, etc. (3) Vaccine dataset: Indicates the number of vaccinated individuals, etc. (4) Policy dataset: Consists of indicators related to government responses, such as school closures or lockdown measures, etc. Some of these variables are illustrated in Fig. 7.3.(e). The problem is then to analyze this data and identify possible hidden functional relations between these variables. Fig. 7.3.(f) shows the noise-to-signal ratio (and its increments) as function of the number of ancestors of the "cumulative number of hospitalized patients" variable. Even for this real dataset, the proposed approach gives a clear signal for stopping the pruning process. Fig. 7.3.(g) shows the full recovered graph, which is highly clustered. Fig. 7.3.(h) shows the cluster corresponding to the variable "schools closing" revealing that the government either implemented multiple restrictive measures simultaneously or lifted them in unison (except for mask mandates that were on the verge of being identified as noise). The vaccination cluster (Fig. 7.3.(j)) reveals a linear relationship between variables (signaling redundant information) and the

[2]The dataset can be accessed here

hospitalization cluster (Fig. 7.3.(i)) reveals a nonlinear one. Eliminating redundant nodes leads to the sparse graph shown in Fig. 7.3.(k), which is interpretable and amenable to (both quantitative and qualitative) analysis,

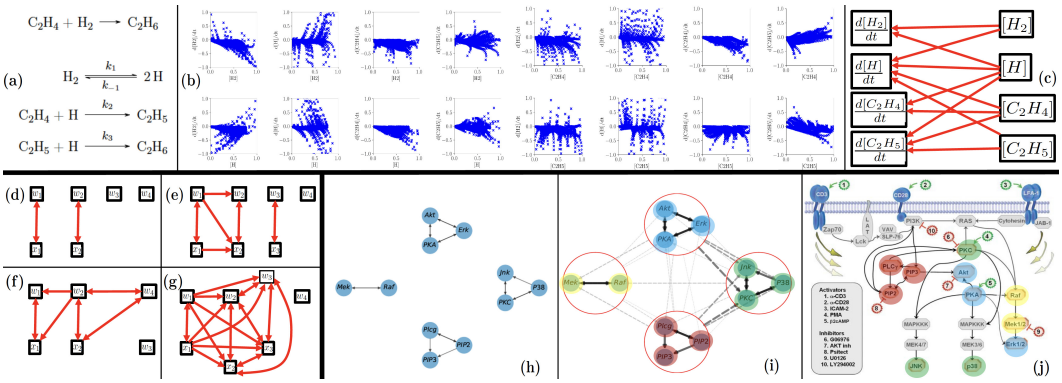


Figure 7.4: (a-c) Chemical reaction network. (d-g) Algebraic equations. (h-j) Cell signaling network.

Chemical reaction network.

In this example, we consider the recovery of a chemical reaction network from concentration snapshots. The reaction network, illustrated in Fig. 7.4.(a) is that of the hydrogenation of ethylene (C_2H_4) into ethane (C_2H_6). The problem is that of recovering the underlying chemical reaction network from snapshots (illustrated in Fig. 7.4.(b)) of concentrations $[H_2]$, $[H]$, $[C_2H_4]$, and $[C_2H_5]$ and their time derivatives. $\frac{d[H_2]}{dt}$, $\frac{d[H]}{dt}$, $\frac{d[C_2H_4]}{dt}$, and $\frac{d[C_2H_5]}{dt}$. The proposed approach leads to a perfect recovery of the computational graph (shown in Fig. 7.4.(c)) and a correct identification of quadratic functional dependencies between variables.

Algebraic equations.

Fig. 7.4.(a-d) illustrate the application of the proposed approach to the recovery of functional dependencies from data satisfying hidden algebraic equations. In all these examples, we have $d = 6$ or $d = 7$ variables and $N = 1000$ samples from those variables. For $d = 6$ the variables are $w_1, w_2, w_3, w_4, x_1, x_2$. For $d = 7$ the variables are $w_1, w_2, w_3, w_4, x_1, x_2, x_3$. The samples from the variables w_1 to w_4 are i.i.d. $\mathbb{N}(0, 1)$ random variables, and the samples from x_1, x_2 (and x_3 for $d = 7$) are functionally dependent on the other variables. In the first example, $d = 6$ and the samples from x_1 and x_2 satisfy the equations $x_1 = w_1$ and $x_2 = w_2$. The algorithm selects the linear kernel and Fig. 7.4.(a) shows the recovered graph (which is exact). In the second example, $d = 7$ and the samples from x_1, x_2 and x_3 satisfy the equations $x_1 = w_1, x_2 = x_1^2 + 1 + 0.1w_2$, and $x_3 = w_3$. The algorithm selects the quadratic kernel

and Fig. 7.4.(b) shows the recovered graph (which is exact). Even though x_2 can trace back its origin to either x_1 and w_2 or w_1 and w_2, the algorithm recognizes x_1, w_1, and w_2 as its ancestors underscoring the importance of eliminating redundant variables when aiming at deriving the sparsest graph. In the third example, $d = 6$ and the samples from x_1 and x_2 satisfy the equations $x_1 = w_1 w_2$ and $x_2 = w_2 \sin(w_4)$. The algorithm selects the nonlinear kernel and Fig. 7.4.(c) shows the recovered graph (which is exact). In the fourth example, $d = 7$ and the samples from x_1, x_2 and x_3 satisfy the equations $x_1 = w_1$, $x_2 = x_1^3 + 1 + 0.1w_2$ and $x_3 = (x_1 + 2)^3 + 0.1w_3$. Although these equations appear to be cubic, the algorithm correctly selects the quadratic kernel and makes an exact recovery of the graph shown in Fig. 7.4. (d) revealing hidden quadratic dependencies between variables.

Cell signaling network

Next, we apply the proposed framework to the example illustrated in Fig. 7.1.(l) from (Sachs et al., 2005) and discover a hierarchy of functional dependencies in biological cellular signaling networks. We use single-cell data consisting of the $d = 11$ phosphoproteins and phospholipids levels in the human immune system T-cells that were measured using flow cytometry. This dataset was studied from a probabilistic modeling perspective in previous works. While (Sachs et al., 2005) learned a directed acyclic graph to encode causal dependencies, (Friedman, Hastie, and Tibshirani, 2008) learned an undirected graph of conditional independencies between the d molecule levels by assuming the underlying data follows a multivariate Gaussian distribution. The latter analysis encodes acyclic dependencies but does not identify directions. In this work, we aim to identify the functional dependencies without imposing strong distributional assumptions on the data. We simply use $N = 2,000$ samples chosen uniformly at random from the dataset consisting of 11 proteins and 7446 samples of their expressions. We apply the algorithm in two stages. The first stage of the algorithm uses only linear and quadratic kernels and recovers the graph shown in Fig. 7.4.(h). It consists of four disconnected clusters where the molecule levels in each cluster are closely related by linear or quadratic dependencies (all connections are linear except for the connection between Akt and PKA, which is quadratic). These edges match a subset of the edges found in the gold standard model identified in (Sachs et al., 2005). With perfect noiseless dependencies, one can define constraints that reduce the total number of variables in the system. Second, we learn the connections between groups of variables within each cluster with nonlinear kernels and obtain the graph shown in Fig. 7.4.(i) in which

solid arrows indicate strong intra-cluster connections identified in the first level, and dashed lines indicate weaker connections between nodes and clusters identified in the second level. The width and grayscale intensities of each edge correspond to its signal-to-noise ratio. We emphasize that while causal graph recovery methods rely on the control of the sampling of the underlying variables (i.e., the simultaneous measurement of multiple phosphorylated protein and phospholipid components in thousands of individual primary human immune system cells, and perturbing these cells with molecular interventions), the reconstruction obtained by our method did not use this information and recovered functional dependencies rather than causal dependencies. Interestingly, the information recovered through our method appears to complement and enhance the findings presented in (Sachs et al., 2005) (e.g., the linear and noiseless dependencies between variables in the JNK cluster is not something that could easily be inferred from the graph produced in (Sachs et al., 2005) shown in Fig. 7.1.(j) where we have colored the clusters for comparison).

Comparisons. Using the expected graph reported in (Sachs et al., 2005) as the ground truth (acknowledging that it may not be entirely accurate), we compare the edges our approach incrementally added to the true graph. Figure 7.5.(a) reports the number of additional edges that have been added and are not present in the ground truth (false positives) and edges removed that are present in the ground truth graph (false negatives). The added edges are based on the two-stage procedure described above, where we first add the ten intra-cluster connections, followed by inter-cluster connections. Edges are added in decreasing order of signal-to-noise ratio, starting with the strongest. In the reported results, we do not account for the recovery of the direction of ground-truth edges. We note that, up to direction, all intra-cluster connections, along with the inter-cluster connections with the strongest signals are found in the ground truth graph, leading to the initial decrease in false negatives with only one false positive edge (the linear connection P38 $\rightarrow$ Jnk that is not reported in the true graph). With the addition of the remaining (possibly non-spurious) edges, the number of false negatives drops to one, having recovered all edges, except for the one between PKC and Raf, which is identified to be statistically non-informative in our approach.

A large-scale chemical reaction network: the BCR reaction benchmark

Lastly, we stress-test the scalability of our approach by applying it to a large-scale chemical reaction network: the BCR reaction benchmark from (Loman et al., 2023), which encompasses 1122 species. The dataset comprises 2,400 snapshots of species

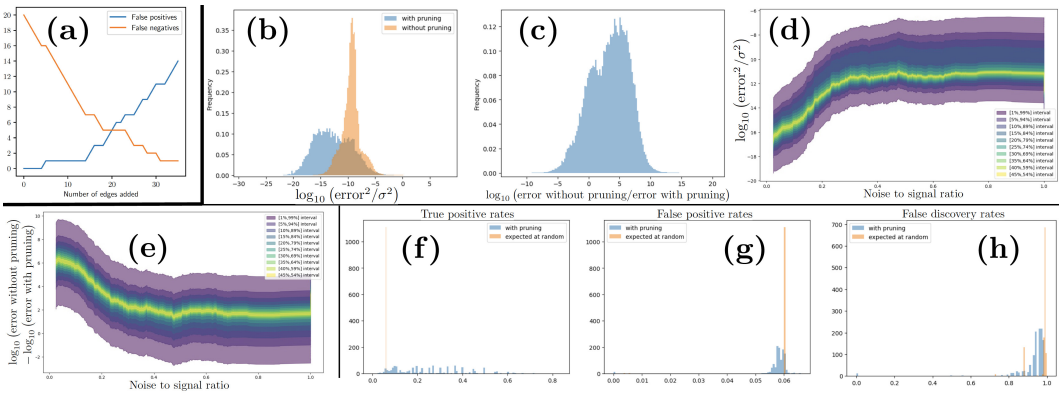


Figure 7.5: (a) Cell signaling network comparisons. (b-h) The BCR reaction benchmark.

concentrations and their corresponding time derivatives. We leveraged JAX's inherent parallelization capabilities (Bradbury et al., 2018) to accelerate our computations, allowing for the simultaneous pruning of multiple nodes while abstracting the complexity of parallel execution. While the scaling with respect to the number of data points is straightforward, scaling with the number of variables introduces a trade-off between computational speed and memory footprint. Specifically, the process of identifying the ancestors of various nodes can be expedited by storing a large array for all nodes. Using a DGX workstation equipped with four Nvidia V100 GPUs, each with 32GB of memory, pruning 190 nodes took approximately three days, projecting a total experiment duration of around one month. Nonetheless, we can mitigate this computational burden by optimizing the computation of terms of the form $y^T K y$ for the specific quadratic kernel identified for this example. We include the details of such optimization in the supplementary material. By implementing this optimization, the duration of the entire experiment was reduced to just one hour.

In the first experiment, we simulated five trajectories of the associated system of ODEs, recording 1000 snapshots per trajectory. Out of these 5000 snapshots, 2,400 were randomly selected as training data, and 2600 as testing data. Writing TP, TN, FP and FN for True/False Positives/Negatives and using the metrics True Positive Rate (TPR=TP/(TP+FN)), False Positive Rate (FPR=FP/(FP+TN)), and False Discovery Rate (FDR=FP/(TP+FP)), we observed a TPR of 39.9%, an FPR of 16.4%, and an FDR of 97.2% (indicating that 97.2% of predicted positives are false). This high FDR can be attributed to the limited exploration of the full variable range—1,122 in total—by the five trajectories. The trajectories explored a subset of the possible space

(near a limit cycle attractor), which led to the recovery of functional dependencies that represent both the chemical reactions and the specific subspace visited. Furthermore, with 1,122 variables, the 630,003 coefficients of the underlying quadratic equations are vastly under-determined with only 2,400 data points. Despite the high FDR in the recovered graph, as illustrated in Fig. 7.5.(b-c), the CHD pruning process vastly improves the accuracy (by orders of magnitude) of the estimated functions on the 2600 unseen snapshots by reducing the dimension of the regression problem whenever possible. We denote y_i as an observed data point, σ^2 as the variance of the observed data, $\widehat{y}_i$ for a predicted data point without pruning, and $\bar{y}_i$ for a predicted data point post-pruning. Fig. 7.5.(b) illustrates the histogram of the log-normalized squared errors before and after pruning, expressed as $\log_{10}\left(|y_i - \widehat{y}_i|^2/\sigma^2\right)$ and $\log_{10}\left(|y_i - \bar{y}_i|^2/\sigma^2\right)$. The 99th percentile of the normalized squared error is less than 10^{-2} for all species. Fig. 7.5.(c) displays the histogram of the log-normalized squared error improvements due to pruning, calculated as $\log_{10}\left(|y_i - \widehat{y}_i|^2/|y_i - \bar{y}_i|^2\right)$. Fig. 7.5.(d-e) display the quantiles of the histograms post-pruning, conditioned on the noise-to-signal ratio observed at the final pruning step. These plots reveal a clear trend: a higher noise-to-signal ratio at the time of pruning correlates with increased error and diminished improvements in accuracy.

In a second experiment, we formed the data by randomly sampling concentrations uniformly in $[0, 1]$ (independently across species and snapshots) and recorded the resulting time derivatives. While this sampling increased the variability of the 2,400 snapshots, the model remained vastly underdetermined. The noise-to-signal and bootstrapped (Z-test) ratios remained close to 0.5, suggesting insufficient data for statistically significant variable importance assessments. Nonetheless, as depicted in Fig. 7.5.(f-h), significant insights can still be gleaned from the activations, showing notable improvements when comparing the histograms of the values of TPR, FPR, and FDR obtained with pruning based on these ratios and pruning at random. This analysis reveals that even with high dimensionality and scarce data, between 10% and 80% of the true ancestors can still be accurately identified.

Discussions

Limitations

In its present form, the proposed approach is limited by several factors. (1) Without access to the sampling of the data, the direction of some edges may not be identifiable. For instance the functional relationship $x - 2y = 0$ can be represented as both $y = 2x$ ($x \to y$) and $x = y/2$ ($y \to x$). (2) It assumes an additive noise W

on the functional relationship $y = f(x) + W$ between the variables x and y. In a fully probabilistic setting, this structure may be non-additive, i.e., of the form $y = f(x, W)$, which implies discovering a general transition kernel, i.e., a non-Gaussian generative model. Although our method achieves polynomial complexity, in settings where one has access to the distribution of the data, the price to pay, when compared with information-theoretic methods, is a reduction in generality imposed by the stronger assumption made on the data-generating process. Furthermore, the price to pay for the weaker data requirements (i.e., the absence of interventional data) is that our method recovers functional relationships rather than causal ones or conditional dependencies. (3) If the (noisy) functional relationship $y = f(x) + W$ is associated with a non-regular (e.g., discontinuous) function f then the kernels discussed above (linear, quadratic, and fully nonlinear) will be misspecified and may lead to false negatives. The kernel selection and hyperparameter tuning problems in misspecified settings require further work. (4) As demonstrated in the BCR reaction application, while the method scales well computationally with an increase in the number of variables, it may still be impacted by the curse of dimensionality. This occurs particularly if the dataset only covers a limited subset of the full range of variable values. Given the results displayed in Fig. 7.5.(b-h) we suspect that this impact could be mitigated by adopting more advanced strategies in place of our current top-down pruning method. Such strategies could involve grouping variables and integrating both top-down and bottom-up iterative approaches.

Conclusions

We have developed a comprehensive Gaussian Process framework for solving Type 3 (hypergraph discovery) problems, which is interpretable and amenable to analysis. The breadth and complexity of Type 3 problems significantly surpass those encountered in Type 2 (hypergraph completion), and the initial numerical examples we present serve as a motivation for the scope of Type 3 problems and the broader applications made possible by this approach. Our proposed algorithm is designed to be fully autonomous, yet it offers the flexibility for manual adjustments to refine the graph's structure recovery. We emphasize that our proposed approach is not intended to supplant causal inference methods (Pearl, 2009); see Methods for a complete overview. Instead, it aims to incorporate a distinct kind of information into the graph's structure, namely, the functional dependencies among variables rather than their causal relationships. Additionally, our method eliminates the need for a predetermined ordering of variables, a common requirement in acyclic probabilistic

models where determining an optimal order is an NP-hard problem usually tackled using heuristic approaches. Furthermore, our approach can actually be utilized to generate such an ordering by quantifying the strength of the connections it recovers. The Uncertainty Quantification properties of the underlying Gaussian Processes are integral to the method and could also be employed to quantify uncertainties in the structure of the recovered graph. We also observe that forming clusters from highly interdependent variables helps to obtain a sparser graph. Additionally, the precision of the pruning process is enhanced by avoiding the division of node activation within the cluster among its separate constituents. We employed this strategy in the recovery of the gene expression graph in Fig. 7.4.(i). Given the polynomial complexity of our method, promising avenues for future work include applications to large datasets in genomics and in systems biology, particularly in the reconstruction and intervention of metabolic pathways. These applications benefit from the ability to handle large-scale datasets efficiently, enabling the analysis of complex biological networks.

Data availability

The data in the paper and the Supplementary Information are available in the Github repository of the paper.

Code availability

The code for the algorithm and its application to various examples are available for download (and as as an installable python library/package) in the Github repository of the paper.

Online content

Supplementary Information is available for this paper.

Acknowledgements

HO, TB, PB, XY, and RB acknowledge support from the Air Force Office of Scientific Research under MURI award number FA9550-20-1-0358 (Machine Learning and Physics-Based Modeling and Simulation). Additionally, HO, TB, and PB acknowledge support by the Department of Energy under award number DE-SC0023163 (SEA-CROGS: Scalable, Efficient and Accelerated Causal Reasoning Operators, Graphs and Spikes for Earth and Embedded Systems) and by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. HO and PB further acknowledge

support by Beyond Limits (Learning Optimal Models) through CAST (The Caltech Center for Autonomous Systems and Technologies). TB acknowledges support from the Kortschak Scholar Fellowship Program. NR acknowledges support from the JPL Researchers on Campus (JROC) program. HO is grateful for the Department of Defense Vannevar Bush Faculty Fellowship. We are grateful to two referees for their insightful comments and valuable suggestions.

Supplementary information

This supplementary document provides an overview of refinements and generalizations on our proposed approach (Sec. 7.1) detailed in subsequent sections. It includes a summary of the principal components of our algorithm (Sec. 7.2). It includes a reminder on Type 2 problems (Sec. 7.3) and their common GP-based solutions. It discusses the hardness of Type 3 problems, presents an overview of causal inference methods, and a well-posed formulation of Type 3 problems (Sec. 7.4). Additionally, this document offers an in-depth description of our developed GP-based solution specifically designed for Type 3 problems (Section 7.5), along with the corresponding algorithmic pseudo-codes (Section 7.6). It also includes an analysis of the signal-to-noise ratio (SNR) test that is integral to our method (Section 7.7), and furnishes supplementary details concerning the examples discussed in the main manuscript (Section 7.8).

7.1 Additional details on our proposed approach.

The efficacy of our proposed approach is enhanced through a series of refinements (implemented in all our examples), which are summarized below and detailed in sections 7.5, 7.6, and 7.7.

Ancestor pruning.

As discussed earlier, rather than using a threshold on the signal-to-noise ratio to prune ancestors, we order the ancestors in decreasing contribution to the signal, the final number q of ancestors is determined as the maximizer of noise to signal ratio increment $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q+1) - \frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$.

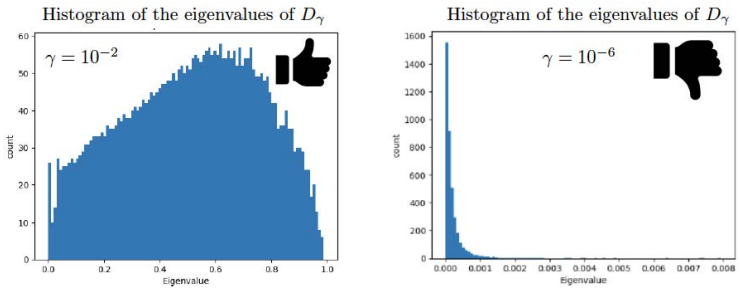


Figure 7.6: Histogram of the eigenvalues of D_γ=(7.11) for $\gamma = 10^{-2}$ (good choice) and $\gamma = 10^{-6}$ (bad choice).

Parameter Selection.

The choice of the parameter γ in (7.2) is a critical aspect of our proposed approach. We provide a structured approach for selecting γ based on the characteristics of

the kernel matrix K_s. Specifically, when K_s is derived from a finite-dimensional feature map ψ (i.e., when $K_s(x, x') := \psi(x)^T \psi(x')$ where the range of ψ is finite-dimensional) and the data cannot be interpolated exactly with K_s (the dimension of the range of ψ is smaller than the number of data points), we employ the regression residual to determine γ as follows:

$$\gamma = \min_v \left\| v^T \psi(X) - Y \right\|^2_{\mathbb{R}^N} . \tag{7.10}$$

Write $K_s(X, X)$ for the $N \times N$ matrix with entries $K_s(X_i, X_j)$. Alternatively, when the data can be interpolated exactly with K_s (e.g., when K_s is a universal kernel), we select γ (see Fig. 7.6) by maximizing the variance of the eigenvalue histogram of the $N \times N$ matrix

$$D_\gamma := \gamma \big(K_s(X, X) + \gamma I \big)^{-1} , \tag{7.11}$$

whose eigenvalues are bounded between 0 and 1 and converge towards 0 as $\gamma \downarrow 0$ and towards 1 as $\gamma \uparrow \infty$. We can also select γ as the median of the eigenvalues of D_γ.

Z-test quantiles.

The noise-to-signal ratio $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}$ associated with (7.2) admits the representer formula $\frac{Y^T D_\gamma^2 Y}{Y^T D_\gamma Y}$. Therefore if the data is only composed of noise (if $Y \sim \sigma^2 Z$ where Z is a random vector with i.i.d. $\mathcal{N}(0, 1)$ entries), then the distribution of the noise-to-signal ratio follows that of the random variable

$$B := \frac{Z^T D_\gamma^2 Z}{Z^T D_\gamma Z} . \tag{7.12}$$

Therefore, the quantiles of B can be used as an interval of confidence on the noise-to-signal ratio if $Y \sim \sigma^2 Z$. Fig. 7.3.(c) shows these Z-test quantiles (in the absence of signal, the noise-to-signal ratio should fall within the shaded area with probability 0.9).

Generalizations on our proposed approach.

Complexity Reduction with Kernel PCA Variant.

Write K for the kernel associated with the RKHS $\mathcal{H}$ in Problem 4. We use a variant of Kernel PCA (Mika et al., 1998) to significantly reduces the computational complexity of our proposed method, making it primarily dependent on the number of principal nonlinear components in the kernel matrix $K(X, X)$ (the $N \times N$ matrix with entries $K(X_i, X_j)$) rather than the number of data points. To describe this

write $\lambda_1 \geq \cdots \geq \lambda_r > 0$ for the nonzero eigenvalues of $K(X, X)$ indexed in decreasing order and write $\alpha_{\cdot,i}$ for the corresponding unit-normalized eigenvectors, i.e. $K(X, X)\alpha_{\cdot,i} = \lambda_i \alpha_{\cdot,i}$. Then $|f(X)|^2 = |f(\phi)|^2$, where $f(\phi)$ is the r vector with entries $f(\phi_i) := \sum_{s=1}^{N} f(X_s)\alpha_{s,i}$. Furthermore, writing $r' \leq r$ for the smallest index i such that $\lambda_i/\lambda_1 < \epsilon$ where $\epsilon > 0$ is some small threshold, the complexity of the problem can be further reduced (as in PCA) by truncating $f(\phi)$ to $f(\phi') = (f(\phi_1), \ldots, f(\phi_{r'}))$ and approximating $\mathcal{F}$ with the space of functions $f \in \mathcal{H}$ such that $|f(\phi')|^2 \approx 0$.

Generalizing Descendants and Ancestors with Kernel Mode Decomposition.

We can extend the concept of descendants and ancestors to cover more complex functional dependencies between variables, including implicit ones. This generalization is achieved through a Kernel-based adaptation of Row Echelon Form Reduction (REFR), initially designed for affine systems, and leveraging the principles of Kernel Mode Decomposition (Houman Owhadi, Clint Scovel, and Yoo, 2021). To describe the connection with REFR consider the example in which $\mathcal{M}$ is the manifold of $\mathbb{R}^3$ defined by the affine equations $x_1 + x_2 + 3x_3 - 2 = 0$ and $x_1 - x_2 + x_3 = 0$, which is equivalent to selecting $\mathcal{F} = \text{span}\{f_1, f_2\}$ with $f_1(x) = x_1 + x_2 + 3x_3 - 2$ and $f_2(x) = x_1 - x_2 + x_3$ in the problem formulation 4. Then, irrespective of how we recover the manifold from data, the hypergraph representation of that manifold is equivalent to the row echelon form reduction of the affine system, and this representation and this reduction require a possibly arbitrary choice of free and dependent variables. So, for instance, if we declare x_3 to be the free variables and x_1 and x_2 to be the dependent variables, then we can represent the manifold via the equations $x_1 = 1 - 2x_3$ and $x_2 = 1 - x_3$ which have the hypergraph representation depicted in Fig. 7.11.(b). To describe the kernel generalization of REFR assume that the kernel K can be decomposed as the additive kernel

$$K = K_a + K_s + K_z\,, \tag{7.13}$$

and write $\mathcal{H}_a$, $\mathcal{H}_s$, and $\mathcal{H}_z$ for the RKHS induced by the kernels K_a, K_s, K_z. Then a function $f \in \mathcal{H}$ can be decomposed as $f = f_a + f_s + f_z$ with $(f_a, f_s, f_z) \in \mathcal{H}_a \times \mathcal{H}_s \times \mathcal{H}_z$. Then, generalizing REFR we can approximate the manifold $\mathcal{M}$ via a manifold parametrized by equations of the form

$$f_a + f_s + f_z = 0 \Leftrightarrow g_a = f_s\,, \tag{7.14}$$

where $f_a = -g_a$ and g_a is a given function in $\mathcal{H}_a$ representing a dependent mode, $f_z = 0$ represents a zero mode, and $f_s \in \mathcal{H}_s$ is identified (regularized) as the

minimizer of the following variational problem

$$\min_{f_s \in \mathcal{H}_s} \|f_s\|_{K_s}^2 + \frac{1}{\gamma}\left|(-g_a + f_s)(\phi)\right|^2 . \tag{7.15}$$

Taking $g_a(x) = x_1$ and $\mathcal{H}_s + \mathcal{H}_z$ to be a space of functions that does not depend on x_1 recovers our initial example (7.1) (with the pruning process encoded into the selection of $\mathcal{H}_z$). This generalization is motivated by its potential to recover implicit equations. For example, consider the implicit equation $x_1^2 + x_2^2 = 1$, which can be retrieved by setting the mode of interest to be $g_a(x) = x_1^2$ and allowing f_s to depend only on the variable x_2.

7.2 Algorithm Overview for Type 3 problems: An Informal Summary

In this section, we provide an accessible overview of our algorithm's key components, which are further detailed in Algorithms 11 and 12 in Section 7.6. Our method focuses on determining the edges within a hypergraph. To achieve this, we consider each node individually, finding its ancestors and establishing edges from these ancestors to the node in question. While we present the algorithm for a single node, it can be applied iteratively to all nodes within the graph.

Algorithm for finding the ancestors of a node:

1. **Initialization:** We start by assuming that all other nodes are potential ancestors of the current node.

2. **Selecting a Kernel**: We choose a kernel function, such as linear, quadratic, or fully nonlinear kernels (refer to Example 7.5.4). The kernel selection process is analogous to the subsequent pruning steps, involving the determination of a parameter γ, regression analysis, and evaluation based on signal-to-noise ratios.

    - **Kernel Selection Method:** The choice of kernel follows a process similar to the subsequent pruning steps, including γ selection, regression analysis, and signal-to-noise ratio evaluation.
    - **Low Signal-to-Noise Ratio for All Kernels:** If the signal-to-noise ratio is insufficient for all possible kernels, the algorithm terminates, indicating that the node has no ancestors.

3. **Pruning Process:** While there are potential ancestors left to consider (details in Section 7.5):

a) **Identify the Least Important Ancestor:** Ancestors are ranked based on their contribution to the signal (see Sec. 7.5).

b) **Noise prior:** Determine the value of γ (see Section 7.7).

c) **Regression Analysis:** Predict the node's value using the current set of ancestors, excluding the least active one (i.e., the one contributing the least to the signal). We employ Kernel Ridge Regression with the selected kernel function and parameter γ (see Sec. 7.5 and 7.5).

d) **Evaluate Removal:** Compute the regression signal-to-noise ratio (see Sec. 7.5 and 7.7):

- **Low Signal-to-Noise Ratio:** If the signal-to-noise ratio falls below a certain threshold, terminate the algorithm and return the current set of ancestors (see Section 7.5).
- **Adequate Signal-to-Noise Ratio:** If the signal-to-noise ratio is sufficient, remove the least active ancestor and continue the pruning process.

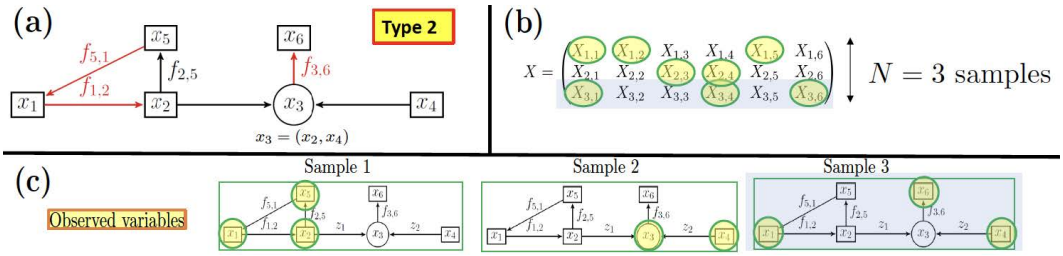


Figure 7.7: Formal description of Type 2 problems.

7.3 Type 2 problems: Formal description and GP-based Computational Graph Completion

Formal description of Type 2 problems

Consider a computational graph (as illustrated in Fig. 7.7.(a)) where nodes represent variables and edges are directed and they represent functions. These functions may be known or unknown. In Fig. 7.7.(a), edges associated with unknown functions ($f_{5,1}$, $f_{1,2}$, $f_{3,6}$) are colored in red, and those associated with known functions ($f_{2,5}$) are colored in black. Round nodes are utilized to symbolize variables, which are derived from the concatenation of other variables (e.g, in Fig. 7.7.(a), $x_3 = (x_2, x_4)$). Therefore, the underlying graph is, in fact, a hypergraph where functions may map groups of variables to other groups of variables, and we use round nodes to

illustrate the grouping step. Given partial observations derived from N samples of the graph's variables, we introduce a problem, termed a Type 2 problem, focused on approximating all unobserved variables and unknown functions. Using Fig. 7.7.(a)-(b) as an illustration we call a vector $(X_{s,1}, \ldots, X_{s,6})$ a sample from the graph if its entries are variables satisfying the functional dependencies imposed by the structure of the graph (i.e., $X_{s,1} = f_{5,1}(X_{s,5})$, $X_{s,2} = f_{1,2}(X_{2,s})$, $X_{s,3} = (X_{s,2}, X_{s,4})$, $X_{s,5} = f_{s,5}(X_{s,s})$, and $X_{s,6} = f_{3,6}(X_{s,3})$. These samples can be seen as the rows of given matrix X illustrated in Fig. 7.7.(b) for $N = 3$. By partial observations, we mean that only a subset of the entries of each row may be observed, as illustrated in Fig. 7.7.(b)-(c). Note that a Type 2 problem combines a regression problem (approximating the unknown functions of the graph) with a matrix completion/data imputation problem (approximating the unobserved entries of the matrix X).

Reminder on Computational Graph Completion for Type 2 problems

Within the context of Sec. 7.3, the proposed GP solution to Type 2 problems is to simply replace unknown functions by GPs and compute their Maximum A Posteriori (MAP)/Maximum Likelihood Estimation (MLE) estimators given available data and constraints imposed by the structure of the graph. Taking into account the example depicted in Fig. 7.7, and substituting $f_{5,1}$, $f_{1,2}$, and $f_{3,6}$ with independent GPs, each with kernels K, G, and Γ respectively, the objective of this MAP solution becomes minimizing $\|f_{5,1}\|_K^2 + \|f_{1,2}\|_G^2 + \|f_{3,6}\|_\Gamma^2$ (writing $\|f\|_K$ for the RKHS norm of f induced by the kernel K) subject to the constraints imposed by the data and the functional dependencies encoded into the structure of the graph.

A system identification example.

In order to exemplify Computational Graphical Completion (CGC), consider the system identification problem depicted in Fig. 7.8, sourced from (Houman Owhadi, 2022). Our objective is to identify a nonlinear electric circuit, as illustrated in Fig. 7.8.(a), from scarce measurement data. The nonlinearity of the circuit emanates from the resistance, capacitance, and inductances, which are nonlinear functions of currents and voltages, as shown in Fig. 7.8.(b). Assuming these functions to be unknown, along with all currents and voltages as unknown time-dependent functions, we operate the circuit between times 0 and 10. Measurements of a subset of variables, representing the system's state, are taken at times $t_s = s/10$ for $s \in 0, \ldots, 99$. Given these measurements, the challenge arises in approximating all unknown functions that define currents and voltages as time functions, capacitance as a voltage function,

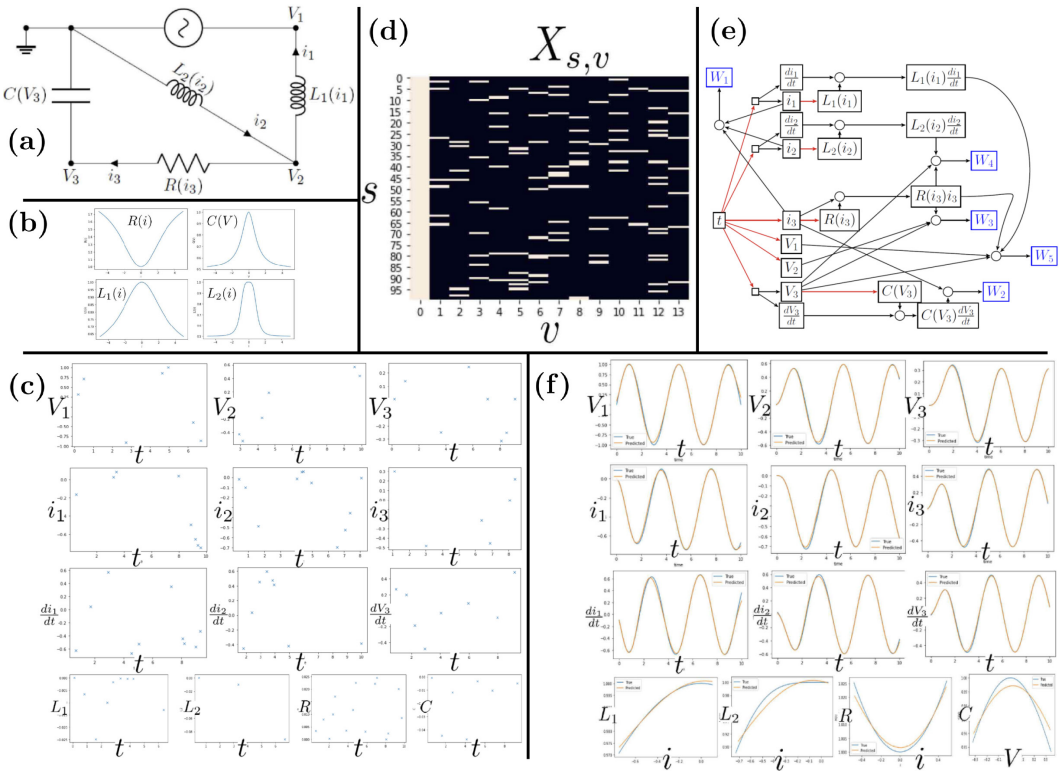


Figure 7.8: (a) Electric circuit. (b) Resistance, capacitance, and inductances are nonlinear functions of currents and voltages (c) Measurements. (d) Kirchhoff's circuit laws. (e) The computational graph with unknown functions represented as red edges. (f) Recovered functions.

and inductances and resistance as current functions. Fig. 7.8.(c) displays the available measurements, which are notably sparse, preventing us from reconstructing the underlying unknown functions independently. Thus, their interdependencies must be utilized for approximation. It is crucial to note that the system's state variables are interconnected through functional relations, as per Kirchhoff's laws for this nonlinear electric circuit, illustrated in Fig. 7.8.(d). These functional dependencies can be conceptualized as a computational graph, depicted in Fig. 7.8.(e), where nodes represent variables and directed edges represent functions. Known functions are colored in black, unknown functions in red, and round nodes aggregate variables, meaning edges map groups of variables, forming a hypergraph. The CGC solution involves substituting the graph's unknown functions with Gaussian Processes (GPs), which may be independent or correlated, and then approximating the unknown functions with their Maximum A Posteriori (MAP) estimators, given the available data and the functional dependencies embedded in the graph's structure. Fig. 7.8.(f) showcases the true and recovered functions, demonstrating a notably accurate approximation

despite the data's scarcity.

This simple example generalizes to an abstract framework detailed in (Houman Owhadi, 2022). This framework has a wide range of applications because most problems in CSE can also be formulated as completing computational graphs representing dependencies between functions and variables, and they can be solved in a similar manner by replacing unknown functions with GPs and by computing their MAP/EB estimator given the data. These problems include those illustrated in Fig. 7.1.(d-h).

7.4 Hardness and well-posed formulation of Type 3 problems.

In this subsection, we describe why Type 3 problems are challenging and why they can even be intractable if not formalized and approached properly.

Curse of combinatorial complexity.

First, the problem suffers from the curse of combinatorial complexity in the sense that the number of hypergraphs associated with N nodes blows up rapidly with N. As an illustration, Fig. 7.9 shows some of the hypergraphs associated with only three nodes. A lower bound on that number is the A003180 sequence, which answers the following question (Ishihara, 2001): given N unlabeled vertices, how many different hypergraphs in total can be realized on them by counting the equivalent hypergraphs only once? For $N = 8$, this lower bound is $\approx 2.78 \times 10^{73}$.

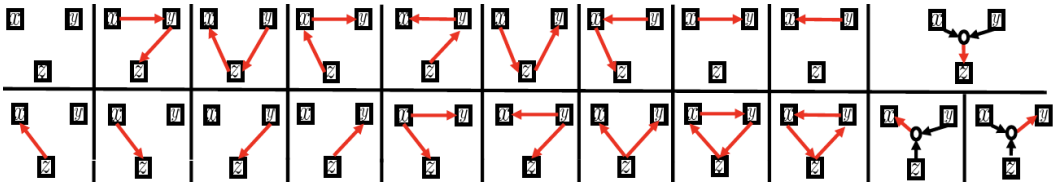


Figure 7.9: Computational Hypergraph Discovery with three variables

Nonidentifiability and implicit dependencies.

Secondly, it is important to note that, even with an infinite amount of data, the exact structure of the hypergraph might not be identifiable. To illustrate this point, let's consider a problem where we have N samples from a computational graph with variables x and y. The task is to determine the direction of functional dependency between x and y. Does it go from x to y (represented as $\boxed{x} \xrightarrow{f} \boxed{y}$), or from y to x (represented as $\boxed{y} \xrightarrow{f} \boxed{x}$)?

If we refer to Fig. 7.10.(a), we can make a decision because y can only be expressed

as a function of x. In contrast, if we examine Fig. 7.10.(b), the decision is also straightforward because x can solely be written as a function of y. However, if the data mirrors the scenario in Fig. 7.10.(c), it becomes challenging to decide as we can write both y as a function of x and x as a function of y. Further complicating matters is the possibility of implicit dependencies between variables. As illustrated in Fig. 7.10.(d), there might be instances where neither y can be derived as a function of x, nor x can be represented as a function of y.

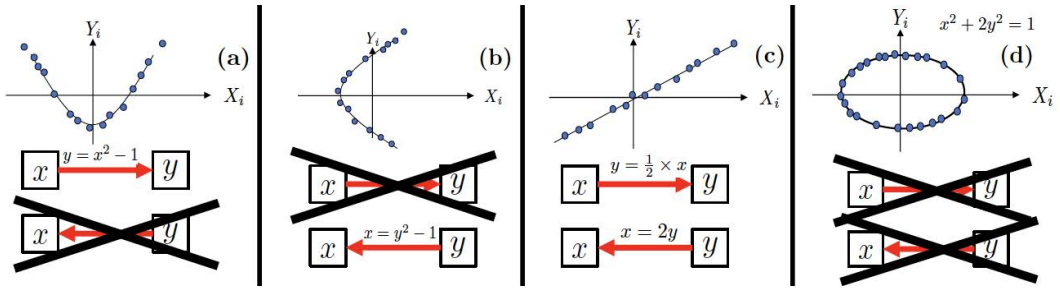


Figure 7.10: The structure of the hypergraph is identifiable in (a), (b), and non-identifiable in (c). The relationship between variables is implicit in (d).

Causal inference and probabilistic graphs.

Causal inference methods broadly consist of two approaches: constraint and score-based methods. While constraint-based approaches are asymptotically consistent, they only learn the graph up to an equivalence class (Spirtes and C. Glymour, 1991). Instead, score-based methods resolve ambiguities in the graph's edges by evaluating the likelihood of the observed data for each graphical model. For instance, they may assign a higher evidence to $y \rightarrow x$ over $x \rightarrow y$ if the conditional distribution $x|y$ exhibits less complexity than $y|x$. The complexity of searching over all possible graphs, however, grows super-exponentially with the number of variables. Thus, it is often necessary to use approximate, but more tractable, search-based methods (Chickering, 2002; J. Peters, Janzing, and Schölkopf, 2017) or alternative criteria based on sensitivity analysis (Data et al., 2016). For example, the preference could lean towards $y \rightarrow x$ rather than $x \rightarrow y$ if y demonstrates less sensitivity to errors or perturbations in x. In contrast, our proposed GP method avoids the growth in complexity by performing a guided pruning process that assesses the contribution of each node to the signal. We also emphasize that our method is not limited to learning acyclic graph structures as it can identify feedback loops between variables. Alternatively, methods for learning probabilistic undirected graphical models, also known as Markov networks, identify the graph structure by assuming the data is

randomly drawn from some probability distribution (Drton and Maathuis, 2017). In this case, edges in the graph (or lack thereof) encode conditional dependencies between the nodes. A common approach learns the graph structure by modeling the data as being drawn from a multivariate Gaussian distribution with a sparse inverse covariance matrix, whose zero entries indicate pairwise conditional independencies (Friedman, Hastie, and Tibshirani, 2008). Recently, this approach has been extended using models for non-Gaussian distributions, e.g., in (Baptista et al., 2021; Ren et al., 2021), as well as kernel-based conditional independence tests (K. Zhang et al., 2011). In this work, we learn functional dependencies rather than causality or probabilistic dependence. We emphasize that we also do not assume the data is randomized or impose strong assumptions, such as additive noise models, in the data-generating process.

We complete this paragraph by comparing the hypergraph discovery framework to structure learning for Bayesian networks and structural equation models (SEM). Let $x \in \mathbb{R}^d$ be a random variable with probability density function p that follows the autoregressive factorization $p(x) = \prod_{i=1}^{d} p_i(x_i|x_1, \ldots, x_{i-1})$ given a prescribed variable ordering. Structure learning for Bayesian networks aims to find the ancestors of variable x_i, often referred to as the set of parents $Pa(i) \subseteq \{1, \ldots, i-1\}$, in the sense that $p_i(x_i|x_1, \ldots, x_{i-1}) = p_i(x_i|x_{Pa(i)})$. Thus, the variable dependence of the conditional density p_i is identified by finding the parent set so that x_i is conditionally independent of all remaining preceding variables given its parents, i.e., $x_i \perp x_{1:i-1\setminus Pa(i)}|x_{Pa(i)}$. Finding ancestors that satisfy this condition requires performing conditional independence tests, which are computationally expensive for general distributions (SHAH and PETERS, 2020). Alternatively, SEMs assume that each variable x_i is drawn as a function of its ancestors with additive noise, i.e. $x_i = f(x_{Pa(i)}) + \epsilon_i$ for some function f and noise ϵ (J. Peters, Janzing, and Schölkopf, 2017). For Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, each marginal conditional distribution in a Bayesian network is given by $p_i(x_i|x_{1:i-1}) \propto \exp(-\frac{1}{2\sigma^2}\|x_i - f(x_{1:i-1})\|^2)$. Thus, finding the parents for such a model by maximum likelihood estimation corresponds to finding the parents that minimize the expected mean-squared error $\|x_i - f(x_{Pa(i)})\|^2$. Our approach minimizes a related objective, without imposing the strong probabilistic assumptions that are required in SEMs and Bayesian Networks. We also observe that while the graph structure identified in Bayesian networks is influenced by the specific sequence in which variables are arranged (a concept exploited in numerical linear algebra (Florian Schäfer, Timothy John Sullivan, and Houman Owhadi, 2021c; Florian Schäfer and Houman Owhadi, 2021) where Schur complementation

is equivalent to conditioning GPs and a carefully ordering leads to the accuracy of the Vecchia approximation $p_i(x_i|x_1,\ldots,x_{i-1}) \approx p_i(x_i|x_{i-k},\ldots,x_{i-1})$ (Vecchia, 1988)), the graph recovered by our approach remains unaffected by any predetermined ordering of those variables.

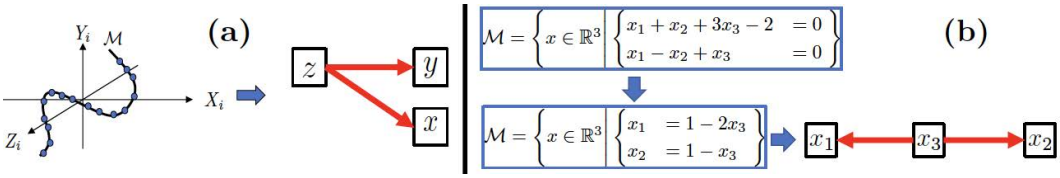


Figure 7.11: (a) CHD formulation as a manifold discovery problem and hypergraph representation, (b) The hypergraph representation of an affine manifold is equivalent to its Row Echelon Form Reduction.

Well-posed formulation of the problem.

In this paper, we focus on a formulation of the problem that remains well-posed even when the data is not randomized, i.e., we formulate the problem as the following manifold learning/discovery problem.

Problem 4. Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) of functions mapping $\mathbb{R}^d$ to $\mathbb{R}$. Let $\mathcal{F}$ be a closed linear subspace of $\mathcal{H}$ and let $\mathcal{M}$ be a subset of $\mathbb{R}^d$ such that $x \in \mathcal{M}$ if and only if $f(x) = 0$ for all $f \in \mathcal{F}$. Given the (possibly noisy and nonrandom) observation of N elements, $X_1,\ldots,X_N$, of $\mathcal{M}$ approximate $\mathcal{M}$.

To understand why problem 4 serves as the appropriate formulation for hypergraph discovery, consider a manifold $\mathcal{M} \subset \mathbb{R}^d$. Suppose this manifold can be represented by a set of equations, expressed as a collection of functions $(f_k)_k$ satisfying $\forall x \in \mathcal{M},\ f_k(x) = 0$. To keep the problem tractable, we assume a certain level of regularity for these functions, necessitating they belong to a RKHS $\mathcal{H}$, ensuring the applicability of kernel methods for our framework. Given that any linear combination of the f_k will also be evaluated to zero on $\mathcal{M}$, the relevant functions are those within the span of the f_k, forming a closed linear subspace of $\mathcal{H}$ denoted as $\mathcal{F}$. The manifold $\mathcal{M}$ can be subsequently represented by a graph or hypergraph (see Fig. 7.11.(a)), whose ambiguity can be resolved through a deliberate decision to classify some variables as free and others as dependent. This selection could be arbitrary, informed by expert knowledge, or derived from probabilistic models or sensitivity analysis.

7.5 A Gaussian Process method for Type 3 problems

Affine case and Row Echelon Form Reduction.

To describe the proposed solution to Problem 4, we start with a simple example. In this example $\mathcal{H}$ is a space of affine functions f of the form

$$f(x) = v^T \psi(x) \text{ with } \psi(x) := \begin{pmatrix} 1 \\ x \end{pmatrix} \text{ and } v \in \mathbb{R}^{d+1}. \tag{7.16}$$

As a particular instantiation (see Fig. 7.11.(b)), we assume $\mathcal{M}$ to be the manifold of $\mathbb{R}^3$ ($d = 3$) defined by the affine equations

$$\mathcal{M} = \left\{ x \in \mathbb{R}^3 \middle| \begin{cases} x_1 + x_2 + 3x_3 - 2 & = 0 \\ x_1 - x_2 + x_3 & = 0 \end{cases} \right\}, \tag{7.17}$$

which is equivalent to selecting $\mathcal{F} = \text{span}\{f_1, f_2\}$ with $f_1(x) = x_1 + x_2 + 3x_3 - 2$ and $f_2(x) = x_1 - x_2 + x_3$ in the problem formulation 4.

Then, irrespective of how we recover the manifold from data, the hypergraph representation of that manifold is equivalent to the row echelon form reduction of the affine system, and this representation and this reduction require a possibly arbitrary choice of free and dependent variables. So, for instance, for the system (7.17), if we declare x_3 to be the free variables and x_1 and x_2 to be the dependent variables, then we can represent the manifold via the equations

$$\mathcal{M} = \left\{ x \in \mathbb{R}^3 \middle| \begin{cases} x_1 & = 1 - 2x_3 \\ x_2 & = 1 - x_3 \end{cases} \right\}, \tag{7.18}$$

which have the hypergraph representation depicted in Fig. 7.11.(b).

Now, in the $N > d$ regime where the number of data points is larger than the number of variables, the manifold can simply be approximated via a variant of PCA. Take $f^* \in \mathcal{F}$, we have $f^*(x) = v^{*T}\psi(x)$ for a certain $v^* \in \mathbb{R}^{d+1}$. Then for $X_s \in \mathcal{M}$, $f^*(X_s) = \psi(X_s)^T v^* = 0$. Defining

$$C_N := \sum_{s=1}^{N} \psi(X_s)\psi(X_s)^T \tag{7.19}$$

we see that $f^*(X_s) = 0$ for all X_s is equivalent to $C_N v^* = 0$. Since $N > d$, we can thus identify $\mathcal{F}$ exactly as $\{v^T\psi \text{ for } v \in Ker(C_N)\}$. We then obtain the manifold

$$\mathcal{M}_N = \left\{ x \in \mathbb{R}^d \mid v^T\psi(x) = 0 \text{ for } v \in \text{Span}(v_{r+1}, \ldots, v_{d+1}) \right\}, \tag{7.20}$$

where $\mathrm{Span}(v_{r+1}, \ldots, v_{d+1})$ is the zero-eigenspace of C_N. Here we write $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_{d+1}$ for the eigenvalues of C_N (in decreasing order), and $v_1, \ldots, v_{d+1}$ for the corresponding eigenvectors ($C_N v_i = \lambda_i v_i$). The proposed approach extends to the noisy case (when the data points are perturbations of elements of the manifold) by simply replacing the zero-eigenspace of the covariance matrix by the linear span of the eigenvectors associated with eigenvalues that are smaller than some threshold $\epsilon > 0$, i.e., by approximating $\mathcal{M}$ with (7.20) where r is such that $\lambda_1 \geq \cdots \geq \lambda_r \geq \epsilon > \lambda_{r+1} \geq \cdots \geq \lambda_{d+1}$. In this affine setting (7.20) allows us to estimate $\mathcal{M}$ directly without RKHS norm minimization/regularization because linear regression does not require regularization in the sufficiently large data regime. Furthermore the process of pruning ancestors can be replaced by that of identifying sparse elements $v \in \mathrm{Span}(v_{r+1}, \ldots, v_{d+1})$ such that $v_i = 1$.

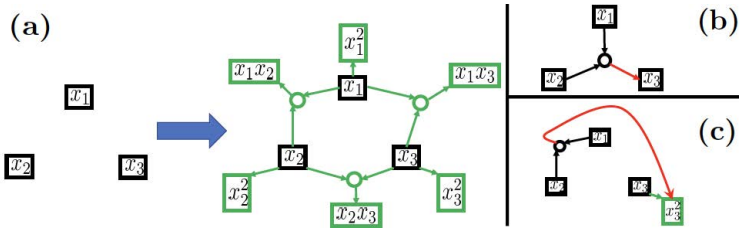


Figure 7.12: Feature map generalization

Feature map generalization.

This simple approach can be generalized by generalizing the underlying feature map ψ used to define the space of functions (writing $d_{\mathcal{S}}$ for the dimension of the range of ψ)

$$\mathcal{H} = \left\{ f(x) = v^T \psi(x) \mid v \in \mathbb{R}^{d_{\mathcal{S}}} \right\}. \tag{7.21}$$

For instance, if we use the feature map

$$\psi(x) := \left(1, \ldots, x_i, \ldots, x_i x_j, \ldots\right)^T \tag{7.22}$$

then $\mathcal{H}$ becomes a space of quadratic polynomials on $\mathbb{R}^d$, i.e.,

$$\mathcal{H} = \left\{ f(x) = v_0 + \sum_i v_i x_i + \sum_{i \leq j} v_{i,j} x_i x_j \mid v \in \mathbb{R}^{d_{\mathcal{S}}} \right\}, \tag{7.23}$$

and, in the large data regime ($N > d_{\mathcal{S}}$), identifying quadratic dependencies between variables becomes equivalent to (1) adding nodes to the hypergraph corresponding to secondary variables obtained from primary variables x_i through known functions (for (7.22), these secondary variables are the quadratic monomials $x_i x_j$, see Fig. 7.12.(a)),

and (2) identifying affine dependencies between the variables of the augmented hypergraph. The problem can, therefore, be reduced to the previous affine case. Indeed, as in the affine case, the manifold can then be approximated in the regime where the number of data points is larger than the dimension $d_{\mathcal{S}}$ of the feature map by (7.20), where $v_r, \ldots, v_N$ are the eigenvectors of C_N =(7.19) whose eigenvalues are zero (noiseless case) or smaller than some threshold $\epsilon > 0$ (noisy case).

Furthermore, the hypergraph representation of the manifold is equivalent to a feature map generalization of Row Echelon Form Reduction to nonlinear systems of equations. For instance, choosing x_3 as the dependent variable and x_1, x_2 as the free variables, $\mathcal{M} = \{x \in \mathbb{R}^3 \mid x_3 - 5x_1^2 + x_2^2 - x_1x_2 = 0\}$ can be represented as in Fig. 7.12.(b) where the round node represents the concatenated variable (x_1, x_2) and the red arrow represents a quadratic function. The generalization also enables the representation of implicit equations by selecting secondary variables as free variables. For instance, selecting x_3^2 as the free variable and x_1, x_2 as the free variables, $\mathcal{M} = \{x \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 - 1 = 0\}$ can be represented as in Fig. 7.12.(c).

Kernel generalization and regularization.

This feature-map extension of the previously discussed affine case can evidently be generalized to arbitrary degree polynomials and to other basis functions. However, as the dimension $d_{\mathcal{S}}$ of the range of the feature map ψ increases beyond the number N of data points, the problem becomes underdetermined: the data only provides partial information about the manifold, i.e., it is not sufficient to uniquely determine the manifold. Furthermore, if the dimension of the feature map is infinite, then we are always in that low data regime, and we have the additional difficulty that we cannot directly compute with that feature map. On the other hand, if $d_{\mathcal{S}}$ is finite (i.e., if the dictionary of basis functions is finite), then some elements of $\mathcal{F}$ (some constraints defining the manifold $\mathcal{M}$) may not be representable or well approximated as equations of the form $v^T\psi(x) = 0$. To address these conflicting requirements, we need to kernelize and regularize the proposed approach (as done in interpolation).

The kernel associated with the feature map.

To describe this kernelization, we assume that the feature map ψ maps $\mathbb{R}^d$ to some Hilbert space $\mathcal{S}$ that could be infinite-dimensional, and we write K for the kernel defined by that feature map. To be precise, we now consider the setting where the feature map ψ is a function from $\mathbb{R}^d$ to a (possibly infinite-dimensional separable)

Hilbert (feature) space $\mathcal{S}$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{S}}$. To simplify notations, we will still write $v^T w$ for $\langle v, w \rangle_{\mathcal{S}}$ and vw^T for the linear operator mapping v' to $v\langle w, v' \rangle_{\mathcal{S}}$. Let

$$\mathcal{H} := \{v^T \psi(x) \mid v \in \mathcal{S}\} \tag{7.24}$$

be the space of functions mapping $\mathbb{R}^d$ to $\mathbb{R}$ defined by the feature map ψ. To avoid ambiguity, assume (without loss of generality) that the identity $v^T \psi(x) = w^T \psi(x)$ holds for all $x \in \mathbb{R}^d$ if and only if $v = w$. It follows that for $f \in \mathcal{H}$ there exists a unique $v \in \mathcal{S}$ such that $f = v^T \psi$. For $f, g \in \mathcal{H}$ with $f = v^T \psi$ and $g = w^T \psi$, we can then define

$$\langle f, g \rangle_{\mathcal{H}} := v^T w \,. \tag{7.25}$$

Observe that $\mathcal{H}$ is a Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. For $x, x' \in \mathcal{X}$, write

$$K(x, x') := \psi(x)^T \psi(x') \,, \tag{7.26}$$

for the kernel defined by ψ and observe that $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is the RKHS defined by the kernel K (which is assumed to contain $\mathcal{F}$ in Problem 4). Observe in particular that for $f = v^T \psi \in \mathcal{H}$, K satisfies the reproducing property

$$\langle f, K(x, \cdot) \rangle_{\mathcal{H}} = v^T \psi(x) = f(x) \,. \tag{7.27}$$

Complexity Reduction with Kernel PCA Variant.

We will now show that the previous feature-map PCA variant (characterizing the subspace of $f \in \mathcal{H}$ such that $f(X) = 0$) can be kernelized as a variant of kernel PCA (Mika et al., 1998). To describe this write $K(X, X)$ for the $N \times N$ matrix with entries $K(X_i, X_j)$. Write $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ for the nonzero eigenvalues of $K(X, X)$ indexed in decreasing order and write $\alpha_{\cdot,i}$ for the corresponding unit-normalized eigenvectors, i.e.

$$K(X, X)\alpha_{\cdot,i} = \lambda_i \alpha_{\cdot,i} \text{ and } |\alpha_{\cdot,i}| = 1 \,. \tag{7.28}$$

Write $f(X)$ for the N vector with entries $f(X_s)$. For $i \leq r$, write

$$\phi_i := \sum_{s=1}^{N} \delta_{X_s} \alpha_{s,i} \tag{7.29}$$

and

$$f(\phi_i) := \sum_{s=1}^{N} f(X_s) \alpha_{s,i} \,. \tag{7.30}$$

Write $f(\phi)$ for the r vector with entries $f(\phi_i)$.

Then, we have the following proposition.

Proposition 7.5.1. The subspace of functions $f \in \mathcal{H}$ such that $f(\phi) = 0$ is equal to the subspace of $f \in \mathcal{H}$ such that $f(X) = 0$. Furthermore for $f \in \mathcal{H}$ with feature map representation $f = v^T\psi$ with $v \in \mathcal{S}$ we have the identity (where C_N =(7.19))

$$v^T C_N v = |f(\phi)|^2 = |f(X)|^2 . \tag{7.31}$$

Proof. Write $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_{\widehat{r}} > 0$ for the nonzero eigenvalues of C_N =(7.19) indexed in decreasing order. Write $v_1, \ldots, v_r$ for the corresponding eigenvectors, i.e.,

$$C_N v_i = \widehat{\lambda}_i v_i . \tag{7.32}$$

Observing that

$$C_N = \sum_{i=1}^{r} \widehat{\lambda}_i v_i v_i^T \tag{7.33}$$

we deduce that the zero-eigenspace of C_N is the set of vectors $v \in \mathcal{S}$ such that $v^T v_i = 0$ for $i = 1, \ldots, r$. Write $f_i := v_i^T \psi$. Observe that for $f = v^T\psi$, we have $v_i^T v = \left\langle f_i, f \right\rangle_K$. Multiplying (7.32) by $\psi^T(x)$ implies

$$\sum_{s=1}^{N} K(x, X_s) f_i(X_s) = \widehat{\lambda}_i f_i(x) \tag{7.34}$$

(7.34) implies that for $f = v^T\psi$

$$v_i^T v = \sum_{s=1}^{N} \widehat{\lambda}_i^{-1} f_i(X_s) \left\langle K(\cdot, X_s), f \right\rangle_K = \sum_{s=1}^{N} \widehat{\lambda}_i^{-1} f_i(X_s) f(X_s), \tag{7.35}$$

where we have used the reproducing property (7.27) of K in the last identity. Write

$$\widehat{\alpha}_{s,i} := \lambda_i^{-1/2} f_i(X_s) . \tag{7.36}$$

Using (7.34) with $x = X_{s'}$ implies that $\widehat{\alpha}_{\cdot,i}$ is an eigenvector of the $N \times N$ matrix $K(X, X)$ with eigenvalue $\widehat{\lambda}_i$. Taking $f = f_i$ in (7.35) implies that $1 = v_i^T v_i = |\widehat{\alpha}_{\cdot,i}|^2$. Therefore, the $\widehat{\alpha}_{\cdot,i}$ are unit-normalized. Summarizing, this analysis (closely related to the one found in kernel PCA (Mika et al., 1998)) shows that the nonzero eigenvalues of $K(X, X)$ coincide with those of C_N and we have $\widehat{r} = r$, $\widehat{\lambda}_i = \lambda_i$ and $\widehat{\alpha}_{\cdot,i} = \alpha_{\cdot,i}$. Furthermore, (7.35) and (7.36) imply that for $i \leq r$, $v \in \mathcal{S}$ and $f = v^T\psi$, we have

$$v_i^T v = \lambda_i^{-1/2} f(X) \alpha_{\cdot,i} . \tag{7.37}$$

The identity (7.37) then implies (7.31). □

Remark 7.5.2. As in PCA the dimension/complexity of the problem can be further reduced by truncating ϕ to $\phi' = (\phi_1, \ldots, \phi_{r'})$ where $r' \leq r$ is identified as the smallest index i such that $\lambda_i/\lambda_1 < \epsilon$ where $\epsilon > 0$ is some small threshold.

Kernel Mode Decomposition.

When the feature map ψ is infinite-dimensional, the data only provides partial information about the constraints defining the manifold in the sense that $f(X) = 0$ or equivalently $f(\phi) = 0$ is a necessary but not sufficient condition for the zero level set of f to be a valid constraint for the manifold (for f to be such that $f(x) = 0$ for all $x \in \mathcal{M}$). So we are faced with the following problems: (1) How to regularize? (2) How do we identify free and dependent variables? (3) How do we identify valid constraints for the manifold? The proposed solution will be based on the Kernel Mode Decomposition (KMD) framework introduced in (Houman Owhadi, Clint Scovel, and Yoo, 2021) (which shares conceptual foundations with Smoothing Spline ANOVA (Wahba, 2003)).

Reminder on KMD We will now present a quick reminder on KMD in the setting of the following mode decomposition problem. So, in this problem, we have an unknown function $f^\dagger$ mapping some input space $\mathcal{X}$ to the real line $\mathbb{R}$. We assume that this function can be written as a sum of m other unknown functions $f_i^\dagger$ which we will call modes, i.e.,

$$f^\dagger = \sum_{i=1}^{m} f_i^\dagger \,. \tag{7.38}$$

We assume each mode $f_i^\dagger$ to be an unknown element of some RKHS $\mathcal{H}_{K_i}$ defined by some kernel K_i. Then we consider the problem in which given the data $f^\dagger(X) = Y$ (with $(X, Y) \in \mathcal{X}^N \times \mathbb{R}^N$) we seek to approximate the m modes composing the target function $f^\dagger$. Then, we have the following theorem.

Theorem 7.5.3. (Houman Owhadi, Clint Scovel, and Yoo, 2021) Using the relative error in the product norm $\|(f_1, \ldots, f_m)\|^2 := \sum_{i=1}^m \|f_i\|_{K_i}^2$ as a loss, the minimax optimal recovery of $(f_1^\dagger, \ldots, f_m^\dagger)$ is $(f_1, \ldots, f_m)$ with

$$f_i(x) = K_i(x, X)K(X, X)^{-1}Y \,, \tag{7.39}$$

where K is the additive kernel

$$K = \sum_{i=1}^{m} K_i \,. \tag{7.40}$$

The GP interpretation of this optimal recovery result is as follows. Let $\xi_i \sim \mathcal{N}(0, K_i)$ be m independent centered GPs with kernels K_i. Write ξ for the additive GP $\xi := \sum_{i=1}^m \xi_i$. (7.39) can be recovered by replacing the modes $f_i^\dagger$ by independent centered

GPs $\xi_i \sim \mathcal{N}(0, K_i)$ with kernels K_i and approximating the mode i by conditioning ξ_i on the available data $\xi(X) = Y$ where $\xi := \sum_{i=1}^m \xi_i$ is the additive GP obtained by summing the independent GPs ξ_i, i.e.,

$$f_i(x) = \mathbb{E}\big[\xi_i(x) \mid \xi(X) = Y\big] . \tag{7.41}$$

Furthermore $(f_1, \ldots, f_m)$ can also be identified as the minimizer of

$$\begin{cases} \text{Minimize} & \sum_{i=1}^m \|f_i\|_{K_i}^2 \\ \text{over} & (f_1, \ldots, f_m) \in \mathcal{H}_{K_1} \times \cdots \times \mathcal{H}_{K_m} \\ \text{s. t.} & (\sum_{i=1}^m f_i)(X) = Y . \end{cases} \tag{7.42}$$

The variational formulation (7.42) can be interpreted as a generalization of Tikhonov regularization which can be recovered by selecting $m = 2$, K_1 to be a smoothing kernel (such as a Matérn kernel) and $K_2(x, y) = \sigma^2 \delta(x - y)$ to be a white noise kernel.

Now, this abstract KMD approach (Houman Owhadi, Clint Scovel, and Yoo, 2021) is associated with a quantification of how much each mode contributes to the overall data or how much each individual GP ξ_i explains the data. More precisely, the activation of the mode i or GP ξ_i can be quantified as

$$p(i) = \frac{\|f_i\|_{K_i}^2}{\|f\|_K^2} , \tag{7.43}$$

where $f = \sum_{i=1}^m f_i$. These activations $p(i)$ satisfy $p(i) \in [0, 1]$ and $\sum_{i=1}^m p(i) = 1$ they can be thought of as a generalization of Sobol sensitivity indices (Sobol, 2001; Sobol', 1993; Owen, 2013) to the nonlinear setting in the sense that they are associated with the following variance representation/decomposition (Houman Owhadi, Clint Scovel, and Yoo, 2021) (writing $\langle \cdot, \cdot \rangle_K$ for the RKHS inner product induced by K):

$$\mathrm{Var}\left[\langle \xi, f \rangle_K\right] = \|f\|_K^2 = \sum_{i=1}^m \|f_i\|_{K_i}^2 = \sum_{i=1}^m \mathrm{Var}\left[\langle \xi_i, f \rangle_K\right]. \tag{7.44}$$

Application to CHD, general case. Now, let us return to our original manifold approximation problem 4 in the kernelized setting of (7.26). Given the data X we cannot regress an element $f \in \mathcal{F}$ directly since the minimizer of $\|f\|_K^2 + \gamma^{-1}\|f(X)\|_{\mathbb{R}^N}^2$ is the null function. To identify the functions $f \in \mathcal{F}$, we need to decompose them into modes that can be interpreted as a generalization of the notion

of free and dependent variables. To describe this, assume that the kernel K can be decomposed as the additive kernel

$$K = K_a + K_s + K_z. \tag{7.45}$$

Then $\mathcal{H}_K = \mathcal{H}_{K_a} + \mathcal{H}_{K_s} + \mathcal{H}_{K_z}$ implies that for all function $f \in \mathcal{H}_K$, f can be decomposed as $f = f_a + f_s + f_z$ with $(f_a, f_s, f_z) \in \mathcal{H}_a \times \mathcal{H}_s \times \mathcal{H}_z$.

Example 7.5.4. As a running example, take K to be the following additive kernel

$$K(x, x') = 1 + \beta_1 \sum_i x_i x_i' + \beta_2 \sum_{i \le j} x_i x_j x_i' x_j' + \beta_3 \prod_i (1 + k(x_i, x_i')), \tag{7.46}$$

that is the sum of a linear kernel, a quadratic kernel, and a fully nonlinear kernel. Take K_a to be the part of the linear kernel that depends only on x_1, i.e.,

$$K_a(x, x') = \beta_1 x_1 x_1' . \tag{7.47}$$

Take K_s to be the part of the kernel that does not depend on x_1, i.e.,

$$K_s = 1 + \beta_1 \sum_{i \neq 1} x_i x_i' + \beta_2 \sum_{i \le j, i, j \neq 1} x_i x_j x_i' x_j' + \beta_3 \prod_{i \neq 1} (1 + k(x_i, x_i')) . \tag{7.48}$$

And take K_z to be the remaining portion,

$$K_z = K - K_a - K_s . \tag{7.49}$$

Therefore the following questions are equivalent:

- Given a function g_a in the RKHS $\mathcal{H}_{K_a}$ defined by the kernel K_a is there a function f_s in the RKHS $\mathcal{H}_{K_s}$ defined by the kernel K_s such that $g_a(x) \approx f_s(x)$ for $x \in \mathcal{M}$?
- Given a function $g_a \in \mathcal{H}_{K_a}$ is there a function f in the RKHS $\mathcal{H}_K$ defined by the kernel K such that $f(x) \approx 0$ for $x \in \mathcal{M}$ and such that its f_a mode is $-g_a$ and its f_z mode is zero?

Then, the natural answer to the questions is to identify the modes of the constraint $f = f_a + f_s + f_z \in \mathcal{H}$ (such that $f(x) \approx 0$ for $x \in \mathcal{M}$) such that $f_a = -g_a$ and $f_z = 0$ by selecting f_s to be the minimizer of the following variational problem

$$\min_{f_s \in \mathcal{H}_s} \|f_s\|_{K_s}^2 + \frac{1}{\gamma} \left| (-g_a + f_s)(\phi) \right|^2 . \tag{7.50}$$

This is equivalent to introducing the additive GP $\xi = \xi_a + \xi_s + \xi_z + \xi_n$ whose modes are the independent GPs $\xi_a \sim \mathcal{N}(0, K_a)$, $\xi_s \sim \mathcal{N}(0, K_s)$, $\xi_z \sim \mathcal{N}(0, K_z)$, $\xi_n \sim \mathcal{N}(0, \gamma\delta(x-y))$ (we use the label "n" in reference to "noise"), and then recovering f_s as

$$f_s = \mathbb{E}\big[\xi_s \mid \xi(X) = 0, \xi_a = -g_a, \xi_z = 0\big] . \tag{7.51}$$

Application to CHD, particular case. Taking $g_a(x) = x_1$ for our running example 7.5.4, the previous questions are, as illustrated in Fig. 7.2(b), equivalent to asking whether there exists a function $f_s \in \mathcal{H}_{K_s}$ that does not depend on x_1 (since K_s does not depend on x_1) such that

$$x_1 \approx f_s(x_2, \ldots, x_d) \text{ for } x \in \mathcal{M} . \tag{7.52}$$

Therefore, the mode f_a can be thought of as a dependent mode (we use the label "a" in reference to "ancestors"), the mode f_s as a free mode (we use the label "s" in reference to "signal"), the mode f_z as a zero mode.

While our numerical illustrations have primarily focused on the scenario where g_a takes the form of $g_a(x) = x_i$, and we aim to express x_i as a function of other variables, the generality of our framework is motivated by its potential to recover implicit equations. For example, consider the implicit equation $x_1^2 + x_2^2 = 1$, which can be retrieved by setting the mode of interest to be $g_a(x) = x_1^2$ and allowing f_s to depend only on the variable x_2.

Signal-to-noise ratio.

Now, we are led to the following question: since the mode f_s (the minimizer of (7.50)) always exists and is always unique, how do we know that it leads to a valid constraint? To answer that question, we compute the activation of the GPs used to regress the data. We write

$$\mathcal{V}(s) := \|f_s\|_{K_s}^2 , \tag{7.53}$$

for the activation of the signal GP ξ_s and

$$\mathcal{V}(n) := \frac{1}{\gamma}\big|(-g_a + f_s)(X)\big|^2 \tag{7.54}$$

for the activation of the noise GP ξ_n, and then these allow us to define a signal-to-noise ratio defined as

$$\frac{\mathcal{V}(s)}{\mathcal{V}(s) + \mathcal{V}(n)} . \tag{7.55}$$

Note that this corresponds to activation ratio of the noise GP defined in (7.43). This ratio can then be used to test the validity of the constraint in the sense that if $V(s)/(V(s)+V(n)) > \tau$ (with $\tau = 0.5$ as a prototypical example), then the data is mostly explained by the signal GP and the constraint is valid. If $V(s)/(V(s)+V(n)) < \tau$, then the data is mostly explained by the noise GP and the constraint is not valid.

Iterating by removing the least active modes from the signal.

If the constraint is valid, then we can next compute the activation of the modes composing the signal. To describe this, we assume that the kernel K_s can be decomposed as the additive kernel

$$K_s = K_{s,1} + \cdots + K_{s,m}\,, \tag{7.56}$$

which results in $\mathcal{H}_{K_s} = \mathcal{H}_{K_{s,1}} + \cdots + \mathcal{H}_{K_{s,m}}$, which results in the fact that $\forall f_s \in \mathcal{H}_s$, f_s can be decomposed as

$$f_s = f_{s,1} + \cdots + f_{s,m}\,, \tag{7.57}$$

with $f_{s,i} \in \mathcal{H}_{K_{s,i}}$. The activation of the mode i can then be quantified as $p(i) = \|f_{s,i}\|^2_{K_{s,i}}/\|f_s\|^2_{K_s}$, which combined with $\|f_s\|^2_{K_s} = \sum_{i=1}^m \|f_{s,i}\|^2_{K_{s,i}}$ leads to $\sum_{i=1}^m p(i) = 1$.

As our running example 7.5.4, we can decompose K_s =(7.48) as the sum of an affine kernel, a quadratic kernel, and a fully nonlinear kernel, i.e., $m = 3$, $K_{s,1} = 1 + \beta_1 \sum_{i\neq 1} x_i x_i'$, $K_{s,2} = \beta_2 \sum_{i\leq j, i,j\neq 1} x_i x_j x_i' x_j'$ and $K_{s,3} = \beta_3 \prod_{i\neq 1}(1 + k(x_i, x_i'))$.

As another example for our running example, we can take K_s to be the sum of the portion of the kernel that does not depend on x_1 and x_2 and the remaining portion, i.e., $m = 2$, $K_{s,1} = 1+\beta_1 \sum_{i\neq 1,2} x_i x_i' + \beta_2 \sum_{i\leq j, i,j\neq 1,2} x_i x_j x_i' x_j' + \beta_3 \prod_{i\neq 1,2}(1+k(x_i, x_i'))$ and $K_{s,2} = K_s - K_{s,1}$.

Then, we can order these sub-modes from most active to least active and create a new kernel K_s by removing the least active modes from the signal and adding them to the mode that is set to be zero (see Fig. 7.13). To describe this, let $\pi(1), \cdots, \pi(m)$ be an ordering of the modes by their activation, i.e., $\|f_{s,\pi(1)}\|^2_{K_{s,\pi(1)}} \geq \|f_{s,\pi(2)}\|^2_{K_{s,\pi(2)}} \geq \cdots$.

Writing $K_t = \sum_{i=r+1}^m K_{s,\pi(i)}$ for the additive kernel obtained from the least active modes (with $r + 1 = m$ as the value used for our numerical implementations), we update the kernels K_s and K_z by assigning the least active modes from K_s to K_z, i.e., $K_s - K_t \to K_s$ and $K_z + K_t \to K_z$ (we zero the least active modes).

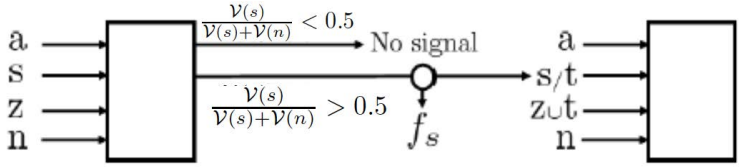


Figure 7.13: Iterating by removing the least active modes from the signal

Finally, we can iterate the process. This iteration can be thought of as identifying the structure of the hypergraph by placing too many hyperedges and removing them according to the activation of the underlying GPs.

For our running example 7.5.4, where we try to identify the ancestors of the variable x_1, if the sub-mode associated with the variable x_2 is found to be least active, then we can try to remove x_2 from the list of ancestors and try to identify x_1 as a function of x_3 to x_d. This is equivalent to selecting $K_a(x, x') = \beta_1 x_1 x_1'$,

$$K_{s/t} = 1 + \beta_1 \sum_{i \neq 1,2} x_i x_i' + \beta_2 \sum_{i \leq j, i,j \neq 1,2} x_i x_j x_i' x_j' + \beta_3 \prod_{i \neq 1,2} (1 + k(x_i, x_i')) \,, \tag{7.58}$$

and $K_{z \cup t} = K - K_a - K_{s/t}$ to assess whether there exists a function $f_s \in \mathcal{H}_K$ that does not depend on x_1 and x_2 s.t. $x_1 \approx f_s(x_3, \ldots, x_d)$ for $x \in \mathcal{M}$.

Alternative determination of the list of ancestors.

Our initial approach to determining the list of ancestors of a given node is to use a fixed threshold (e.g., $\tau = 0.5$) to prune nodes. We propose a refined approach that mimics the strategy employed in Principal Component Analysis (PCA) for deciding which modes should be kept and which ones should be removed. The PCA approach is to order the modes in decreasing order of eigenvalues/variance and (1) either keep the smallest number modes holding/explaining a given fraction (e.g., 90%) of the variance in the data, (2) or use an inflection point/sharp drop in the decay of the eigenvalues to select which modes should be kept. Here, we propose a similar strategy. First we employ an alternative determination of the least active mode: we iteratively remove the mode that leads to the smallest increase in noise-to-signal ratio, i.e., we remove the mode t such that

$$t = \text{argmin}_{\text{t}} \frac{\mathcal{V}(n)}{\mathcal{V}(s/t) + \mathcal{V}(n)} \,. \tag{7.59}$$

For our running example 7.5.4 in which we try to find the ancestors of the variable x_1 this is equivalent to removing the variables or node t whose removal leads to

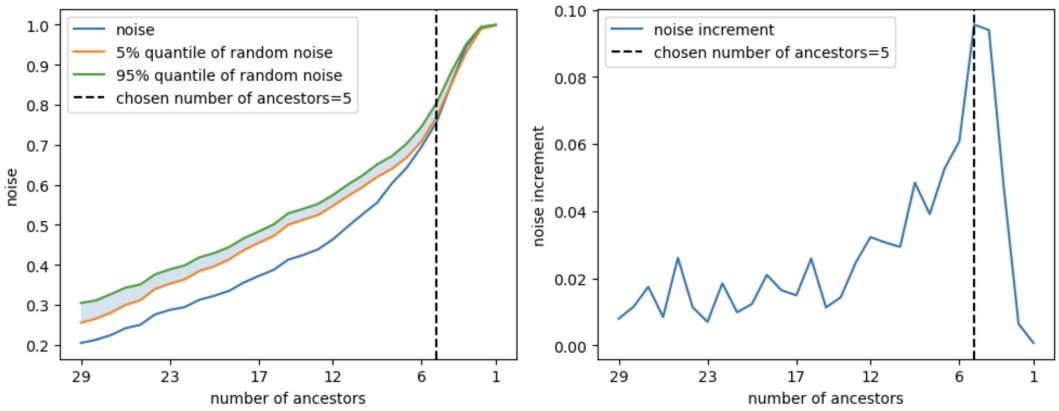


Figure 7.14: Computing the ancestors of the variable $\dot{x}_0$ in the Fermi-Pasta-Ulam-Tsingou problem. (a) Noise-to-Signal Ratio, denoted as $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$, with respect to the number of proposed ancestors, represented by q. Additionally, we include a visualization of the quantiles derived from the Z-test, as described in Section 7.7. Notably, when there is no signal present, the noise-to-signal ratio is expected to fall within the shaded area with a probability of 0.9. (b) Increments in the Noise-to-Signal Ratio, defined as $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q) - \frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q-1)$, as a function of the number of ancestors, denoted as q. The horizontal axis represents the number of proposed ancestors for $\dot{x}_0$. Determining an appropriate stopping point based solely on absolute noise-to-signal ratio levels can be challenging. In contrast, the increments in the noise-to-signal ratio clearly exhibit a discernible maximum, offering a practical point for decision-making.

the smallest loss in signal-to-noise ratio (or increase in noise-to-signal ratio) by selecting

$$K_{s/t} = 1 + \beta_1 \sum_{i \neq 1,t} x_i x_i' + \beta_2 \sum_{i \leq j, i,j \neq 1,t} x_i x_j x_i' x_j' + \beta_3 \prod_{i \neq 1,t} (1 + k(x_i, x_i')) \,.$$

Next, we iterate this process, and we plot (a) the noise-to-signal ratio, and (b) the increase in noise-to-signal ratio as a function of the number of ancestors ordered according to this iteration. Fig. 7.14 illustrates this process and shows that the removal of an essential node leads to a sharp spike in increase in the noise-to-signal ratio (the noise-to-signal ratio jumps from approximately 50-60% to 99%). The identification of this inflection point can be used as a method for effectively and reliably pruning ancestors.

7.6 Algorithm pseudocode.

Our overall method is summarized in the pseudocode Alg. 11 and Alg. 12 that we will now describe. Alg. 11 takes the data D (encoded into the samples $X_1, \ldots, X_N$ of Problem 4) and the set of nodes V as an input and produces, as described in

Algorithm 11 CHD by thresholding the signal-to-noise ratio

Input: Data D, set of nodes V, threshold τ (default $\tau = 0.5$)
Output: Learned hypergraph *Set of ancestors for each node*
1: $D \leftarrow$ NormalizeData(D) *Normalize the data*
2: **for** each $v \in V$ **do**
3: **for** each kernel in `{"linear", "quadratic", "nonlinear"}` *Find the kernel* **do**
4: SetOfAncestors(v) $\leftarrow$ all other nodes
5: SignalToNoiseRatio $\leftarrow$ ComputeSignalToNoiseRatio(kernel, node, D)
6: **if** SignalToNoiseRatio $> \tau$ **then**
7: Choose that kernel and **exit** the for loop
8: **else**
9: Remove all ancestors from node
10: **end if**
11: **end for**
12: **while** SignalToNoiseRatio $> \tau$ *Prune ancestors* **do**
13: Find least important ancestor
14: Recompute SignalToNoiseRatio without that ancestor
15: **if** SignalToNoiseRatio $> \tau$ **then**
16: Remove that ancestor
17: **end if**
18: **end while**
19: **end for**

Sec. 7.5, for each node $i \in V$ its set of minimal ancestors A_i and the simplest possible function f_i such that $x_i \approx f_i\big((x_j)_{j\in A_i}\big)$. It employs the default threshold of 0.5 on the signal-to-noise ratios for its operations. Line 1 normalizes the data (via an affine transformation) so that the samples X_i are of mean zero and variance 1. Given a node with index $i = 1$ in Line 2 (i runs through the set of nodes, and we select $i = 1$ for ease of presentation), the command in Line 3 refers to selecting a signal kernel of the form K_s =(7.48) (where k is selected to be a vanilla RBF kernel such as Gaussian or Matérn), with $1 \geq \beta_1 > 0 = \beta_2 = \beta_3$ for the linear kernel, $1 \geq \beta_1 \geq \beta_2 > 0 = \beta_3$ for the quadratic kernel and $1 \geq \beta_1 \geq \beta_2 \geq \beta_3 > 0$ for the fully nonlinear (interpolative) kernel. The ComputeSignalToNoiseRatio function in Line 5 computes the signal-to-noise ratio with $g_a(x) = x_1$ and with the kernel selected in Line 3. The value of γ is selected automatically by maximizing the variance of the histogram of eigenvalues of D_γ as described in Sec. 7.7 (with the kernel $K = K_s$ =(7.48) selected in Line 3 and $Y = g_a(X)$ with $g_a(x) = x_1$). The value of γ is re-computed whenever a node is removed from the list of ancestors, and K_s is nonlinear. Lines 13, 14 and 16 are described in Sec. 7.5. They correspond

to iteratively identifying the ancestor node t contributing the least to the signal and removing that node from the set of ancestors of the node 1 if the removal of that node t does not send the signal-to-noise ratio below the default threshold 0.5.

Algorithm 12 CHD by inflection point in the noise-to-signal ratio

Input: Data D, set of nodes V, threshold τ (default $\tau = 0.5$)
Output: Learned hypergraph *Set of ancestors for each node*

$D \leftarrow$ NormalizeData(D) *Normalize the data*
for each node $v \in V$ **do**
  for each kernel in `{"linear", "quadratic", "nonlinear"}` *Find the kernel* **do**
    SetOfAncestors $\leftarrow$ all other nodes
    SignalToNoiseRatio $\leftarrow$ ComputeSignalToNoiseRatio(kernel, node, D)
    if SignalToNoiseRatio $> \tau$ **then**
      Choose that kernel and **exit** the for loop
    else
      Remove all ancestors from node
    end if
  end for
  $q \leftarrow$ Cardinal(all other nodes)
  SetOfAncestors(q) $\leftarrow$ all other nodes
  while $q \geq 1$ **do**
    NoiseToSignalRatio(q) $\leftarrow$ ComputeNoiseToSignalRatio(kernel, node, D)
    LeastImportantAncestor $\leftarrow$ Find least important ancestor in SetOfAncestors(q)
    SetOfAncestors($q-1$) $\leftarrow$ SetOfAncestors(q)$\setminus$ LeastImportantAncestor
    $q \leftarrow q - 1$
  end while
  $q^{\dagger} \leftarrow$ inflection point in $q \rightarrow$ NoiseToSignalRatio(q)
      or spike in $q \rightarrow$ NoiseToSignalRatio(q)$-$NoiseToSignalRatio($q-1$)
  FinalSetOfAncestors(v) $\leftarrow$ SetOfAncestors($q^{\dagger}$)
end for

Algorithm 12 distinguishes itself from Algorithm 11 in its approach to pruning ancestors based on signal-to-noise ratios. Instead of using a default threshold of 0.5 like Algorithm 11, Algorithm 12 computes the noise-to-signal ratio, represented as $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$. This ratio is calculated as a function of the number q of ancestors, which are ordered based on their decreasing contribution to the signal. The detailed methodology behind this computation can be found in Section 7.5 and is visually depicted in Figure 7.14. The final number q of ancestors is then determined by finding the value that maximizes the difference between successive noise-to-signal

ratios, $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q+1) - \frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}(q)$.

7.7 Analysis of the signal-to-noise ratio test.

The signal-to-noise ratio depends on the prior on the level of noise.

The signal-to-noise ratio (7.55) depends on the value of γ, which is the variance prior on the level of noise. The goal of this subsection is to answer the following two questions: (1) How do we select γ? (2) How do we obtain a confidence level for the presence of a signal? Or equivalently for a hyperedge of the hypergraph? To answer these questions, we will now analyze the signal-to-noise ratio in the following regression problem in which we seek to approximate the unknown function $f^\dagger : \mathcal{X} \to \mathbb{R}$ based on noisy observations

$$f^\dagger(X) + \sigma Z = Y \tag{7.60}$$

of its values at collocation points X_i ($(X, Y) \in \mathcal{X}^N \times \mathbb{R}^N$, $Z \in \mathbb{R}^N$, and the entries Z_i of Z are i.i.d $\mathbb{N}(0, 1)$). Assuming σ^2 to be unknown and writing γ for a candidate for its value, recall that the GP solution to this problem is approximate $f^\dagger$ by interpolating the data with the sum of two independent GPs, i.e.,

$$f(x) = \mathbb{E}[\xi(x)|\xi(X) + \sqrt{\gamma}Z = Y] \,, \tag{7.61}$$

where $\xi \sim \mathcal{N}(0, K)$ is the GP prior for the signal $f^\dagger$ and $\sqrt{\gamma}Z \sim \mathcal{N}(0, \gamma I_N)$ is the GP prior for the noise σZ in the measurements. Following Sec. 7.5 f can also be identified as a minimizer of

$$\text{minimize}_{f'} \|f'\|_K^2 + \frac{1}{\gamma}\|f'(X) - Y\|_{\mathbb{R}^N}^2 \,, \tag{7.62}$$

the activation of the signal GP can be quantified as $s = \|f\|_K^2$, the activation of the noise GP can be quantified as $\mathcal{V}(n) = \frac{1}{\gamma}\|f(X) - Y\|_{\mathbb{R}^N}^2$. We can then define the noise to signal ratio $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)}$, which admits the following representer formula:

$$\frac{\mathcal{V}(n)}{\mathcal{V}(s) + \mathcal{V}(n)} = \gamma \frac{Y^T\big(K(X, X) + \gamma I\big)^{-2} Y}{Y^T\big(K(X, X) + \gamma I\big)^{-1} Y} \,. \tag{7.63}$$

Observe that when applied to the setting of Sec. 7.5, this signal-to-noise ratio is calculated with $K = K_s$ and $Y = g_a(X)$.

Now we have the following proposition, which follows from (7.63).

Proposition 7.7.1. It holds true that $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)} \in [0, 1]$, and if $K(X, X)$ has full rank,

$$\lim_{\gamma \downarrow 0} \frac{\mathcal{V}(n)}{\mathcal{V}(s) + \mathcal{V}(n)} = 0 \text{ and } \lim_{\gamma \uparrow \infty} \frac{\mathcal{V}(n)}{\mathcal{V}(s) + \mathcal{V}(n)} = 1 \,. \tag{7.64}$$

Therefore, we are led to the following question: if the signal $f^\dagger$ and the level of noise σ^2 are both unknown, how do we select γ to decide whether the data is mostly signal or noise?

How do we select the prior on the level of noise?

Our answer to this question depends on whether the feature-map associated with the base kernel K is finite-dimensional or not.

When the kernel is linear, quadratic, or associated with a finite-dimensional feature map.

If the feature-map associated with the base kernel K is finite-dimensional, then γ can be estimated from the data itself when the number of data-points is sufficiently large (at least larger than the dimension of the feature-space $\mathcal{S}$). A prototypical example (when trying to identify the ancestors of the variable x_1) is $K = K_s$=(7.48) with $\beta_3 = 0$. In the general setting assume that $K(x, x') := \psi(x)^T\psi(x')$ where the range $\mathcal{S}$ of ψ is finite-dimensional. Assume that $f^\dagger$ belongs to the RKHS defined by ψ, i.e., assume that it is of the form $f^\dagger = v^T\psi$ for some v in the feature-space. Then (7.60) reduces to

$$v^T\psi(X) + \sigma Z = Y\,, \tag{7.65}$$

and, in the large data regime, σ^2 can be estimated by

$$\bar{\sigma}^2 := \frac{1}{N}\inf_{w\in\mathcal{S}} \left\|w^T\psi(X) - Y\right\|^2_{\mathbb{R}^N}. \tag{7.66}$$

Our strategy, when the feature map is finite-dimensional, is then to select

$$\gamma = N\bar{\sigma}^2 = \inf_{w\in\mathcal{S}} \left\|w^T\psi(X) - Y\right\|^2_{\mathbb{R}^N}. \tag{7.67}$$

When the kernel is interpolatory (associated with an infinite-dimensional feature map).

If the feature-map associated with the base kernel K is infinite-dimensional (or has more dimensions than we have data points) then it can interpolate the data exactly and the previous strategy cannot be employed since the minimum of (7.66) is zero. A prototypical example (when trying to identify the ancestors of the variable x_1) is $K = K_s$=(7.48) with $\beta_3 > 0$. In this situation, we do not attempt to estimate the level of noise σ but select a prior γ such that the resulting noise-to-signal ratio can effectively differentiate noise from signal. To describe this, observe that the

noise-to-signal ratio (7.63) admits the representer formula

$$\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)} = \frac{Y^T D_\gamma^2 Y}{Y^T D_\gamma Y}\,, \tag{7.68}$$

involving the $N \times N$ matrix

$$D_\gamma := \gamma\big(K(X,X)+\gamma I\big)^{-1}\,. \tag{7.69}$$

Observe that $0 \le D_\gamma \le I$, and

$$\lim_{\gamma\downarrow 0} D_\gamma = 0 \text{ and } \lim_{\gamma\uparrow\infty} D_\gamma = I\,. \tag{7.70}$$

Write (λ_i, e_i) for the eigenpairs of $K(X,X)$ ($K(X,X)e_i = \lambda_i e_i$) where the λ_i are ordered in decreasing order. Then the eigenpairs of D_γ are (ω_i, e_i) where

$$\omega_i := \frac{\gamma}{\gamma+\lambda_i}\,. \tag{7.71}$$

Note that the ω_i are contained in $[0,1]$ and also ordered in decreasing order.

Writing $\bar{Y}_i$ for the orthogonal projection of Y onto e_i, we have

$$\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)} = \frac{\sum_{i=1}^n \omega_i^2 \bar{Y}_i^2}{\sum_{i=1}^n \omega_i \bar{Y}_i^2}\,. \tag{7.72}$$

It follows that if the histogram of the eigenvalues of D_γ is concentrated near 0 or near 1, then the noise-to-signal ratio is non-informative since the prior γ dominates it. To avoid this phenomenon, we select γ so that the eigenvalues of D_γ are well spread out in the sense that the histogram of its eigenvalues has maximum or near-maximum variance (see Fig. 7.6 for a good choice and a bad choice for γ). If the eigenvalues have an algebraic decay, then this is equivalent to taking γ to be the geometric mean of those eigenvalues.

In practice, we use an off-the-shelf optimizer to obtain γ by maximizing the sample variance of $(\omega_i)_{i=1}^n$. If this optimization fails, we default to the median of the eigenvalues. This ensures a balanced, well-spread spectrum for $D\gamma$, with half of the eigenvalues λ_i being lower and half being higher than the median.

Rationale for the choices of γ

The purpose of this section is to present a rationale for the proposed choices for γ in Sec. 7.7 and 7.7. For the choice Sec. 7.7, we present an asymptotic analysis of the

signal-to-noise ratio in the setting of a simple linear regression problem. According to (7.67), γ must scale linearly in N; this scaling is necessary to achieve a ratio that represents the signal-to-noise per sample. Without it (if γ remains bounded as a function of N), this scaling of the signal-to-noise would converge towards 0 as $N \to \infty$. To see how we will now consider a simple example in which we seek to linearly regress the variable y as a function of the variable x, both taken to be scalar (in which case $\psi(x) = x$). Assume that the samples are of the form $Y_i = aX_i + \sigma Z_i$ for $i = 1, \dots, N$, where $a, \sigma \neq 0$, the Z_i are i.i.d. $\mathcal{N}(0,1)$ random variables, and the X_i satisfy $\frac{1}{N}\sum_{i=1}^N X_i = 0$ and $\frac{1}{N}\sum_{i=1}^N X_i^2 = 1$. Then, the signal-to-noise ratio is $\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)}$ with $\mathcal{V}(s) = |v|^2$ and $\mathcal{V}(n) = \frac{1}{\gamma}\sum_{i=1}^N |vX_i - Y_i|^2$ and v is a minimizer of

$$\min_{v\in\mathbb{R}} |v|^2 + \frac{1}{\gamma}\sum_{i=1}^N |vX_i - Y_i|^2 \,. \tag{7.73}$$

In asymptotic $N \to \infty$ regime, we have $v \approx \frac{aN}{\gamma+N}$ and

$$\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)} \approx \frac{\frac{\gamma}{N}a^2}{-a^2(\gamma/N+1) + (a^2+\sigma^2)(\gamma/N+1)^2} \,. \tag{7.74}$$

If γ is bounded independently from N, then $\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)}$ converges towards zero as $N \to \infty$, which is undesirable as it does not represent a signal-to-noise ratio per sample. If $\gamma = N$, then $\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)} \approx \frac{a^2}{4\sigma^2+2a^2}$, which does not converge to 1 as $a \to \infty$ and $\sigma \to 0$, which is also undesirable. If γ is taken as in (7.67), then $\gamma \approx N\sigma^2$ and

$$\frac{\mathcal{V}(s)}{\mathcal{V}(s)+\mathcal{V}(n)} \approx \frac{a^2}{(\sigma^2+1)(a^2+\sigma^2+1)} \,, \tag{7.75}$$

which converges towards 0 as $\sigma \to \infty$ and towards $1/(1+\sigma^2)$ as $a \to \infty$, which has, therefore, the desired properties.

Moving to Sec. 7.7, because the kernel can interpolate the data exactly we can no longer use (7.66) to estimate the level of noise σ. For a finite-dimensional feature map ψ, with data (X, Y), we can decompose $Y = v^T\psi(X) + \sigma Z$ into a signal part Y_s and noise part Y_s, s.t. $Y = Y_s + Y_n$. While Y_s belongs to the linear span of eigenvectors of $K(X, X)$ associated with non-zero eigenvalues, Y_n also activates the eigenvectors associated with with the null space of $K(X, X)$ and the projection of Y onto that null-space is what allows us to derive γ in Sec. 7.7. Since in the interpolatory case, all eigenvalues are strictly positive, we need to choose which eigenvalues are associated with noise differently, as is described in the previous section. With a fixed γ, we see that if $\lambda_i \gg \gamma$, then $\omega_i \approx 0$, which contributes in (7.72) to yield a low noise-to-signal ratio. Similarly, if $\lambda_i \ll \gamma$, this eigenvalue yields a high noise-to-signal

ratio. Thus, we see that the choice of γ assigns a noise level to each eigenvalue. While in the finite-dimensional feature map setting, this assignment is binary, here we perform soft thresholding using $\lambda \mapsto \gamma/(\gamma + \lambda)$ to indicate the level of noise of each eigenvalue. This interpretation sheds light on the selection of γ in equation (7.67). Let ψ represent the feature map associated with K. Assuming the empirical mean of $\psi(X_i)$ is zero, the matrix $K(X, X)$ corresponds to an unnormalized kernel covariance matrix $\psi^T(X)\psi(X)$. Consequently, its eigenvalues correspond to N times the variances of the $\psi(X_i)$ across various eigenspaces. After conducting Ordinary Least Squares regression in the feature space, if the noise variance is estimated as $\bar{\sigma}^2$, then any eigenspace of the normalized covariance matrix whose eigenvalue is lower than $\bar{\sigma}^2$ cannot be recovered due to the noise. Given this, we set the soft thresholding cutoff to be $\gamma = N\bar{\sigma}^2$ for the unnormalized covariance matrix $K(X, X)$.

Z-score/quantile bounds on the noise-to-signal ratio.

If the data is only composed of noise, then an interval of confidence can be obtained on the noise-to-signal ratio. To describe this consider the problem of testing the null hypothesis $\mathbf{H_0} : f^\dagger \equiv 0$ (there is no signal) against the alternative hypothesis $\mathbf{H_1} : f^\dagger \not\equiv 0$ (there is a signal). Under the null hypothesis $\mathbf{H_0}$, the distribution of the noise-to-signal ratio (7.68) is known and it follows that of the random variable

$$B := \frac{Z^T D_\gamma^2 Z}{Z^T D_\gamma Z}. \tag{7.76}$$

Therefore, the quantiles of B can be used as an interval of confidence on the noise-to-signal ratio if $\mathbf{H_0}$ is true. More precisely, selecting β such that $\mathbb{P}[B \leq \beta_\alpha] \approx \alpha$ with $\alpha = 0.05$ as a prototypical example, we expect the noise to signal ratio (7.68) to be, under $\mathbf{H_0}$, to be larger than β_α with probability $\approx 1 - \alpha$. The estimation of β requires Monte-Carlo sampling.

An alternative approach (in the large data regime) to using the quantile β_α is to use the Z-score

$$\mathcal{Z} := \frac{\frac{Y^T D_\gamma^2 Y}{Y^T D_\gamma Y} - \mathbb{E}[B]}{\sqrt{\mathrm{Var}[B]}}, \tag{7.77}$$

after estimating $\mathbb{E}[B]$ and $\mathrm{Var}[B]$ via Monte-Carlo sampling. In particular if $\mathbf{H_0}$ is true then $|\mathcal{Z}| \geq z_\alpha$ should occur with probability $\approx \alpha$ with $z_{0.1} = 1.65$, $z_{0.05} = 1.96$ and $z_{0.01} = 2.58$.

Remark 7.7.2. Although the quantile β_α or the Z-score $\mathcal{Z}$ can be employed to produce an interval of confidence on the noise-to-signal ratio under $\mathbf{H_0}$ we cannot

use them as thresholds for removing nodes from the list of ancestors as discussed in Sec. 7.5 Indeed, observing a noise-to-signal ratio (7.68) below the threshold β_α does not imply that all the signal has been captured by the kernel; it only implies that some signal has been captured by the kernel K. To illustrate this point, consider the setting where one tries to approximate the variable x_1 as a function of the variable x_2. If x_1 is not a function of x_2, but of x_2 and x_3, as in $x_1 = \cos(x_2) + \sin(x_3)$, then applying the proposed approach with Y encoding the values of x_1, X encoding the values of x_2, and the kernel K depending on x_2 could lead to a noise-to-signal ratio below β_α due to the presence of a signal in x_2. Therefore, although we are missing the variable x_3 in the kernel K, we would still observe a possibly low noise-to-signal ratio due to the presence of *some* signal in the data. Summarizing if the data only contains noise then $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)} \geq \beta_\alpha$ should occur with probability $1-\alpha$. If the event $\frac{\mathcal{V}(n)}{\mathcal{V}(s)+\mathcal{V}(n)} < \beta_\alpha$ is observed in the setting of $K = K_{s/t}$=(7.58) where we try to identify the ancestors of x_1, then we can only deduce that $x_3, \ldots, x_d$ contain some signal but perhaps not all of it (we can use this a criterion for pruning x_2).

7.8 Supplementary information on examples.

Algebraic equations.

Although we have used Alg. 12 for the algebraic equations examples presented in Fig. 7.4, Alg. 11 yields the same results with the default signal-to-noise threshold $\tau = 0.5$.

The chemical reaction network.

Consider the chemical reaction network example illustrated in Fig. 7.4.(a). The proposed mechanism for the hydrogenation of ethylene (C_2H_4) to ethane (C_2H_6), is (writing $[H]$ for the concentration of H) modeled by the following system of differential equations:

$$\begin{aligned}
\frac{d[H_2]}{dt} &= -k_1[H_2] + k_{-1}[H]^2 \\
\frac{d[H]}{dt} &= 2k_1[H_2] - 2k_{-1}[H]^2 - k_2[C_2H_4][H] - k_3[C_2H_5][H] \\
\frac{d[C_2H_4]}{dt} &= -k_2[C_2H_4][H] \\
\frac{d[C_2H_5]}{dt} &= k_2[C_2H_4][H] - k_3[C_2H_5][H].
\end{aligned} \tag{7.78}$$

The primary variables are the concentrations $[H_2]$, $[H]$, $[C_2H_4]$, and $[C_2H_5]$ and their time derivatives $\frac{d[H_2]}{dt}$, $\frac{d[H]}{dt}$, $\frac{d[C_2H_4]}{dt}$, and $\frac{d[C_2H_5]}{dt}$. The computational hypergraph encodes the functional dependencies (7.78) associated with the chemical

reactions. The hyperedges of the hypergraph are assumed to be unknown and the primary variables are assumed to be known. Given N samples from the graph of the form

$$\big([H_2](t_i), [H](t_i), [C_2H_4](t_i), [C_2H_5](t_i)\big)_{i=1,\dots,N}, \tag{7.79}$$

our objective is to recover the structure of the hypergraph given by (7.78), representing the functions by hyperedges. We create a dataset of the form (7.79) by integrating 50 trajectories of (7.78) for different initial conditions, and each equispaced 50 times from $t = 0$ to $t = 5$. The dataset is represented in Fig. 7.4.(b) (the time derivatives of concentrations are estimated by taking the derivatives of the interpolants of those concentrations). We impose the information that the derivative variables are function of the non-derivative variables to avoid ambiguity in the recovery, as (7.78) is not the unique representation of the functional relation between nodes in the graph. We implement Alg. 11 with weights $\beta = [0.1, 0.01, 0.001]$ for linear, quadratic, and nonlinear, respectively (Alg. 12 recovers the same hypergraph). The output graph can be seen in Fig. 7.4.(b). We obtain a perfect recovery of the computational graph and a correct identification of the relations being quadratic.

The Google Covid 19 open data.

Consider the example illustrated in Fig. 7.3.(e-k). Categorical data are treated as scalar values, with all variables scaled to achieve a mean of 0 and a variance of 1. We implement three distinct kernel types: linear, quadratic, and Gaussian, with a length scale of 1 for the latter. A weight ratio of $1/10$ is assigned between kernels, signifying that the quadratic kernel is weighted ten times less than the linear kernel. Lastly, the noise parameter, γ, is determined using the optimal value outlined in Sec. 7.7. Initially, a complete graph is constructed using all variables, depicted in Fig. 7.3.(g). This construction is done using only linear and quadratic kernels. The full graph is highly clustered and redundant information is eliminated by selecting representative nodes for each cluster. Eliminating redundant nodes is important for two reasons: firstly, it improves the graph's readability, especially with 31 variables; secondly, it avoids hindering graph discovery. In an extreme case, treating two identical variables as distinct would result in one variable's ancestor simply being its duplicate, yielding an uninformative graph. Subsequently, the graph discovery algorithm is rerun, with reduced variables due to eliminating redundancy, ushering us into a predominantly noisy regime. With fewer variables available, we use additionally the nonlinear kernel. Two indicators are employed to navigate our discovery process: the signal-to-noise ratio and the Z-test. The former quantifies the

degree to which our regression is influenced by noise, while the latter signals the existence of any signal. We follow the procedure in algorithm 12, resulting in the graph presented in Fig. 7.3.(k).

Cell signaling network

Consider the example Fig. 7.1.(l) from (Sachs et al., 2005) and Fig. 7.4.(h-j). To identify the ancestors of each node, we apply the algorithm in two stages. First, we learn the dependencies using only linear and quadratic kernels. Fig. 7.4.(h) identifies the resulting graph learned given a subset of $N = 2,000$ samples chosen uniformly at random from the dataset. We observe that the graph identified by the algorithm consists of four disconnected clusters where the molecule levels in each cluster are closely related by linear or quadratic dependencies (all connections are linear except for the connection between Akt and PKA, which is quadratic). These edges match a subset of the edges found in the gold standard model identified in (Sachs et al., 2005). With perfect dependencies that have no noise, one can define constraints that reduce the total number of variables in the system. For this noisy dataset, we treat these dependencies as forming groups of similar variables and introduce a hierarchical approach to learn the connections between groups. Second, we run the graph discovery algorithm after grouping the molecules into clusters. For each node in the graph, we identified the ancestors of each node by constraining the dependence to be a subset of the clusters. In other words, when identifying the ancestors of a given node i in cluster C, the algorithm is only permitted to (1) use ancestors that do not belong to cluster C, and (2) include all or none of the variables in each cluster (j in cluster $D \neq C$ is listed as an ancestor if and only if all other nodes j' in cluster D are also listed as ancestors). The ancestors were identified using a Gaussian (fully nonlinear) kernel and the number of ancestors were selected manually based on the inflection point in the noise-to-signal ratio. The resulting graph is depicted in Fig. 7.4.(i). Each edge is weighted based on its signal-to-noise ratio. We observe that there is a stronger dependence of the Jnk, PKC, and P38 cluster on the PIP3, Plcg, and PIP2 cluster, which closely matches the gold standard model. As compared to approaches based on acyclic DAGs, however, the graph identified by our algorithm also contains feedback loops between the various molecule levels. Fig. 7.4.(i-j) displays a side-by-side comparison between the graph identified with our method and the graph generated in (Sachs et al., 2005). To aid in this comparison, we have highlighted different clusters in distinct colors. We emphasize that while the Bayesian network analysis in (Sachs et al., 2005)

relied on the control of the sampling of the underlying variables (the simultaneous measurement of multiple phosphorylated protein and phospholipid components in thousands of individual primary human immune system cells, and perturbing these cells with molecular interventions), the reconstruction obtained by our method did not use this information and recovered functional dependencies rather than causal dependencies. Interestingly, the information recovered through our method appears to complement and enhance the findings presented in (Sachs et al., 2005) (e.g., the linear and noiseless dependencies between variables in the JNK cluster is not something that could easily be inferred from the graph produced in (Sachs et al., 2005)).

BCR reaction network

In the high-dimensional example of the BCR reaction network, the computations of terms of the form $y^T k_o(X,X)y$ (i.e., the activations), where $y \in \mathbb{R}^n$ and $k_o(X,X)$ is the o-th coordinate of the quadratic kernel ($k(x_i,x_j) = (1+\langle x_i, x_j\rangle)^2$) becomes the computational bottleneck of our method. If we let $x_1,\ldots,x_n \in \mathbb{R}^p$ be the points and x_i^o be the o-th coordinate of x_i, we can compute the activation of the $o-$th coordinate using

$$k_o(x_i,x_j) = (1+x_i^o x_j^o)^2 - 1 + 2x_i^o x_j^o \langle x_i^{-o}, x_j^{-o}\rangle, \tag{7.80}$$

where k_o is the $o-$th coordinate of the kernel and x_i^{-o} represents the remaining coordinates of x_i. To compute the $n\times n$ kernel matrix of k_o for each $o \in \{1,..,p\}$, we must compute $p\times n\times n$ inner products in $\mathbb{R}^p$, which is a very large computation. Instead, we may use the following reformulation to speed up computations. Notice $\langle x_i,x_j\rangle = x_i^o x_j^o + \langle x_i^{-o}, x_j^{-o}\rangle$, and therefore $k_o(x_i,x_j) = 2x_i^o x_j^o\langle x_i,x_j\rangle + 2x_i^o x_j^o - (x_i^o x_j^o)^2$. Now, define $v^o = (x_i^o y_i)_{i=1}^p$ and $w^o = ((x_i^o)^2 y_i)_{i=1}^p$, and note that

$$y^T K_o y = \sum_{i,j} 2y_i x_i^o y_j x_j^o (1+\langle x_i,x_j\rangle) - \sum_{i,j} y_i y_j (x_i^o x_j^o)^2 \tag{7.81}$$

and so defining $\widetilde{K} = (2(1+\langle x_i,x_j\rangle))_{i,j=1}^n$ we have that

$$y^T K_o y = v^{oT}\widetilde{K}v^o - \left(\sum_{i=1}^p w_i^o\right)^2. \tag{7.82}$$

Note that $\widetilde{K}$ is computed just once for all p, and only v^o and w^o change for every ancestor calculation, which is where the main computational gain comes from. One may find in the GitHub repository of the paper a comparison of the two methods of computations and observe a tenfold speedup. This speedup is even larger in

our implementation of the BCR example, as GPU acceleration enables the second method to run even faster.

BIBLIOGRAPHY

Agrawal, Akshay, Robin Verschueren, Steven Diamond, and Stephen Boyd (2018). "A rewriting system for convex optimization problems". In: *Journal of Control and Decision* 5.1, pp. 42–60.

Aliprantis, C. D. and K. C. Border (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Third. Berlin: Springer, pp. xxii+703.

Aliprantis, C. D. and O. Burkinshaw (1998). *Principles of Real Analysis*. 3rd. Academic Press.

Almroth, Bo O, Perry Stern, and Frank A Brogan (1978). "Automatic choice of global shape functions in structural analysis". In: *Aiaa Journal* 16.5, pp. 525–528.

Altmann, Robert, Patrick Henning, and Daniel Peterseim (2021). "Numerical homogenization beyond scale separation". In: *Acta Numerica* 30, pp. 1–86.

Alvarez, Mauricio A, Lorenzo Rosasco, Neil D Lawrence, et al. (2012). "Kernels for vector-valued functions: A review". In: *Foundations and Trends® in Machine Learning* 4.3, pp. 195–266.

Amsallem, David and Charbel Farhat (2008). "Interpolation method for adapting reduced-order models and application to aeroelasticity". In: *AIAA journal* 46.7, pp. 1803–1813.

Arcangéli, Rémi, María Cruz López de Silanes, and Juan José Torrens (2007). "An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing". In: *Numerische Mathematik* 107.2, pp. 181–211.

Ashton, Greg et al. (2022). "Nested sampling for physical scientists". In: *Nature Reviews Methods Primers* 2.

Babuška, Ivo and John Osborn (2000). "Can a finite element method perform arbitrarily badly?" In: *Mathematics of computation* 69.230, pp. 443–462.

Bădoiu, M., S. Har-Peled, and P. Indyk (2002). "Approximate clustering via core-sets". In: *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*. ACM, pp. 250–257.

Bajgiran, Hamed Hamze, Pau Batlle, Houman Owhadi, Mostafa Samir, Clint Scovel, Mahdy Shirdel, Michael Stanley, and Peyman Tavallali (2022a). "Uncertainty quantification of the 4th kind; optimal posterior accuracy-uncertainty tradeoff with the minimum enclosing ball". In: *Journal of Computational Physics* 471. P.B contributed to the conceptualization, writing, and numerical experiments in Sections 1 and 6., p. 111608. DOI: https://doi.org/10.1016/j.jcp.2022.111608.

Bajgiran, Hamed Hamze, Pau Batlle, Houman Owhadi, Mostafa Samir, Clint Scovel, Mahdy Shirdel, Michael Stanley, and Peyman Tavallali (Dec. 2022b). "Uncertainty quantification of the 4th kind: Optimal posterior accuracy-uncertainty tradeoff with the minimum enclosing ball". In: *Journal of Computational Physics* 471, p. 111608. DOI: 10.1016/j.jcp.2022.111608. URL: https://doi.org/10.1016/j.jcp.2022.111608.

Baptista, Ricardo, Youssef Marzouk, Rebecca E Morrison, and Olivier Zahm (2021). "Learning non-Gaussian graphical models via Hessian scores and triangular transport". In: *arXiv preprint arXiv:2101.03093*.

Bastos, Leonardo S and Anthony O'hagan (2009). "Diagnostics for Gaussian process emulators". In: *Technometrics* 51.4, pp. 425–438.

Batlle, Pau, Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart (2023). "Error analysis of kernel/GP methods for nonlinear and parametric PDEs". In: *Journal of Computational Physics*.
P.B contributed to numerical results of Section 4, and writing. DOI: https://doi.org/10.1016/j.jcp.2024.113488.

Batlle, Pau, Matthieu Darcy, Bamdad Hosseini, and Houman Owhadi (2023). "Kernel methods are competitive for operator learning". In: *Journal of Computational Physics*.

Batlle, Pau, Michael Stanley, Pratik Patil, Mikael Kuusela, and Houman Owhadi (2023). "Optimization-based frequentist confidence intervals for functionals in constrained inverse problems: Resolving the Burrus conjecture". In: *arXiv preprint arXiv:2310.02461*.

Beck, Joakim, Raul Tempone, Fabio Nobile, and Lorenzo Tamellini (2012). "On the optimal polynomial approximation of stochastic PDEs by Galerkin and collocation methods". In: *Mathematical Models and Methods in Applied Sciences* 22.09, p. 1250023.

Becker, Roland, Maximilian Brunner, Michael Innerberger, Jens Markus Melenk, and Dirk Praetorius (2023). "Cost-optimal adaptive iterative linearized FEM for semilinear elliptic PDEs". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 57.4, pp. 2193–2225.

Bendsoe, Martin Philip and Ole Sigmund (2003). *Topology optimization: theory, methods, and applications*. Springer Science & Business Media.

Berger, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Berger, Roger L. and Dennis D. Boos (1994). "P Values Maximized Over a Confidence Set for the Nuisance Parameter". In: *Journal of the American Statistical Association* 89, pp. 1012–1016.

Berlinet, Alain and Christine Thomas-Agnan (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bertsekas, D. P., A. Nedić, and A. Ozdaglar (2003). *Convex Analysis and Optimization*. Athena Scientific Optimization and Computation Series. Athena Scientific.

Bhattacharya, Kaushik, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart (2021). "Model Reduction And Neural Networks For Parametric PDEs". en. In: *The SMAI Journal of computational mathematics* 7, pp. 121–157.

Bogachev, Vladimir Igorevich (1998). *Gaussian measures*. American Mathematical Society.

Böhmer, Klaus and Robert Schaback (2013). "A nonlinear discretization theory". In: *Journal of computational and applied mathematics* 254, pp. 204–219.

– (2020). "A nonlinear discretization theory for meshfree collocation methods applied to quasilinear elliptic equations". In: *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik* 100.10, e201800170.

Boncoraglio, Gabriele and Charbel Farhat (2021). "Active manifold and model-order reduction to accelerate multidisciplinary analysis and optimization". In: *AIAA Journal* 59.11, pp. 4739–4753.

Bonnans, J. F. and A. Shapiro (2000). *Perturbation Analysis of Optimization Problems*. Springer-Verlag.

Boucheron, Stéphane and Maud Thomas (2012). "Concentration inequalities for order statistics". In: *Electronic Communications in Probability* 17.none, pp. 1–12. DOI: 10.1214/ECP.v17-2210. URL: https://doi.org/10.1214/ECP.v17-2210.

Boullé, Nicolas and Alex Townsend (2022). "Learning elliptic partial differential equations with randomized linear algebra". In: *Foundations of Computational Mathematics*, pp. 1–31.

Bourdais, Théo, Pau Batlle, Xianjin Yang, Ricardo Baptista, Nicolas Rouquette, and Houman Owhadi (2023). "Codiscovering graphical structure and functional relationships within data: A Gaussian Process framework for connecting the dots". In: *Proceedings of the National Academy of Sciences (PNAS)*.
P.B contributed to conceptualization, early prototyping, and writing. DOI: https://doi.org/10.1073/pnas.2403449121.

Boyd, Stephen and Lievan Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Bradbury, James, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang (2018). *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. URL: http://github.com/google/jax.

Brunton, Steven L, Joshua L Proctor, and J Nathan Kutz (2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the national academy of sciences* 113.15, pp. 3932–3937.

Buchner, Johannes (2023). "Nested Sampling Methods". In: *arXiv preprint arXiv:2101.09675v4*.

Burrus, W. R. (1965). *Utilization of a Priori Information by Means of Mathematical Programming in the Statistical Interpretation of Measured Distributions*. Oak Ridge National Laboratory. URL: https://books.google.com/books?id=ubUPYiYzrRcC.

Carpenter, James (1999). "Test inversion bootstrap confidence intervals". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.1, pp. 159–172.

Casella, G. and R. L. Berger (2002). *Statistical Inference*. Thomson Learning Inc.

Casella, George and Roger L. Berger (2002). *Statistical Inference*. Second. Duxbery.

Cash, Webster (1979). "Parameter estimation in astronomy through application of the likelihood ratio". In: *The Astrophysical Journal* 228, pp. 939–947.

Chen, Jiong, Florian Schäfer, Jin Huang, and Mathieu Desbrun (2021). "Multiscale cholesky preconditioning for ill-conditioned problems". In: *ACM Transactions on Graphics (TOG)* 40.4, pp. 1–13.

Chen, Tianping and Hong Chen (1995a). "Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks". In: *IEEE Transactions on Neural Networks* 6.4, pp. 904–910.

– (1995b). "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems". In: *IEEE Transactions on Neural Networks* 6.4, pp. 911–917.

Chen, Yifan, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart (2021a). *Solving and Learning Nonlinear PDEs with Gaussian Processes*.

– (2021b). "Solving and learning nonlinear PDEs with Gaussian processes". In: *Journal of Computational Physics* 447, p. 110668. ISSN: 0021-9991.

– (2021c). "Solving and learning nonlinear PDEs with Gaussian processes". In: *Journal of Computational Physics* 447, p. 110668.

Chen, Yifan, Houman Owhadi, and Florian Schäfer (2024). "Sparse Cholesky factorization for solving nonlinear PDEs via Gaussian processes". In: *Mathematics of Computation*.

Chen, Yifan, Houman Owhadi, and Andrew Stuart (2021a). "Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation". In: *Mathematics of Computation*.

Chen, Yifan, Houman Owhadi, and Andrew Stuart (2021b). "Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation". In: *Mathematics of Computation* 90.332, pp. 2527–2578.

Chen, Yuansi, Raaz Dwivedi, Martin J. Wainwright, and Bin Yu (2018). "Fast MCMC Sampling Algorithms on Polytopes". In: *Journal of Machine Learning Research* 19, pp. 1–86.

Cheung, Ka Chun, Leevan Ling, and Robert Schaback (2018). "H^2-Convergence of Least-Squares Kernel Collocation Methods". In: *SIAM Journal on Numerical Analysis* 56.1, pp. 614–633.

Chickering, David Maxwell (2002). "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov, pp. 507–554.

Chkifa, Abdellah, Albert Cohen, Ronald DeVore, and Christoph Schwab (Nov. 2012). "Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 47.1, pp. 253–280.

Chkifa, Abdellah, Albert Cohen, and Christoph Schwab (2014). "High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs". In: *Foundations of Computational Mathematics* 14.4, pp. 601–633.

Chkrebtii, Oksana A, David A Campbell, Ben Calderhead, and Mark A Girolami (2016). "Bayesian solution uncertainty quantification for differential equations". In: *Bayesian Analysis* 11.4, pp. 1239–1267.

Cialenco, Igor, Gregory E Fasshauer, and Qi Ye (2012). "Approximation of stochastic partial differential equations by a kernel-based collocation method". In: *International Journal of Computer Mathematics* 89.18, pp. 2543–2561.

CMS Collaboration (2016). "Measurement of differential cross sections for Higgs boson production in the diphoton decay channel in pp collisions at $\sqrt{s} = 8$ TeV". In: *The European Physical Journal C* 76.13. DOI: https://doi.org/10.1140/epjc/s10052-015-3853-3.

– (2019). "Measurement of inclusive and differential Higgs boson production cross sections in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2019.183. DOI: https://doi.org/10.1007/JHEP01(2019)183.

Cockayne, Jon, Chris Oates, Tim Sullivan, and Mark Girolami (2017). "Probabilistic numerical methods for PDE-constrained Bayesian inverse problems". In: *AIP Conference Proceedings*, p. 060001.

Cockayne, Jon, Chris J Oates, Timothy John Sullivan, and Mark Girolami (2019). "Bayesian probabilistic numerical methods". In: *SIAM Review* 61.4, pp. 756–789.

Cohen, Albert and Ronald DeVore (2015). "Approximation of high-dimensional parametric PDEs". In: *Acta Numerica* 24, pp. 1–159.

Cohen, Albert, Ronald DeVore, and Christoph Schwab (2010). "Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs". In: *Foundations of Computational Mathematics* 10.6, pp. 615–646.

Cranmer, Kyle, Johann Brehmer, and Gilles Louppe (2020). "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48, pp. 30055–30062.

Cranmer, Kyle, Juan Pavez, and Gilles Louppe (2016). *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*. arXiv:1506.02169 [stat.AP].

Dalmasso, Niccolo, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B. Lee (2024). "Likelihood-Free Frequentist Inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference". In: *Electronic Journal of Statistics*, pp. 5045–5090.

Dalmasso, Niccolò, Rafael Izbicki, and Ann B. Lee (2020). "Confidence sets and hypothesis testing in a likelihood-free inference setting". In: *International Conference on Machine Learning*. PMLR, pp. 2323–2334.

Dalmasso, Niccolò, Luca Masserano, David Zhao, Rafael Izbicki, and Ann B. Lee (2023). "Likelihood-free frequentist inference: Bridging Classical Statistics and Machine Learning for Reliable Simulator-Based Inference". In: *arXiv preprint arXiv:2107.03920*.

Darcy, Matthieu, Boumediene Hamzi, Giulia Livieri, Houman Owhadi, and Peyman Tavallali (2023). "One-shot learning of stochastic differential equations with data adapted kernels". In: *Physica D: Nonlinear Phenomena* 444, p. 133583.

Data, MIT Critical, Justin D Salciccioli, Yves Crutain, Matthieu Komorowski, and Dominic C Marshall (2016). "Sensitivity analysis and model validation". In: *Secondary analysis of electronic health records*, pp. 263–271.

De Hoop, Maarten, Daniel Zhengyu Huang, Elizabeth Qian, and Andrew M Stuart (2022). "The cost-accuracy trade-off in operator learning with neural networks". In: *arXiv preprint arXiv:2203.13181*.

De Ryck, Tim and Siddhartha Mishra (2022). "Error analysis for physics-informed neural networks (PINNs) approximating Kolmogorov PDEs". In: *Advances in Computational Mathematics* 48.6, pp. 1–40.

Deng, Beichuan, Yeonjong Shin, Lu Lu, Zhongqiang Zhang, and George Em Karniadakis (2022). "Approximation rates of DeepONets for learning operators arising from advection–diffusion equations". In: *Neural Networks* 153, pp. 411–426.

Diamond, Steven and Stephen Boyd (2016). "CVXPY: A Python-embedded modeling language for convex optimization". In: *Journal of Machine Learning Research* 17.83, pp. 1–5.

Dirmeyer, Paul A, Pierre Gentine, Michael B Ek, and Gianpaolo Balsamo (2019). "Land surface processes relevant to sub-seasonal to seasonal (S2S) prediction". In: *Sub-Seasonal to Seasonal Prediction*. Elsevier, pp. 165–181.

Donoho, David L. (1994). “Statistical estimation and optimal recovery”. In: *The Annals of Statistics* 22.1, pp. 238–270.

Doostan, Alireza and Houman Owhadi (2011). “A non-adapted sparse approximation of PDEs with stochastic inputs”. In: *Journal of Computational Physics* 230.8, pp. 3015–3034.

Drton, Mathias and Marloes H Maathuis (2017). “Structure learning in graphical modeling”. In: *Annual Review of Statistics and Its Application* 4, pp. 365–393.

Dudley, R. (2009). *Statistics for Applications*. 18.443. MIT OpenCourseWare.

Economon, Thomas D, Francisco Palacios, Sean R Copeland, Trent W Lukaczyk, and Juan J Alonso (2016). “SU2: An open-source suite for multiphysics simulation and design”. In: *Aiaa Journal* 54.3, pp. 828–846.

Fan, Yuwei, Cindy Orozco Bohorquez, and Lexing Ying (2019). “BCR-Net: A neural network based on the nonstandard wavelet form”. In: *Journal of Computational Physics* 384, pp. 1–15.

Fan, Yuwei, Jordi Feliu-Faba, Lin Lin, Lexing Ying, and Leonardo Zepeda-Núnez (2019). “A multiscale neural network based on hierarchical nested bases”. In: *Research in the Mathematical Sciences* 6.2, pp. 1–28.

Fan, Yuwei, Lin Lin, Lexing Ying, and Leonardo Zepeda-Núnez (2019). “A multiscale neural network based on hierarchical matrices”. In: *Multiscale Modeling & Simulation* 17.4, pp. 1189–1213.

Fasshauer, Gregory E (1999). “Solving differential equations with radial basis functions: multilevel methods and smoothing”. In: *Advances in computational mathematics* 11.2, pp. 139–159.

Feischl, Michael and Daniel Peterseim (2020). “Sparse compression of expected solution operators”. In: *SIAM Journal on Numerical Analysis* 58.6, pp. 3144–3164.

Feldman, Gary J. and Robert D. Cousins (Apr. 1998). “Unified approach to the classical statistical analysis of small signals”. In: *Physical Review D* 57.7, pp. 3873–3889. DOI: 10.1103/physrevd.57.3873. URL: https://doi.org/10.11032Fphysrevd.57.3873.

Feyel, Frédéric and Jean-Louis Chaboche (2000). “FE2 multiscale approach for modelling the elastoviscoplastic behaviour of long fibre SiC/Ti composite materials”. In: *Computer methods in applied mechanics and engineering* 183.3-4, pp. 309–330.

Feynman, Richard P (1998). “Cargo cult science”. In: *The art and science of analog circuit design*. Elsevier, pp. 55–61.

Fish, Jacob, Kamlun Shek, Muralidharan Pandheeradi, and Mark S Shephard (1997). "Computational plasticity for composite structures based on mathematical homogenization: Theory and practice". In: *Computer methods in applied mechanics and engineering* 148.1-2, pp. 53–73.

Fisher, Eyal, Regev Schweiger, and Saharon Rosset (2020). "Efficient construction of test inversion confidence intervals using quantile regression". In: *Journal of Computational and Graphical Statistics* 29.1, pp. 140–148.

Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch (2015). "ECCO version 4: an integrated framework for non-linear inverse modeling and global ocean state estimation". In: *Geoscientific Model Development*.

Fornberg, Bengt and Natasha Flyer (2015). "Solving PDEs with radial basis functions". In: *Acta Numerica* 24, pp. 215–258.

Franke, Carsten and Robert Schaback (1998a). "Convergence order estimates of meshless collocation methods using radial basis functions". In: *Advances in computational mathematics* 8.4, pp. 381–399.

– (1998b). "Solving partial differential equations by collocation using radial basis functions". In: *Applied Mathematics and Computation* 93.1, pp. 73–82.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3, pp. 432–441.

Fuselier, Edward and Grady B Wright (2012). "Scattered data interpolation on embedded submanifolds with restricted positive definite kernels: Sobolev error estimates". In: *SIAM Journal on Numerical Analysis* 50.3, pp. 1753–1776.

Garthwaite, Paul H. and Stephen T. Buckland (1992). "Generating Monte Carlo Confidence Intervals by the Robbins–Monro Process". In: *Journal of the Royal Statistical Society* 41.1, pp. 159–171.

Gärtner, B. (1999). "Fast and robust smallest enclosing balls". In: *European Symposium on Algorithms*. Springer, pp. 325–338.

Geng, Xinbo and Le Xie (2019a). "Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization". In: *Annual Reviews in Control* 47, pp. 341–363. ISSN: 1367-5788. DOI: 10.1016/j.arcontrol.2019.05.005. URL: http://dx.doi.org/10.1016/j.arcontrol.2019.05.005.

– (2019b). "Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization". In: *Annual reviews in control* 47, pp. 341–363.

Ghanem, Roger G and Pol D Spanos (2003). *Stochastic finite elements: a spectral approach*. Dover Publications.

Giesl, Peter and Holger Wendland (2007). "Meshless collocation: Error estimates with application to dynamical systems". In: *SIAM Journal on Numerical Analysis* 45.4, pp. 1723–1741.

Gilbarg, David and Neil S Trudinger (1977). *Elliptic Partial Differential Equations of second order*. Springer.

Gin, Craig R, Daniel E Shea, Steven L Brunton, and J Nathan Kutz (2021). "DeepGreen: deep learning of Green's functions for nonlinear boundary value problems". In: *Scientific reports* 11.1, p. 21614.

Gittell, Jody Hoffer and Hebatallah Naim Ali (2021). *Relational analytics: Guidelines for analysis and action*. Routledge.

Glymour, Madelyn, Judea Pearl, and Nicholas P Jewell (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Goel, A., P. Indyk, and K. R. Varadarajan (2001). "Reductions among high dimensional proximity problems". In: *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, pp. 769–778.

Gouriéroux, Christian, Alberto Holly, and Alain Monfort (1982). "Likelihood Ratio Test, Wald Test, and Kuhn–Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters". In: *Econometrica* 50, pp. 63–80.

Grossmann, Tamara G, Urszula Julia Komorowska, Jonas Latz, and Carola-Bibiane Schönlieb (2023). "Can Physics-Informed Neural Networks beat the Finite Element Method?" In: *arXiv preprint arXiv:2302.04107*.

Gunzburger, Max D, Clayton G Webster, and Guannan Zhang (2014). "Stochastic finite element methods for partial differential equations with random input data". In: *Acta Numerica* 23, pp. 521–650.

Gutmann, Michael U. and Jukka Corander (2015). *Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models*. eprint: `arXiv:1501.03291`.

Hahn, Gerald J. and William Q. Meeker (1991). *Statistical Intervals*. Wiley.

Hamzi, Boumediene, Romit Maulik, and Houman Owhadi (2021a). "Simple, low-cost and accurate data-driven geophysical forecasting with learned kernels". In: *Proceedings of the Royal Society A* 477.2252, p. 20210326.

– (2021b). "Simple, low-cost and accurate data-driven geophysical forecasting with learned kernels". In: *Proceedings of the Royal Society A* 477.2252, p. 20210326.

Hamzi, Boumediene and Houman Owhadi (2021). "Learning dynamical systems from data: a simple cross-validation perspective, part I: parametric kernel flows". In: *Physica D: Nonlinear Phenomena* 421, p. 132817.

Hamzi, Boumediene, Houman Owhadi, and Yannis Kevrekidis (2023). "Learning dynamical systems from data: A simple cross-validation perspective, part iv: case with partial observations". In: *Physica D: Nonlinear Phenomena* 454, p. 133853.

Hansen, Per Christian (1992). "Analysis of discrete ill-posed problems by means of the L-curve". In: *SIAM Review* 34.4, pp. 561–580.

Hansen, Per Christian and Dianne Prost O'Leary (1993). "The use of the L-curve in the regularization of discrete ill-posed problems". In: *SIAM Journal on Scientific Computing* 14.6, pp. 1487–1503.

Heinrich, Lukas (2022). "Learning Optimal Test Statistics in the Presence of Nuisance Parameters". In: *arXiv preprint arXiv:2203.13079*.

Hesthaven, J.S. and S. Ubbiali (2018). "Non-intrusive reduced order modeling of nonlinear problems using neural networks". In: *Journal of Computational Physics* 363, pp. 55–78.

Hesthaven, Jan S, Gianluigi Rozza, Benjamin Stamm, et al. (2016). *Certified reduced basis methods for parametrized partial differential equations*. Vol. 590. Springer.

Höcker, Andreas and Vakhtang Kartvelishvili (1996). "SVD approach to data unfolding". In: *Nuclear Instruments and Methods in Physics Research A* 372, pp. 469–481.

Hon, YC and Robert Schaback (2008). "Solvability of partial differential equations by meshless kernel methods". In: *Advances in Computational Mathematics* 28.3, pp. 283–299.

Huang, Daniel Zhengyu, Tapio Schneider, and Andrew M Stuart (2022). "Iterated Kalman methodology for inverse problems". In: *Journal of Computational Physics* 463, p. 111262.

Ishihara, Toru (2001). "Enumeration of hypergraphs". In: *European Journal of Combinatorics* 22.4, pp. 503–509.

Javanmard, Adel and Andrea Montanari (2014). "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression". In: *Journal of Machine Learning Research* 15.82, pp. 2869–2909. URL: http://jmlr.org/papers/v15/javanmard14a.html.

Kadri, Hachem, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren (2016). "Operator-valued kernels for learning from functional response data". In.

Kalmikov, Alexander G. and Patrick Heimbach (2014). "A Hessian-based method for uncertainty quantification in global ocean state estimation". In: *SIAM Journal on Scientific Computing* 36, S267–S295.

Kannan, Ravindran and Hariharan Narayanan (2012). "Random Walks on Polytopes and an Affine Interior Point Method for Linear Programming". In: *Mathematics of Operations Research* 37, pp. 1–20.

Kansa, Edward J (1990a). “Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics—I surface approximations and partial derivative estimates”. In: *Computers & Mathematics with applications* 19.8-9, pp. 127–145.

– (1990b). “Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics—II solutions to parabolic, hyperbolic and elliptic partial differential equations”. In: *Computers & mathematics with applications* 19.8-9, pp. 147–161.

Karr, A. F. (1983). “Extreme points of certain sets of probability measures, with applications”. In: *Mathematics of Operations Research* 8.1, pp. 74–85.

Kaveh, Hojjat, Pau Batlle, Mateo Acosta, Pranav Kulkarni, Stephen J Bourne, and Jean Philippe Avouac (2024). “Induced seismicity forecasting with uncertainty quantification: Application to the Groningen gas field”. In: *Seismological Research Letters* 95.2A.
P.B contributed to the conceptualization of the Uncertainty Quantification algorithm and writing of its corresponding section, pp. 773–790. DOI: https://doi.org/10.1785/0220230179.

Kempf, Rüdiger, Holger Wendland, and Christian Rieger (2019). “Kernel-based reconstructions for parametric PDEs”. In: *IWMMPDE 2017: Meshfree Methods for Partial Differential Equations IX 9*. Springer, pp. 53–71.

Kempthorne, P. J. (1987). “Numerical specification of discrete least favorable prior distributions”. In: *SIAM Journal on Scientific and Statistical Computing* 8.2, pp. 171–184.

Kennedy, Marc C and Anthony O’Hagan (2001). “Bayesian calibration of computer models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3, pp. 425–464.

Khoo, Yuehaw, Jianfeng Lu, and Lexing Ying (2021). “Solving parametric PDE problems with artificial neural networks”. In: *European Journal of Applied Mathematics* 32.3, pp. 421–435.

Khoo, Yuehaw and Lexing Ying (2019). “SwitchNet: a neural network model for forward and inverse scattering problems”. In: *SIAM Journal on Scientific Computing* 41.5, A3182–A3201.

Kibzun, A I and Y S Kan (Aug. 1997). “Stochastic Programming Problems with Probability and Quantile Functions”. In: *Journal of the Operational Research Society* 48.8, pp. 849–849. ISSN: 1476-9360. DOI: 10.1057/palgrave.jors.2600833. URL: http://dx.doi.org/10.1057/palgrave.jors.2600833.

Kingma, D. P. and J. Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

Kissas, Georgios, Jacob H Seidman, Leonardo Ferreira Guilhoto, Victor M Preciado, George J Pappas, and Paris Perdikaris (2022). "Learning operators with coupled attention". In: *Journal of Machine Learning Research* 23.215, pp. 1–63.

Koenker, R. and G. Bassett Jr (1978). "Regression quantiles". In: *Econometrica: journal of the Econometric Society*, pp. 33–50.

Koenker, Roger (2005). *Quantile regression*. Vol. 38. Cambridge university press.

Koenker, Roger and Kevin F. Hallock (2001). "Quantile Regression". In: *Journal of Economic Perspectives* 15.4, pp. 143–156.

Kovachki, Nikola, Samuel Lanthaler, and Siddhartha Mishra (2021). "On universal approximation and error bounds for fourier neural operators". In: *The Journal of Machine Learning Research* 22.1, pp. 13237–13312.

Kovachki, Nikola, Burigede Liu, Xingsheng Sun, Hao Zhou, Kaushik Bhattacharya, Michael Ortiz, and Andrew Stuart (2022). "Multiscale modeling of materials: Computing, data science, uncertainty and goal-oriented optimization". In: *Mechanics of Materials* 165, p. 104156.

Krischer, K, R Rico-Martínez, IG Kevrekidis, HH Rotermund, G Ertl, and JL Hudson (1993). "Model identification of a spatiotemporally varying catalytic reaction". In: *AIChE Journal* 39.1, pp. 89–98.

Kröpfl, Fabian, Roland Maier, and Daniel Peterseim (2022). "Operator compression with deep neural networks". In: *Advances in Continuous and Discrete Models* 2022.1, pp. 1–23.

Kuusela, Mikael (July 2016). "Uncertainty quantification in unfolding elementary particle spectra at the Large Hadron Collider". PhD thesis. École Polytechnique Fédérale de Lausanne.

Kuusela, Mikael and Victor M Panaretos (2015a). "Statistical unfolding of elementary particle spectra: Empirical Bayes estimation and bias-corrected uncertainty quantification". In: *The Annals of Applied Statistics* 9.3, pp. 1671–1705.

– (2015b). "Statistical Unfolding of Elementary Particle Spectra: Empirical Bayes Estimation and Bias-Corrected Uncertainty Quantification". In: *The Annals of Applied Statistics* 9.3, pp. 1671–1705.

Kuusela, Mikael and Philip B. Stark (2017a). "Shape-constrained uncertainty quantification in unfolding steeply falling elementary particle spectra". In: *Annals of Applied Statistics*.

– (2017b). "Shape-constrained uncertainty quantification in unfolding steeply falling elementary particle spectra". In: *The Annals of Applied Statistics* 11.3, pp. 1671–1710.

Larkin, F. M. (1972). "Gaussian measure in Hilbert space and applications in numerical analysis". In: *Journal of Mathematics* 2.3.

Le Maitre, Olivier and Omar M Knio (2010). *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media.

Lee, Jaehoon, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein (2017). "Deep neural networks as Gaussian processes". In: *arXiv preprint arXiv:1711.00165*.

Lee, John M (2012). *Introduction to Smooth Manifolds*. Springer.

Lehmann, E. L. and Joseph P. Romano (2008). *Testing Statistical Hypotheses*. Springer.

Li, Zongyi, Nikola Kovachki, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar (2020). "Fourier Neural Operator for Parametric Partial Differential Equations". In: *International Conference on Learning Representations*.

Li, Zongyi, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar (2020). *Fourier Neural Operator for Parametric Partial Differential Equations*.

Lim, T. and R. J. McCann (2021). "Geometrical Bounds for Variance and Recentered Moments". In: *Mathematics of Operations Research*.

Ling, Leevan, Roland Opfer, and Robert Schaback (2006). "Results on meshless collocation techniques". In: *Engineering Analysis with Boundary Elements* 30.4, pp. 247–253.

Ling, Leevan and Robert Schaback (2008). "Stable and convergent unsymmetric meshless collocation methods". In: *SIAM Journal on Numerical Analysis* 46.3, pp. 1097–1115.

Liu, Junjie, Kevin Bowman, Lee Meemong, et al. (2016). "Carbon Monitoring system flux estimation and attribution: impact of ACOS-GOSAT X_{CO_2} sampling on the inference of terrestrial biospheric sources and sinks". In: *Tellus B: Chemical and Physical Meteorology* 66. DOI: https://doi.org/10.3402/tellusb.v66.22486.

Loman, Torkel E., Yingbo Ma, Vasily Ilin, Shashi Gowda, Niklas Korsbo, Nikhil Yewale, Chris Rackauckas, and Samuel A. Isaacson (Oct. 2023). *Catalyst: Fast and flexible modeling of reaction networks*. en. Ed. by Christos A. Ouzounis. DOI: 10.1371/journal.pcbi.1011530. URL: http://dx.doi.org/10.1371/journal.pcbi.1011530.

Long, Da, Nicole Mrvaljevic, Shandian Zhe, and Bamdad Hosseini (2022). "A Kernel Approach for PDE Discovery and Operator Learning". In: *arXiv preprint arXiv:2210.08140*.

Long, Da, Zheng Wang, Aditi Krishnapriyan, Robert Kirby, Shandian Zhe, and Michael Mahoney (2022). "AutoIP: A United Framework to Integrate Physics into Gaussian Processes". In: *International Conference on Machine Learning*. PMLR, pp. 14210–14222.

Lopez-Paz, David, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin (2015). "Towards a learning theory of cause-effect inference". In: *International Conference on Machine Learning*. PMLR, pp. 1452–1461.

Lotka, A. J. (1920). "Analytical Note on Certain Rhythmic Relations in Organic Systems". In: *Proceedings of the National Academy of Sciences of the United States of America* 6.7, pp. 410–415.

Lovasz, Laszlo (1999). "Hit-and-run mixes fast". In: *Mathematical Programming* 86, pp. 443–461.

Lovasz, Laszlo and Santosh Vempala (2006). "Hit-and-Run from a Corner". In: *SIAM Journal on Computing* 35, pp. 985–1005.

Lu, Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis (Mar. 2021). "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators". In: *Nature Machine Intelligence* 3.3, pp. 218–229.

Lu, Lu, Xuhui Meng, Shengze Cai, Zhiping Mao, Somdatta Goswami, Zhongqiang Zhang, and George Em Karniadakis (2022). "A comprehensive and fair comparison of two neural operators (with practical extensions) based on FAIR data". In: *Computer Methods in Applied Mechanics and Engineering* 393, p. 114778. ISSN: 0045-7825.

Lu, Yiping, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet (2021). "Machine learning for elliptic PDEs: fast rate generalization bound, neural scaling law and minimax optimality". In: *arXiv preprint arXiv:2110.06897*.

Lucia, David J, Philip S Beran, and Walter A Silva (2004). "Reduced-order modeling: new approaches for computational physics". In: *Progress in aerospace sciences* 40.1-2, pp. 51–117.

Maday, Yvon, Anthony T Patera, and Gabriel Turinici (2002). "A priori convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations". In: *Journal of Scientific Computing* 17, pp. 437–446.

Målqvist, Axel and Daniel Peterseim (2014). "Localization of elliptic multiscale problems". In: *Mathematics of Computation* 83.290, pp. 2583–2603.

Martin, James, Lucas C Wilcox, Carsten Burstedde, and Omar Ghattas (2012). "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion". In: *SIAM Journal on Scientific Computing* 34.3, A1460–A1487.

Martinez-Cantin, Ruben (2014). *BayesOpt: A Bayesian Optimization Library for Nonlinear Optimization, Experimental Design and Bandits*. eprint: arXiv:1405.7430.

Martins, Joaquim RRA and Andrew B Lambe (2013). "Multidisciplinary design optimization: a survey of architectures". In: *AIAA journal* 51.9, pp. 2049–2075.

Marzari, Nicola, Arash A Mostofi, Jonathan R Yates, Ivo Souza, and David Vanderbilt (2012). "Maximally localized Wannier functions: Theory and applications". In: *Reviews of Modern Physics* 84.4, p. 1419.

Masserano, Luca, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, and Ann Lee (2023). "Simulation-Based Inference with WALDO: Confidence Regions by Leveraging Prediction Algorithms or Posterior Estimators for Inverse Problems". In: *The International Conference on Artificial Intelligence and Statistics*.

Masserano, Luca, Alexander Shen, Michele Doro, Tommaso Dorigo, Rafael Izbiki, and Ann Lee (2024). "Classification Under Nuisance Parameters and Generalized Label Shift in Likelihood-Free Inference". In: *arXiv preprint arXiv:2402.05330*.

McLean, William (2000). *Strongly elliptic systems and boundary integral equations*. Cambridge university press.

Meinshausen, Nicolai (2006). "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7, pp. 983–999.

Melenk, Jens M (2000). "On n-widths for elliptic problems". In: *Journal of mathematical analysis and applications* 247.1, pp. 272–289.

Meng, Rui and Xianjin Yang (2022). "Sparse Gaussian processes for solving nonlinear PDEs". In: *arXiv preprint arXiv:2205.03760*.

Micchelli, Charles A and Theodore J Rivlin (1977). "A survey of optimal recovery". In: *Optimal estimation in approximation theory*, pp. 1–54.

Mika, Sebastian, Bernhard Schölkopf, Alexander J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch (1998). "Kernel PCA and De-noising in feature spaces." In: *NIPS*. Vol. 11, pp. 536–542.

Molenberghs, Geert and Geert Verbeke (2007). "Likelihood ratio, score, and Wald tests in a constrained parameter space". In: *The American Statistician* 61.1, pp. 22–27.

Montel, Noemi Anau, James Alvey, and Christoph Weniger (2023). "Scalable Inference with Autoregressive Neural Ratio Estimation". In: *arXiv preprint arXiv:2308:08597v1*.

Morgan, Stephen L and Christopher Winship (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Mou, Chenchen, Xianjin Yang, and Chao Zhou (2022). "Numerical methods for mean field games based on Gaussian processes and Fourier features". In: *Journal of Computational Physics* 460, p. 111188.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. (2017). "Kernel mean embedding of distributions: A review and beyond". In: *Foundations and Trends® in Machine Learning* 10.1-2, pp. 1–141.

Murphy, S. A. (1995). "Likelihood Ratio-Based Confidence Intervals in Survival Analysis". In: *Journal of the American Statistical Association* 90.432, pp. 1399–1405. DOI: 10.1080/01621459.1995.10476645. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1995.10476645. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1995.10476645.

Narayanan, Hariharan (2016). "Ranomized Interior Point Methods for Sampling and Optimization". In: *The Annals of Applied Probability* 26, pp. 597–641.

Neal, Radford M (1996). "Priors for infinite networks". In: *Bayesian learning for neural networks*, pp. 29–53.

Neale, Michael C. and Michael B. Miller (1997). "The use of likelihood-based confidence intervals in genetic models". In: *Behavior Genetics* 27, pp. 113–120.

Neumann, J. von (1928). "Zur Theorie der Gesellschaftsspiele". In: *Math. Ann.* 100.1, pp. 295–320. ISSN: 0025-5831. DOI: 10.1007/BF01448847. URL: http://dx.doi.org/10.1007/BF01448847.

Nobile, Fabio, Raul Tempone, and Clayton G Webster (2008). "An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data". In: *SIAM Journal on Numerical Analysis* 46.5, pp. 2411–2442.

Nobile, Fabio, Raúl Tempone, and Clayton G Webster (2008). "A sparse grid stochastic collocation method for partial differential equations with random input data". In: *SIAM Journal on Numerical Analysis* 46.5, pp. 2309–2345.

Nocedal, J. and S. Wright (2006). *Numerical Optimization*. Springer Science & Business Media.

Noor, Ahmed K and Jeanne M Peters (1980). "Reduced basis technique for nonlinear analysis of structures". In: *Aiaa journal* 18.4, pp. 455–462.

O'Leary, Dianne P. and Bert W. Rust (1986). "Confidence Intervals for Inequality-Constrained Least Squares Problems, with Applications to Ill-Posed Problems". In: *SIAM Journal on Scientific Computing* 7.2, pp. 473–489.

Ok, Efe A. (2007). *Real Analysis with Economic Applications*. English. 1st. Econometric Society Monographs. Includes bibliographical references and index. Princeton, NJ: Princeton University Press, p. 832. ISBN: 978-0-691-12949-5. URL: https://press.princeton.edu/books/hardcover/9780691129495/real-analysis-with-economic-applications.

Owen, Art B (2013). "Variance components and generalized Sobol'indices". In: *SIAM/ASA Journal on Uncertainty Quantification* 1.1, pp. 19–41.

Owhadi, H. and C. Scovel (2016). "Brittleness of Bayesian inference and new Selberg formulas". In: *Communications in Mathematical Sciences* 14.1, pp. 83–145.

– (2017a). "Extreme points of a ball about a measure with finite support". In: *Communications in Mathematical Sciences* 15.1. arXiv:1504.06745, pp. 77–96.

– (2017b). "Qualitative robustness in Bayesian inference". In: *ESAIM: Probability and Statistics* 21, pp. 251–274.

– (2017c). "Toward Machine Wald". In: *Handbook of Uncertainty Quantification*. Ed. by Owhadi H. Ghanem R. Higdon D. arXiv:1508.02449. Springer, pp. 157–191.

– (2019). *Operator Adapted Wavelets, Fast Solvers, and Numerical Homogenization, from a game theoretic approach to numerical approximation and algorithm design*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

Owhadi, H., C. Scovel, and F. Schäfer (2019). "Statistical Numerical Approximation". In: *Notices of the AMS* 66.10.

Owhadi, H., C. Scovel, and T. Sullivan (2015a). "Brittleness of Bayesian inference under finite information in a continuous world". In: *Electronic Journal of Statistics* 9.1, pp. 1–79.

– (2015b). "On the brittleness of Bayesian inference". In: *SIAM Review* 57.4, pp. 566–582.

Owhadi, H., C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz (2013a). "Optimal Uncertainty Quantification". In: *SIAM Review* 55.2, pp. 271–345. DOI: 10.1137/10080782X.

– (2013b). "Optimal uncertainty quantification". In: *Siam Review* 55.2, pp. 271–345.

Owhadi, Houman (2015a). "Bayesian numerical homogenization". In: *Multiscale Modeling & Simulation* 13.3, pp. 812–828.

– (2015b). "Bayesian numerical homogenization". In: *Multiscale Modeling & Simulation* 13.3, pp. 812–828.

– (2017). "Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games". In: *Siam Review* 59.1, pp. 99–149.

– (2022). "Computational graph completion". In: *Research in the Mathematical Sciences* 9.2, p. 27.

– (2023a). "Do ideas have shape? Idea registration as the continuous limit of artificial neural networks". In: *Physica D: Nonlinear Phenomena* 444, p. 133592.

– (2023b). "Do ideas have shape? Idea registration as the continuous limit of artificial neural networks". In: *Physica D: Nonlinear Phenomena* 444, p. 133592.

Owhadi, Houman and Clint Scovel (2019a). *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

– (2019b). *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*. Cambridge University Press.

Owhadi, Houman, Clint Scovel, and Tim Sullivan (2015a). "Brittleness of Bayesian inference under finite information in a continuous world". In: *Electronic Journal of Statistics* 9, pp. 1–79.

– (2015b). "On the brittleness of Bayesian inference". In: *SIAM Review* 57.4, pp. 566–582.

Owhadi, Houman, Clint Scovel, and Gene Ryan Yoo (2021). *Kernel Mode Decomposition and the programming of kernels*. Springer.

Owhadi, Houman and Gene Ryan Yoo (2019a). "Kernel flows: from learning kernels from data into the abyss". In: *Journal of Computational Physics* 389, pp. 22–47.

– (2019b). "Kernel flows: from learning kernels from data into the abyss". In: *Journal of Computational Physics* 389, pp. 22–47.

Owhadi, Houman and Lei Zhang (2007). "Metric-based upscaling". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 60.5, pp. 675–723.

– (2017). "Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients". In: *Journal of Computational Physics* 347, pp. 99–128.

Pakman, Ari and Liam Paninski (2014). "Exact Hamiltonian Monte Carlo for Truncated Multivariate Gaussians". In: *Journal of Computational and Graphical Statistics* 23, pp. 518–542.

Palais, Richard S. (1997). *The Symmetries of Solitons*. arXiv: dg-ga/9708004 [dg-ga].

Panaretos, Victor M. (2016). *Statistics for Mathematicians*. Springer. DOI: 10.1007/978-3-319-28341-8. URL: https://doi.org/10.1007/978-3-319-28341-8.

Pass, B. (2020). "Generalized barycenters and variance maximization on metric spaces". In: *arXiv:2006.02984*.

Paszke, A., S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H.

Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-%20style-high-performance-deep-learning-library.pdf.

Patil, Pratik, Mikael Kuusela, and Jonathan Hobbs (2022). "Objective frequentist uncertainty quantification for atmospheric CO_2 retrievals". In: *SIAM/ASA Journal on Uncertainty Quantification* 10. DOI: 10.1137/20M1356403.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pena-Ordieres, Alejandra, James R Luedtke, and Andreas Wachter (2020). "Solving chance-constrained problems via a smooth sample-based nonlinear approximation". In: *SIAM Journal on Optimization* 30.3, pp. 2221–2250.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Pinkus, Allan (2012). *N-widths in Approximation Theory*. Vol. 7. Springer Science & Business Media.

Popoviciu, T. (1935). "Sur les équations algébriques ayant toutes leurs racines réelles". In: *Mathematica (Cluj)*, pp. 129–145.

Prasanth, Sai, Ziad Haddad, Jouni Susiluoto, Amy Braverman, Houman Owhadi, Boumediene Hamzi, Svetla Hristova-Veleva, and Joseph Turk (2021). "Kernel flows to infer the structure of convective storms from satellite passive microwave observations". In: *AGU Fall Meeting Abstracts*. Vol. 2021, A55F–1445.

Raissi, Maziar, Paris Perdikaris, and George E Karniadakis (2019). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378, pp. 686–707.

Raissi, Maziar, Paris Perdikaris, and George Em Karniadakis (2018). "Numerical Gaussian processes for time-dependent and nonlinear partial differential equations". In: *SIAM Journal on Scientific Computing* 40.1, A172–A198.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, pp. I–XVIII, 1–248. ISBN: 026218253X.

Ren, Christopher X, Sidhant Misra, Marc Vuffray, and Andrey Y Lokhov (2021). "Learning Continuous Exponential Families Beyond Gaussian". In: *arXiv preprint arXiv:2102.09198*.

Reznikov, A and Edward B Saff (2016). "The covering radius of randomly distributed points on a manifold". In: *International Mathematics Research Notices* 2016.19, pp. 6065–6094.

Richter, Lorenz, Leon Sallandt, and Nikolas Nüsken (2021). "Solving high-dimensional parabolic PDEs using the tensor train format". In: *arXiv preprint arXiv:2102.11830*.

Robertson, Tim, F. T. Wright, and Richard Dykstra (1988). *Order Restricted Statistical Inference*. Wiley.

Roch, Sebastien (2024). *Modern Discrete Probability: An Essential Toolkit*. To be published by Cambridge University Press. URL: `https://people.math.wisc.edu/~roch/mdp/index.html`.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rockafellar, R. T. and R. J.-B. Wets (1998). *Variational Analysis*. Vol. 317. Grundlehren der Mathematischen Wissenschaften. Berlin: Springer-Verlag.

Rodgers, Clive D. (2000). *Inverse Methods for Atmospheric Sounding*. World Scientific Publishing.

Rojo, J. (2006). "Optimality: The Second Erich L. Lehmann Symposium". In: IMS.

– (2009). "Optimality: The Third Erich L. Lehmann Symposium". In: IMS.

Rojo, J. and V. Pérez-Abreu (2004). "The First Erich L. Lehmann Symposium: Optimality". In: IMS.

Rossi, R. J. (2018). *Mathematical Statistics: an Introduction to Likelihood Based Inference*. John Wiley & Sons.

Rust, Bert W. and Dianne P. O'Leary (1994). "Confidence intervals for discrete approximations to ill-posed problems". In: *Journal of Computational and Graphical Statistics* 3.1, pp. 67–96.

Rust, Burt W. and Walter R. Burrus (1972). *Mathematical Programming and the Numerical Solution of Linear Equations*. American Elsevier.

Sachs, Karen, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan (2005). "Causal protein-signaling networks derived from multiparameter single-cell data". In: *Science* 308.5721, pp. 523–529.

Särkkä, Simo (2011). "Linear operators and stochastic partial differential equations in Gaussian process regression". In: *International Conference on Artificial Neural Networks*. Springer, pp. 151–158.

Schaäfer, Florian, Matthias Katzfuss, and Houman Owhadi (2021). "Sparse Cholesky Factorization by Kullback–Leibler Minimization". In: *SIAM Journal on scientific computing* 43.3, A2019–A2046.

Schaback, Robert (2007). "Convergence of unsymmetric kernel-based meshless collocation methods". In: *SIAM Journal on Numerical Analysis* 45.1, pp. 333–351.

Schaback, Robert (2015). "A computational tool for comparing all linear PDE solvers: Error-optimal methods are meshless". In: *Advances in Computational Mathematics* 41, pp. 333–355.

– (2016). "All well-posed problems have uniformly stable and convergent discretizations". In: *Numerische Mathematik* 132.3, pp. 597–630.

Schaback, Robert and Holger Wendland (2006). "Kernel techniques: from machine learning to meshless methods". In: *Acta numerica* 15, p. 543.

Schafer, Chad M. and Philip B. Stark (2009). "Constructing confidence regions of optimal expected size". In: *Journal of the American Statistical Association* 104.487, pp. 1080–1089.

Schäfer, Florian, Matthias Katzfuss, and Houman Owhadi (2021). "Sparse Cholesky Factorization by Kullback–Leibler Minimization". In: *SIAM Journal on Scientific Computing* 43.3, A2019–A2046.

Schäfer, Florian and Houman Owhadi (2021). "Sparse recovery of elliptic solvers from matrix-vector products". In: *arXiv preprint arXiv:2110.05351*.

– (2023). "Sparse recovery of elliptic solvers from matrix-vector products". In: *SIAM Journal on Scientific Computing*.

Schäfer, Florian, Timothy John Sullivan, and Houman Owhadi (2021a). "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity". In: *Multiscale Modeling & Simulation* 19.2, pp. 688–730.

– (2021b). "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity". In: *Multiscale Modeling & Simulation* 19.2, pp. 688–730.

– (2021c). "Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity". In: *Multiscale Modeling & Simulation* 19.2, pp. 688–730.

Schervish, Mark J. (1995). *Theory of Statistics*. Springer.

Schmitt, S. (2012). "TUnfold, an algorithm for correcting migration effects in high energy physics". In: *Journal of Instrumentation* 7, T10003.

Scholkopf, Bernhard and Alexander J Smola (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press.

Schölkopf, Bernhard, Ralf Herbrich, and Alex J. Smola (2001). "A Generalized Representer Theorem". In: *Computational Learning Theory*. Ed. by David Helmbold and Bob Williamson. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 416–426. ISBN: 978-3-540-44581-4.

Schott, J. R. (2016). *Matrix Analysis for Statistics*. John Wiley & Sons.

Schweiger, Regev, Eyal Fisher, Elior Rahmani, Liat Shenhav, Saharon Rosset, and Eran Halperin (2018). "Using Stochastic Approximation Techniques to Efficiently Construct Confidence Intervals for Heritability". In: *Journal of Computational Biology* 25.7, pp. 794–808.

Schweitzer, Frank, Giorgio Fagiolo, Didier Sornette, Fernando Vega-Redondo, Alessandro Vespignani, and Douglas R White (2009). "Economic networks: The new challenges". In: *science* 325.5939, pp. 422–425.

SHAH, RAJEN D and JONAS PETERS (2020). "THE HARDNESS OF CONDITIONAL INDEPENDENCE TESTING AND THE GENERALISED COVARIANCE MEASURE". In: *The Annals of Statistics* 48.3, pp. 1514–1538.

Shaked, M. and J.G. Shanthikumar (2007). *Stochastic Orders*. Springer Series in Statistics. Springer New York. ISBN: 9780387346755.

Shapiro, A. and A. Kleywegt (2002). "Minimax analysis of stochastic problems". In: *Optimization Methods and Software* 17.3, pp. 523–542.

Shapiro, Alexander (1988). "Towards a unified theory of inequality constrained testing in multivariate analysis". In: *International Statistical Review*, pp. 49–62.

Shin, Yeonjong, Jerome Darbon, and George Em Karniadakis (2020). "On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs". In: *arXiv preprint arXiv:2004.01806*.

Silvapulle, Mervyn J. and Pranab Kumar Sen (2011). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley.

Sion, M. (1958). "On general minimax theorems". In: *Pacific J. Math* 8.1, pp. 171–176.

Skilling, J. (2004). "Nested Sampling". In: *AIP Conference Proceedings* 735, p. 295.

Smith, Robert L. (1984). "Efficient Monte Carlo Procedures for Generating Points Uniformly Distributed over Bounded Regions". In: *Operations Research* 32, pp. 1296–1308.

Sobol, Ilya M (2001). "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates". In: *Mathematics and computers in simulation* 55.1-3, pp. 271–280.

Soboĺ, IM (1993). "Sensitivity estimates for nonlinear mathematical models". In: *Math. Model. Comput. Exp.* 1, p. 407.

Spano, Francesco (2014). "Unfolding in particle physics: a window on solving inverse problems." en. In: *Proc. of Top 2013*, 256–261, DESY. DOI: 10.3204/DESY-PROC-2014-02/52. URL: http://www-library.desy.de/preparch/desy/proc/proc14-02/P52.pdf.

Spirtes, Peter and Clark Glymour (1991). "An algorithm for fast recovery of sparse causal graphs". In: *Social science computer review* 9.1, pp. 62–72.

Sprott, D. A. (2008). *Statistical Inference in Science*. Springer Verlag.

Stanley, Michael, Pau Batlle, Pratik Patil, Houman Owhadi, and Mikael Kuusela (2025). "Confidence intervals for functionals in constrained inverse problems via data-adaptive sampling-based calibration". In: *arXiv preprint arXiv:2502.02674*. P.B contributed to conceptualization, proof of the theoretical results, and numerical experiments.

Stanley, Michael, Pratik Patil, and Mikael Kuusela (2022). "Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals". In: *Journal of Instrumentation* 17. DOI: 10.1088/1748-0221/17/10/P10013.

Stark, P. (2020). "Your Prior Can Bite You on the Posterior: Contrasting Bayesian and Frequentist Measures of Uncertainty". In: *JPL Science Visitor and Colloquium Program - Earth Science Seminar, Sept. 1, 2020*. https://www.stat.berkeley.edu/~stark/Seminars/uqJPL20.slides.html/.

Stark, Philip B. (1992a). "Inference in Infinite-Dimensional Inverse Problems: Discretization and Duality". In: *Journal of Geophysical Research* 97.B10, pp. 14055–14082.

– (1992b). "Inference in infinite-dimensional inverse problems: discretization and duality". In: *Journal of Geophysical Research: Solid Earth* 97.B10, pp. 14055–14082.

– (1994). *Simultaneous confidence intervals for linear estimates of linear functionals*. Tech. rep. Citeseer.

– (2015). "Constraints versus Priors". In: *SIAM/ASA Journal on Uncertainty Quantification* 3 (1).

Stegle, Oliver, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf (2010). "Probabilistic latent variable models for distinguishing between cause and effect". In: *Advances in neural information processing systems* 23.

Steinwart, Ingo and Andreas Christmann (2011). "Estimating conditional quantiles with the help of the pinball loss". In: *Bernoulli* 117.1, pp. 211–225.

Storn, R. and K. Price (1997). "Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces". In: *Journal of Global Optimization* 11.4, pp. 341–359.

Stuart, Andrew M. (2010). "Inverse Problems: A Bayesian Perspective". In: *Acta Numerica* 19, pp. 451–559.

Sudret, Bruno, Stefano Marelli, and Joe Wiart (2017). "Surrogate models for uncertainty quantification: An overview". In: *2017 11th European conference on antennas and propagation (EUCAP)*. IEEE, pp. 793–797.

Susiluoto, Jouni, Amy Braverman, Philip Brodrick, Boumediene Hamzi, Maggie Johnson, Otto Lamminpaa, Houman Owhadi, Clint Scovel, Joaquim Teixeira, and Michael Turmon (2021). "Radiative transfer emulation for hyperspectral imaging retrievals with advanced kernel flows-based gaussian process emulation". In: *AGU Fall Meeting Abstracts*. Vol. 2021, NG25A–0506.

Swiler, Laura P, Mamikon Gulian, Ari L Frankel, Cosmin Safta, and John D Jakeman (2020). "A Survey of Constrained Gaussian Process Regression: Approaches and Implementation Challenges". In: *Journal of Machine Learning for Modeling and Computing* 1.2.

Sylvester, J. J. (1857). "A question in the geometry of situation". In: *Quarterly Journal of Pure and Applied Mathematics* 1.1, pp. 79–80.

Takeuchi, Ichiro, Quoc V. Le, Timothy D. Sears, and Alexander J. Smola (2006). "Nonparametric Quantile Estimation". In: *Journal of Machine Learning* 7, pp. 1231–1264.

Tarantola, Albert (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Taylor, Michael (2013). *Partial differential equations I: Basic theory*. Springer Science & Business Media.

Tenorio, L., A. Fleck, and K. Moses (2007). "Confidence intervals for linear discrete inverse problems with a non-negativity constraint". In: *Inverse Problems* 23.2, p. 669.

Thomas, Owen, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann (2022). "Likelihood-Free Inference by Ratio Estimation". In: *Bayesian Analysis* 17.1, pp. 1–31.

Vaart, Aad W van der and J Harry van Zanten (2008). "Reproducing Kernel Hilbert Spaces of Gaussian priors". In: *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*. Institute of Mathematical Statistics, pp. 200–222.

Vecchia, Aldo V (1988). "Estimation and model identification for continuous spatial processes". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 50.2, pp. 297–312.

Venzon, D. J. and S. H. Moolgavkar (1988). "A method for computing profile-likelihood-based confidence intervals". In: *Journal of the Royal Statistical Society: Series C* 37.1, pp. 87–94.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors (2020).

"SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Voelker, Aaron R., Jan Gosmann, and Terrence C. Stewart (2017). *Simultaneous confidence intervals for linear estimates of linear functionals*. Tech. rep. Centre for Theoretical Neuroscience.

Wahba, Grace (2003). "An introduction to smoothing spline anova models in rkhs, with examples in geographical data, medicine, atmospheric sciences and machine learning". In: *IFAC Proceedings Volumes* 36.16, pp. 531–536.

Wald, A. (1945). "Statistical decision functions which minimize the maximum risk". In: *Annals of Mathematics*, pp. 265–280.

Wald, A. and J. Wolfowitz (1951). "Characterization of the Minimal Complete Class of Decision Functions When the Number of Distributions and Decisions Is Finite". In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, Calif.: University of California Press, pp. 149–157.

Wald, Abraham (1943). "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large". In: *Transactions of the American Mathematical Society* 54, pp. 426–482.

Wang, Junyang, Jon Cockayne, Oksana Chkrebtii, Timothy John Sullivan, Chris Oates, et al. (2021). "Bayesian numerical methods for nonlinear partial differential equations". In: *Statistics and Computing* 31.5, pp. 1–20.

Wang, Sifan, Hanwen Wang, and Paris Perdikaris (2021). "Learning the solution operator of parametric partial differential equations with physics-informed deeponets". In: *Science advances* 7.40, eabi8605.

– (2022). "Improved architectures and training algorithms for deep operator networks". In: *Journal of Scientific Computing* 92.2, p. 35.

Wasserman, Larry (2004). *All of Statistics: A Concise Course in Statistical Inference*. Vol. 26. Springer.

Wasserman, Larry, Aaditya Ramdas, and Sivaraman Balakrishnan (2020). "Universal Inference". In: *PNAS* 117, pp. 16880–16890.

Weinan, E (2011). *Principles of multiscale modeling*. Cambridge University Press.

Weinan, E, Jiequn Han, and Arnulf Jentzen (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.

Weinan, E and Bing Yu (2018). "The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems". In: *Communications in Mathematics and Statistics* 6.1, pp. 1–12.

Weizsäcker, H. von and G. Winkler (1979). "Integral representation in the set of solutions of a generalized moment problem". In: *Mathematische Annalen* 246.1, pp. 23–32.

Welzl, E. (1991). "Smallest enclosing disks (balls and ellipsoids)". In: *New Results and New Trends in Computer Science*. Springer, pp. 359–370.

Wendland, Holger (2004). *Scattered Data Approximation*. Cambridge University Press.

Williams, Christopher K. I. and Carl Edward Rasmussen (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Wilson, Andrew Gordon, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing (2016). "Deep kernel learning". In: *Artificial Intelligence and Statistics*. PMLR, pp. 370–378.

Wolak, Frank A. (1987). "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model". In: *Journal of the American Statistical Association* 82, pp. 782–793.

– (1989). "Testing Inequality Constraints in Linear Econometric Models". In: *Journal of Econometrics* 41, pp. 205–235.

Wu, Zong-min and Robert Schaback (1993). "Local error estimates for radial basis function interpolation of scattered data". In: *IMA journal of Numerical Analysis* 13.1, pp. 13–27.

Xiu, Dongbin (2010). *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press.

Xiu, Dongbin and George Em Karniadakis (2002). "The Wiener–Askey polynomial chaos for stochastic differential equations". In: *SIAM journal on scientific computing* 24.2, pp. 619–644.

Xiu, Dongbin and Jie Shen (2009). "Efficient stochastic Galerkin methods for random diffusion equations". In: *Journal of Computational Physics* 228.2, pp. 266–281.

Ye, Qi (2013). "Kernel-based methods for stochastic partial differential equations". In: *arXiv preprint arXiv:1303.5381*.

Yildirim, E. A. (2008). "Two algorithms for the minimum enclosing ball problem". In: *SIAM Journal on Optimization* 19.3, pp. 1368–1391.

Yu, Ming, Varun Gupta, and Mladen Kolar (2019). "Constrained high dimensional statistical inference". In: *arXiv preprint arXiv:1911.07319*.

Zhang, K, J Peters, D Janzing, and B Schölkopf (2011). "Kernel-based Conditional Independence Test and Application in Causal Discovery". In: *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. AUAI Press, pp. 804–813.

Zhu, Yinhao and Nicholas Zabaras (2018). “Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification”. In: *Journal of Computational Physics* 366, pp. 415–447.