

©Copyright 2026
Alexander Win Hsu

Generative and variational approaches to scientific machine learning,
inverse problems and PDEs

Alexander Win Hsu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Bamdad Hosseini, Chair

Aleksandr Aravkin, Chair

Tom Trogdon

Dan Mikulincer (Graduate School Representative)

Program Authorized to Offer Degree:
Department of Applied Mathematics

University of Washington

Abstract

Generative and variational approaches to scientific machine learning, inverse problems and PDEs

Alexander Win Hsu

Co-Chairs of the Supervisory Committee:

Assistant Professor Bamdad Hosseini
Applied Mathematics

Associate Professor Aleksandr Aravkin
Applied Mathematics

This dissertation collects four pieces of research conducted during my doctoral studies. The chapters are largely independent, except that Chapter 3 builds directly on the framework developed in Chapter 2.

Chapter 1 develops a theory of conditional optimal transport in separable infinite-dimensional function spaces, motivated by simulation-based Bayesian inference. We study the solvability of block-triangular Monge maps and their Kantorovich relaxations, derive regularity estimates for the posterior-conditioning maps arising in Bayesian inverse problems, and demonstrate the framework on a Darcy-flow benchmark.

Chapter 2 develops a kernel-based framework for learning differential equations from data, together with quantitative worst-case error bounds. The state and the unknown governing operator are both represented in reproducing kernel Hilbert spaces, with the differential equation enforced via least-squares collocation; we also discuss connections of this framework to operator learning and PDE solvers.

Chapter 3 builds on the kernel collocation framework of Chapter 2 to develop an all-at-once method for jointly identifying sparse ordinary differential equations and reconstructing the underlying state from scarce, partial and noisy observations. The unknown dynamics are parametrized as a sparse linear combination over a fixed function library, and the resulting nonconvex problem is solved by alternating Levenberg–Marquardt steps on a fixed-support nonlinear least squares problem with an outer sparse-regression update.

Chapter 4 determines the sharp value of the hot spots ratio in every dimension and characterizes its asymptotic behavior. The upper bound combines a comparison principle, Talenti’s rearrangement inequality, and a quantitative form of the Szegő–Weinberger eigenvalue inequality; the lower bound follows from a Neumann-sieve construction of extremizing sequences.

TABLE OF CONTENTS

	Page
List of Figures	iii
Introduction	1
Chapter 1: Conditional Optimal Transport on Function Spaces	2
Preface	2
1.1 Introduction and setup	3
1.2 Preliminaries	11
1.3 General theory of conditional OT	14
1.4 Application to BIPs	24
1.5 Numerical Experiments	27
1.6 Conclusions	34
Chapter 2: Data-Efficient Kernel Methods for Learning Differential Equations and Their Solution Operators	35
Preface	35
2.1 Introduction	36
2.2 Kernel equation learning	40
2.3 Implementation and algorithms	46
2.4 Experiments	49
2.5 Theoretical analysis	55
2.6 Discussion	64
2.7 Details of theoretical analysis	65
2.8 Details of algorithms	71
2.9 Details of numerical experiments	77
Chapter 3: A Joint Optimization Approach to Identifying Sparse Dynamics Using Least Squares Kernel Collocation	87
Preface	87
3.1 Introduction	87
3.2 Joint state estimation and system identification	91

3.3	Algorithms and implementation details	97
3.4	Numerical results	100
3.5	Discussion and conclusion	118
3.6	Theory	121
3.7	Computational tools	122
Chapter 4: Sharp Bounds on the Failure of the Hot Spots Conjecture		123
	Preface	123
4.1	Introduction	124
4.2	The upper bound	128
4.3	The lower bound construction	131
4.4	The effective problem	133
4.5	The Neumann sieve problem	139
4.6	Asymptotics as $d \rightarrow \infty$	144
Bibliography		145

LIST OF FIGURES

Figure Number		Page
1.1	A depiction of conditional sampling using triangular maps obtained from Algorithm 1. The left panel shows a scatter plot of samples from ν that were used to solve the conditional OT problem while the blue, red, and black lines denote slices along which conditional samples are generated. The other panels show the conditional histograms with dotted lines showing the true histograms and solid lines showing the numerical approximations.	6
1.2	Conditional OT experiments on two dimensional benchmarks using Algorithm 1: (Left column) the empirical samples from ν . The blue and black lines indicate slices along which we condition the measures. (Middle column) comparison between the ground truth histogram indicated with dotted lines and the conditional OT histogram indicated with solid lines. (Right column) The conditioning maps $\hat{T}_{\mathcal{U}}(y, \cdot)$ for each slice computed using two nearest neighbor interpolation. The ground truth conditional densities were computed by performing rejection sampling to generate 50,000 samples in the intervals $(y - 0.01, y + 0.01)$, and treating these as conditional samples from $\nu(\cdot y)$. Kernel density estimation was then used on these samples to plot the conditional distributions.	29
1.3	Summary of our experiments for Darcy flow inverse problem comparing pCN and WaMGAN. From left, the first column shows the ground truth field $u^\dagger$, second column shows the corresponding $p(u^\dagger)$ and measurement locations x_j denoted as black dots. Third and fourth columns show posterior means for WaMGAN and pCN respectively while the fifth and sixth columns show pointwise variances of the posteriors in the same order. In all cases we used the same color range for pCN and WaMGAN to ensure proper comparison between the two methods.	32
1.4	Scatterplot comparison of pointwise posterior variances calculated by WaMGAN and pCN for the four examples in Figure 1.3. The diagonal line represents perfect agreement between the two methods.	32
1.5	Numerical convergence plot of transport maps for the Darcy flow example. We compared models trained on an increasing number of KL modes ($\hat{T}^N$) to a high resolution map trained with 64 modes ($\hat{T}^{64}$), which was taken as the numerical ground truth. We measured the error of the maps with respect to the $L^2(\eta)$ norm which was computed using an independent set of reference samples.	33

1.6	A comparison of posterior samples from pCN (top row) vs WaMGAN (bottom row) for one of our experiments.	33
2.1	(a) Schematic depiction of our equation learning problem. Red elements are the unknowns of interest, blue objects are the given data, while black objects are assumed to be known. The left and right panels show the solution and right-hand side of an example second order PDE depending on $y, u, \partial_y u$, and $\partial_y^2 u$ while the middle panel shows $\Phi(y, u)$; (b) The computational graph of 2-step KEqL. Red edges are unknown nonlinearities of interest, blue boxes denote data that is known for various nodes with dashed lines denoting where the data is injected, black edges are known. Note that the graphs for u^m and P are disconnected, hence the learning of u^m and P is performed sequentially in two steps; (c) The computational graph for 1-step KEqL. Coloring conventions follow panel (b) with the main difference being that the u^m and P are now connected and have to be learned simultaneously.	42
2.2	Representative numerical results for the Duffing oscillator (2.18): (a) Shows the training data and the ground truth state of the system u in comparison to the reconstructed state $\hat{u}$ using 1-step KEqL and 2-step methods; (b) Quantitative values of relative state reconstruction and operator learning errors, the latter is reported for different time windows. These values were averaged over three random initial conditions; (c) Visualization of three extrapolated dynamics used to compute $\mathcal{R}_{\text{opl}}$	52
2.3	Representative numerical results for Burgers' PDE (2.19): (a) State reconstruction and equation learning errors computed on the training functions for 1-step KEqL, kernel interpolant + SINDy, and PINN-SR using $M = 1$ training pairs and varying the number of interior observations N_y ; (b) Similar experiment as panel (a) but with randomized initial conditions; (c) State reconstruction results from scarce data using parameters that lead to a near shock comparing 1-step KEqL, PINN-SR and a kernel interpolation; (d) Similar setup to row (c) with a smooth solution and observations only on the boundary; (e) Computed solutions to the PDEs identified by 1-step KEqL, PINN-SR, and SINDy from the data in row (d) for a new initial condition.	54
2.4	Representative numerical results for Darcy's flow PDE (2.20): The top two figures from the left show the equation learning errors computed over training and ID test sets. The OOD test errors are shown in the next row along with the ID operator learning error. The 1-step KEqL was used and the labels on the graphs denote the number of interior observation points N_y	55
2.5	Convergence history of LM for 1-step KEqL for the Duffing ODE (2.18).	78
2.6	Initial conditions sampled from the process (2.45) for examples of Burgers' PDE (2.19) increasing number of observations: (A) Condition used in the fixed IC case; (B-D) Sample conditions for the varying IC case.	79

2.7	Samples of training solutions along with their N observation points (red) and the collocation points (white) for Burgers' PDE (2.19) with ICs sampled from the GP (2.45): (A) Solution obtained using a fixed IC depicted in Figure 2.6(A); (B) Three solutions obtained from solving the PDE with different ICs depicted in Figure 2.6(B-D).	79
2.8	Solution of the Burgers' PDE (2.19) with initial condition (2.47). Sampled observation points are shown in red, while collocation points are overlaid in white.	80
2.9	Representative numerical results for an ablation study on Burgers' PDE (2.19) with initial condition (2.47), examining the effect of removing the diffusion feature across the 1-step KEqL, PINN-SR, and kernel interpolation methods: (A) solutions recovered using the reduced feature set $\{1, u, \partial_x u\}$; (B) solutions recovered using the expanded feature set $\{1, u, \partial_x u, \partial_x^2 u, \partial_{tx} u, \partial_t^2 u\}$	82
2.10	Behavior of solutions at initial and final time for one-shot learning examples of Burgers' PDE (2.19): (A) Shock development (Section 2.9.2); (B) Smooth transition (Section 2.9.2); (C) Solution behavior to the PDE in (B) for the depicted new initial condition.	82
2.11	Quantitative values of relative filtering and operator learning errors for Burgers' PDE Equation (2.19) with different ICs corresponding to the one-shot learning examples Ex. 1 (Section 2.9.2), its ablation studies Ex. 1a-b, in which features are removed and added (Section 2.9.2), and Ex. 2 (Section 2.9.2) in Figure 2.3(C-E).	83
2.12	Convergence history of the LM algorithm for 1-step KEqL for Burgers' PDE (2.19): (A-B) Sample cases using fixed and varying IC, using $M = 1$ training pairs with number of observations $N_y = 20$ corresponding to the two cases depicted in the first column in Figure 2.7; (C) Shock development (Section 2.9.2); (D) One-shot smooth case where only observed on the boundary (Section 2.9.2).	83
2.13	Samples of training tuples (u^m, f^m) with different number of observation points for Darcy's flow PDE (2.20): (A-C) Training tuples observation points (red) and collocation points (white).	84
2.14	Samples of test solutions u (ID and OOD) and their contours of the equation and operator learning errors for 1-step and 2-step KEqL methods for Darcy's flow PDE (2.20): (A) Sample solutions u ID and OOD for the PDE ;(B) Contour plots for equation learning errors; (C) Contour plots for operator learning errors.	86
2.15	Convergence history of the block arrowhead matrix structure LM-type algorithm for the reduced 1-step KEqL when using $M = 8$ solution tuples at $N_y = 8$ observation points in the interior for Darcy's flow PDE (2.20).	86

3.1	Results for Lorenz 63 system. Observations sampled at $\Delta t = 0.05$ until $t = 10$ with added Gaussian noise of variance $\sigma^2 = 4$. Simulated dynamics from eq. (3.43) are compared against true dynamics from $t = 10$ until $t = 15$.	103
3.2	Lorenz 63 system with observations sampled at $\Delta t = 0.5$ until $t = 10$ with no added noise. Simulated dynamics of (3.44) are plotted against true, unseen, trajectory from $t = 10$ to $t = 25$.	104
3.3	Results for the Lotka–Volterra system of section 3.4.1. Samples are taken at the rate $\Delta t = 3$ up to $t = 48$, including additive noise with $\sigma = 0.2$. Simulated dynamics are shown up to $t = 100$, comparing the results from (3.46) with the true dynamics.	105
3.4	Left: True Lotka–Volterra plotted from multiple initial conditions on phase portrait using (3.45). Red trajectory and measurements are those from Figure 3.3. Right: Using learned model (3.46) to simulate same trajectories provided same initial conditions.	106
3.5	Results for Lorenz 63 system with streams of partial observations. Observations alternate between each coordinate every 10 samples at sampling rate $\Delta t = 0.025$ up to $t = 10$ with added noise set to $\sigma^2 = 0.1$. Learned dynamics are simulated from $t = 10$ until $t = 20$.	107
3.6	Results for Lorenz 63 system with an unobserved coordinate. Observations alternate between first two coordinates at sampling rate $\Delta t = 0.025$ up to $t = 10$, with no knowledge of the third, except that the system is known to be three dimensional. Our algorithm constructs a third, latent coordinate and corresponding dynamics which, when evolved together with the first two coordinates, reproduces the observed dynamics. Further, we can effectively extrapolate the time series forward by using our learned dynamics and imputing the value of our artificial coordinate as a new initial condition, generating good predictions for the first two coordinates.	107
3.7	Results for the Van der Pol oscillator. Sampling is done only on system state, $\mathbf{x}$, with sampling rate $\Delta t = 2$ up to $t = 100$ with simulated dynamics from $t = 100$ to $t = 200$ using the learned equations from eq. (3.50).	109
3.8	Left: True Van der Pol oscillator trajectories plotted from multiple initial conditions on phase portrait using eq. (3.49). Right: Using learned model eq. (3.50) to simulate same trajectories provided same initial conditions.	110
3.9	Position observations sampled at rate $\Delta t = 0.2$ up to $t = 30$ with added noise set to $\sigma^2 = 0.04$ equating to a relative noise of 2.55%. Learned and true dynamics from eq. (3.55) are simulated from $t = 30$ to $t = 50$.	111

3.10	Results for estimating time derivatives of the coupled Duffing system. The relative mean square errors of the estimates of the first, second, third, and fourth derivatives are $2.32 \cdot 10^{-3}$, $3.68 \cdot 10^{-3}$, $6.57 \cdot 10^{-3}$ and $1.1 \cdot 10^{-2}$ respectively. In this case, our approach resolves the fine structure in estimating the state.	113
3.11	Results for misspecified linear and cubic models for nonlinear pendulum. Observations were sampled at a rate $\Delta t = 1$ up to $t = 100$ with added noise with variance $\sigma^2 = 0.01$. Learned dynamics for both linear and cubic model, eq. (3.57) and eq. (3.58), are simulated from $t = 100$ to $t = 200$ where the linear model successfully captures periodicity of the true trajectory, but fails to capture the correct amplitude, whereas this is flipped for the cubic model.	114
3.12	Results for the nonlinear damped oscillator with full state observations, $\Delta t = 0.05$ in the interval $[0, 10]$, and $\sigma = 0.32$ additive noise. While the learned equations in eq. (3.60) do not match the true dynamics in terms of the coefficient values, JSINDy effectively infers the states, and produces a reasonable extrapolation in forward simulation.	116
3.13	Results for the damped nonlinear oscillator across varying levels of noise. Left: coefficient error $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$, Right: state estimation error $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x})$. In this case, ensembling significantly improves the coefficient error for both weak SINDy and JSINDy. For JSINDy, including ensembling inside of the sparsifier brings the coefficient error from noticeably worse than ODR-BINDy to slightly better. Overall, JSINDy performs very well in state estimation across varying levels of noise.	116
3.14	Experiment across different trajectory lengths and noise levels, averaged over 128 independent trials and $\Delta t = 0.01$. Top: Proportion of trials where the correct support was exactly recovered Bottom: Relative coefficient error $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ of identified system.	117
4.1	(Left) Plot of S_d as a function of d with the asymptotic value $S_d \rightarrow \sqrt{e}$ marked by a red line. (Right) Approximate values of S_d for different dimensions d along with the previously best known bounds. The value > 1 means that no specific value had been computed (to the knowledge of the authors), but the value is known to be > 1 . Results marked with * were <i>folklore</i> results. The fact that $\liminf_{d \rightarrow \infty} S_d > 1$, without a specific computed value, follows from the proof of [348]. The lower bound in $d = 2$ comes from numerical experiments [360]. All previously best known upper bounds follow from [342].	125
4.2	(Left) Sketch of the Neumann sieve domains. (Right) First Neumann Laplace eigenfunction for a sequence of sieve domains. In the limit when the number of holes goes to infinity one recovers a radially symmetric <i>effective problem</i> . The effective problem is studied in Section 4.4, and the fact that the convergence holds is studied in Section 4.5.	131

- 4.3 (Left) Hot spots ratio in dimension 3 as a function of β and δ . Note that the actual value of S_3 is approximately 2.3861. The black curve represents the boundary of the allowed region, above which the first non-trivial Neumann eigenfunction is not radial and the effective problem yields a hot spots constant equal to 1. The optimal asymptotic domain is achieved by taking $\delta \rightarrow 0$ and $\beta \rightarrow \mu_{B_1(\mathbb{R}^3)}$. (Right) Examples of the radial part of the eigenfunctions $\frac{\psi_{\beta,\delta}^{(1)}(r)}{\psi_{\beta,\delta}^{(1)}(1^-)}$, where β is taken as large as possible while keeping the first eigenfunction radial. This normalization sets the maximum on the boundary to 1, in contrast to the L^2 normalization. As we take $\delta \rightarrow 0$ and $\beta \rightarrow \mu_{B_1(\mathbb{R}^3)}$, $\frac{\psi_{\beta,\delta}^{(1)}(0)}{\psi_{\beta,\delta}^{(1)}(1^-)}$ will approach S_3 while $\frac{\psi_{\beta,\delta}^{(1)}(1^+)}{\psi_{\beta,\delta}^{(1)}(1^-)}$ will tend to $-\infty$ at a rate $\sim -\delta^{-1}$. . 134
- 4.4 The domain Ω_ϵ consists of $B_{1+\delta} := B_{1+\delta}(\mathbb{R}^d)$ with the region $\{x \in B_{1+\delta} \text{ s.t. } |x| \in (1 - \epsilon/2, 1 + \epsilon/2)\}$ removed along with thin channels through that region. We refer to the bulk as W_ϵ and the channels as N_ϵ . The width of the channels must be much smaller than their length, ϵ , being that they contain only a proportion $\epsilon^2 \alpha$ of the volume in the shell while being equidistributed on $\mathbb{S}^{d-1}$ at scales $\sim \exp(-\frac{1}{\epsilon})$. In the case $d = 2$, this will consist of exponentially many channels that are getting exponentially narrower as $\epsilon \rightarrow 0$ 139

ACKNOWLEDGMENTS

Completing this PhD would not have been possible without the support, guidance, and love of everyone along the way. Naturally, many people will fall into multiple categories in these acknowledgements—I will mostly name each person only once besides the list of collaborators. First, I’d like to thank my wonderful advisors Bamdad Hosseini and Sasha Aravkin. Their wisdom, patience, and generosity underlie all of my accomplishments during this PhD. I’m grateful to the others who have served on my committee: Tom Trogdon, Dan Mikulincer, and Stefan Steinerberger.

I’d like to thank my family: Mom, Dad, Sophie, Lauren, Poh Poh, and Nai Nai. Their love and support have been constant and indispensable throughout this journey.

Everyone who has taught me math has shaped the mathematician I am today: Xiao Wu, Sarah Marzen, Lenny Fukshansky, Marina Chugunova, Asuman Aksoy, Mike O’Neill, Anne Greenbaum, Bernard Deconinck, and Heather Wilber. I owe special thanks to Tom for welcoming me into his applied analysis reading group and to Stefan for sharing so much wisdom in many lectures and conversations. My time in grad school has been enriched by my many friends, officemates, and colleagues in the applied math department: Ike, Ellie, Cade, Charbel, James, Johnson, Juan, Sara, Wietse, Kaitlynn, Rohin, Lemon, Saba, Jenny, Payton, Maddie, Becks, Laurel, Ziqi, Mohammed, Mike, Emily, Biraj, Matt, Damien, and Cooper.

I’ve had the privilege of working with many fantastic collaborators during this time: Charbel Abi Younes, Aleksandr Aravkin, Aras Bacho, Ricardo Baptista, Théo Bourdais, Edoardo Calvello, Matthieu Darcy, Jaume de Dios Pont, Ariane Ducellier, Laura Dwyer-Lindgren, Ethan Epperly, Lenny Fukshansky, Ike Griss Salas, Bill Gustafson, Bamdad Hosseini, Yasamin Jalalian, Alexander Johnson-Vázquez, Parkes Kendrick, J. Nathan Kutz, Sarah Marzen, Christopher Murray, Juan Felipe Osorio Ramirez, Houman Owadi, Aleksei Sorokin, Jacob Stevens-Haas, Amirhossein Taghvaei, Mitchell Taylor, Xianjin Yang, and Peng Zheng.

I gratefully acknowledge support from National Science Foundation awards DMS-2208535, DMS-2337678, DMS-2133244, and DMS-2230648, and a Carl E. Pearson Fellowship.

Finally, I must acknowledge the very hospitable working conditions in the office owing to the climate control furnished by Eric Shea-Brown, without which my productivity would have been severely diminished over the summer months.

INTRODUCTION

This dissertation collects four pieces of research carried out during my PhD. The chapters are largely independent and can be read in any order.

Chapter 1 develops a theory of conditional optimal transport in separable infinite-dimensional function spaces, motivated by simulation-based Bayesian inference. The chapter studies the solvability of block-triangular Monge maps and their Kantorovich relaxations under general cost functions, specializes the resulting theory to Bayesian inverse problems with regularity estimates on the posterior-conditioning map, and demonstrates the framework on a Darcy-flow benchmark. This work was joint with Bamdad Hosseini and Amirhossein Taghvaei [1].

Chapter 2 develops a kernel-based framework for learning differential equations from data, together with quantitative worst-case error bounds. The state and the unknown governing operator are both represented in reproducing kernel Hilbert spaces, with the differential equation enforced via least-squares collocation; the chapter also discusses the connections of this framework to operator learning and PDE solvers. This work was joint with Yasamin Jalalian, Juan Felipe Osorio Ramirez, Bamdad Hosseini and Houman Owhadi [2].

Chapter 3 builds on the kernel collocation framework of Chapter 2 to develop an all-at-once method for jointly identifying sparse ordinary differential equations and reconstructing the underlying state from scarce, partial and noisy observations. The unknown right-hand side is parametrized as a sparse linear combination over a fixed function library, and the resulting nonconvex problem is solved by alternating Levenberg–Marquardt steps on a fixed-support nonlinear least squares problem with an outer sparse-regression update. This work was joint with Ike Griss Salas, Jacob M. Stevens-Haas, J. Nathan Kutz, Aleksandr Aravkin and Bamdad Hosseini [3].

Chapter 4 concerns Rauch’s hot spots conjecture, which posits that the extrema of the first nontrivial Neumann eigenfunction of the Laplacian on a Lipschitz domain are attained on the boundary. The conjecture is false in general, and the degree of its failure is measured by the hot spots ratio: the maximum of the eigenfunction over the domain divided by its maximum on the boundary. The chapter determines the sharp value of this ratio in every dimension and characterizes its asymptotic behavior as the dimension grows. This work was joint with Jaume de Dios Pont and Mitchell A. Taylor [4].

Chapter 1

CONDITIONAL OPTIMAL TRANSPORT ON FUNCTION SPACES

Preface

In this chapter, from [1] which was joint work with Bamdad Hosseini and Amirhossein Taghvaei, we develop the theory of conditional optimal transport in the context of simulation based, or amortized Bayesian inference. We preface this article with a brief primer on Bayesian inference from the point of view of generative modeling. In its most abstract form, Bayesian inference amounts to positing a joint distribution $\nu(\mathrm{d}u, \mathrm{d}y)$ on $\mathcal{U} \times \mathcal{Y}$ between parameters of interest u , and data y , and performing inference by sampling or computing statistics of the *posterior* distribution, defined as the conditional $\nu(\mathrm{d}u \mid y)$ after observing data y .

Traditionally, this joint measure is specified through its disintegration into a prior μ and a likelihood kernel $\nu(\mathrm{d}y \mid u)$; when these admit densities, Bayes' theorem yields the familiar representation $p(u \mid y) \propto p(u)p(y \mid u)$ of the posterior. Inference is then performed using values and gradients of this unnormalized density, for instance computing a maximum a posteriori estimate $\arg \max_u \log p(y \mid u) + \log p(u)$, or running a Markov chain Monte-Carlo algorithm to draw samples.

The simulation based approach to Bayesian inference eschews the use of such a density entirely. Instead of assuming access to any (unnormalized) density function, it requires only the ability to draw joint samples $\{(u_i, y_i)\}_{i=1}^N \stackrel{iid}{\sim} \nu$ from which to extract information about the conditional law $U \mid Y = y$. As an example of how this works, we might consider the task of computing a posterior moment $\mathbb{E}[\phi(U) \mid Y = y]$. Taking this as a function $F_\phi(y) := \mathbb{E}[\phi(U) \mid Y = y]$ of the possible observed data, we have $F_\phi = \arg \min_{g \in L^2_{\nu_y}(\mathcal{Y})} \mathbb{E}[\|g(Y) - \phi(U)\|^2]$. As such, F_ϕ may be approximated by solving the empirical regression problem

$$\hat{F}_\phi = \arg \min_{g \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \|g(y_i) - \phi(u_i)\|^2, \quad (1.1)$$

after which $F_\phi(y) \approx \hat{F}_\phi(y)$.

The generative variant of this approach goes one step further: we fit a *conditional generative model* from the dataset that gives an approximation $\hat{\nu}(\mathrm{d}u \mid y) \approx \nu(\mathrm{d}u \mid y)$ in an appropriate sense, allowing us to cheaply generate samples from the posterior for any y after training. This embeds the entire computational task of inference into *measure valued regression*.

This immediately raises the question of how to represent probability measures in a way that allows for convenient sampling. We have efficient ways to represent functions — basis expansions, neural networks, kernel methods — and could parametrize a measure through its (log-)density, but doing so leaves us with the original problem of how to sample from

it. Transport maps offer an alternative: given a *reference distribution* η and a target measure ν , a transport map is a function $T : \mathcal{U} \rightarrow \mathcal{U}$ which transforms η into ν via its pushforward, denoted with a $\#$ symbol and defined by $T\#\eta(A) = \eta(T^{-1}(A))$. For T to be a transport map from η to ν we require that $T\#\eta = \nu$, or in other words, $T(V) \stackrel{d}{=} U$ in the distributional sense, for $V \sim \eta$ and $U \sim \nu$. Abstractly, we may therefore approach transport-based generative modeling as the following task: given a reference measure η and samples $\{u_i\} \stackrel{iid}{\sim} \nu$, find a map $\hat{T}$ such that $\hat{T}\#\eta \approx \nu$.

In our setting of simulation based inference, however, we must go even further and approximate *conditional* measures $\nu(\mathrm{d}u \mid y)$ from joint samples. The key observation is that when the map $T : \mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{Y} \times \mathcal{U}$ possesses a (block) triangular structure

$$T(y, u) = (y, T_{\mathcal{U}}(y, u)), \quad (1.2)$$

and acts as a valid transport map in the sense that

$$T\#(\nu_{\mathcal{Y}} \otimes \eta_{\mathcal{U}}) = \nu, \quad (1.3)$$

where $\nu_{\mathcal{Y}}$ denotes the $\mathcal{Y}$ -marginal of ν and $\eta_{\mathcal{U}} \in \mathcal{P}(\mathcal{U})$ is a reference on $\mathcal{U}$, then T naturally generates conditionals of the posterior, so that

$$T_{\mathcal{U}}(y, \cdot)\#\eta_{\mathcal{U}} = \nu(\cdot \mid y). \quad (1.4)$$

For any given pair of measures, there will exist many transport maps between them. Optimal transport maps are one set of choices, solving the optimization problem

$$\min_T \int_{\mathcal{U}} c(v, T(v)) \mathrm{d}\eta(v) \quad \text{subject to} \quad T\#\eta = \nu, \quad (1.5)$$

for a cost function $c : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}$. The chapter that follows develops the existence, uniqueness, and stability theory of these conditional optimal transport maps in separable Hilbert spaces, and presents applications to Bayesian inverse problems.

To summarize, in the simulation based/amortized approach to Bayesian inference we seek to approximate the posterior distribution parametrically across the space of possible data observations y . In doing so, we fit a map from the data y to a desired quantity of interest; in the generative framework, this quantity is the entire posterior distribution, which is a probability measure. Phrasing this in terms of transport maps casts the problem of representing measures as one of representing functions, and triangular transport pushes this further to introduce y -parametric dependence for a measure valued regression problem. This article develops conditional optimal transport as the instantiation of triangular transport in the framework of optimal transport.

1.1 Introduction and setup

This chapter presents a systematic study of conditional optimal transport (OT) on separable infinite dimensional function spaces towards likelihood-free and amortized Bayesian inference. We introduce and study the solvability of the conditional Kantorovich problem

for general costs and measures followed by the analogous Monge formulation. We then give refined results on conditional OT maps under stronger assumptions on the underlying measures with applications to Bayesian inverse problems (BIPs). Our analysis leads to algorithms for the computation of optimal (block) triangular maps for likelihood-free and amortized solution of partial differential equation (PDE) inverse problems.

Sampling and simulation of infinite or high dimensional probability measures is a fundamental task in data science, computational statistics, and uncertainty quantification (UQ), with Markov chain Monte Carlo (MCMC) [5–15] and variational inference (VI) [16–18] being the most broadly used approaches for this task. In recent years, a lot of interest has been generated around the family of transport-based sampling algorithms [19] due to the rise of generative models [20, 21] such as generative adversarial networks (GANs) [22–24], normalizing flows (NFs) [25–27], and most recently, diffusion models (DMs) [28–32]. Broadly speaking, these transport-based algorithms proceed as follows: given Hilbert spaces $\mathcal{U}, \mathcal{V}$, a *target* Borel probability measure $\nu \in \mathbb{P}(\mathcal{U})$, and a *reference* Borel probability measure $\eta \in \mathbb{P}(\mathcal{V})$, a (parametric and possibly stochastic) transport map T^* is computed such that $T^*\#\eta \approx \nu$. The reference η is often chosen to be trivial to simulate, such as a Gaussian measure, so that ν -samples can be (approximately) generated by drawing $v \sim \eta$ and then evaluating $T^*(v)$. The map T^* is often computed by minimizing a statistical divergence between the pushforward $T^*\#\eta$ and the target ν or their empirical approximations. To this end, such transport-based algorithms can be categorized under VI as minimum divergence estimators [33].

Naturally, this line of thinking leads one to the idea of characterizing T^* as an OT map pushing η to ν [34], connecting the fields of generative modeling and sampling to computational OT [35]. Exploring this connection has become an area of intense interdisciplinary research with ideas from OT finding use in generative modeling [36–42] and, conversely, Machine Learning (ML) technology inspiring algorithms for OT [43–48]. We should note that the overwhelming majority of contributions in the aforementioned areas are focused on the finite-dimensional setting with the exception of [49–53] that focus specifically on infinite dimensional spaces.

Our goal in this chapter is to study a particular type of transport aimed at characterizing conditional measures. We pose the following problem:

Problem 1.1 (Conditional transport). Consider separable Hilbert spaces $\mathcal{Y}, \mathcal{U}, \mathcal{V}$ and Borel probability measures $\nu \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ and $\eta_{\mathcal{V}} \in \mathbb{P}(\mathcal{V})$. Let $\nu(\cdot \mid y)$ denote conditional of ν on the y variable. Find a transport map $T_{\mathcal{U}}^* : \mathcal{Y} \times \mathcal{V} \rightarrow \mathcal{U}$ such that $T_{\mathcal{U}}^*(y, \cdot)\#\eta_{\mathcal{V}} = \nu(\cdot \mid y)$.

Here, the measure $\eta_{\mathcal{V}}$ is a reference measure on $\mathcal{V}$ that is assumed to be easy to simulate. Assuming that we could solve the above problem, then we obtain a simple recipe for the simulation of samples not just from ν , but any of its conditionals $\nu(\cdot \mid y)$ for a.s. choice of y since we simply have to draw $v \sim \eta_{\mathcal{V}}$ and evaluate $T_{\mathcal{U}}^*(y, v) \sim \nu(\cdot \mid y)$. Such a strategy falls under the category of *amortized inference* [17, 54], where an upfront cost is paid for the calculation of $T_{\mathcal{U}}^*$ which can later generate conditional samples from $\nu(\cdot \mid y)$ for any choice of y . Such a strategy is inherently different from more traditional approaches such as MCMC, which directly target $\nu(\cdot \mid y)$ for a single choice of y .

Note that we may view the solution $T_{\mathcal{U}}^*(y, \cdot)$ to Problem 1.1 simply as a family of trans-

port maps from $\eta_{\mathcal{U}}$ to $\nu(\cdot \mid y)$ parameterized by the variable y . While such maps can be defined for each value of y individually, our goal in conditional transport is to simultaneously characterize these maps for all values of y towards applications in amortized Bayesian inference.

For example, consider the Darcy flow PDE that arises in various applications, such as the subsurface flows [55] or electrical impedance tomography [56]:

$$-\operatorname{div} \exp(u^\dagger) \nabla p = f \quad \text{in } \Omega \quad \text{and} \quad p = 0 \quad \text{on } \partial\Omega. \quad (1.6)$$

Here, $\Omega \subset \mathbb{R}^d$ is a smooth domain with boundary $\partial\Omega$, $u^\dagger$ is a smooth field that models the permeability of the rock, p is the pressure field, and f is the source term. Suppose we have limited and noisy measurements $y = (p(x_1), \dots, p(x_m)) + \epsilon \in \mathbb{R}^m$ of the pressure field p at a set of points $\{x_1, \dots, x_m\} \subset \Omega$ and with Gaussian noise $\epsilon \sim N(0, \sigma^2 I)$. Our goal is to estimate the log-permeability field $u^\dagger$ given the data y . Employing the Bayesian perspective, we choose a prior measure $\mu \in \mathbb{P}(\mathcal{U})$, e.g., a Gaussian random field, and model $u^\dagger$ with the posterior measure $\nu(\cdot \mid y) \in \mathbb{P}(\mathcal{U})$ defined via Bayes' rule [6]:

$$\frac{d\nu(\cdot \mid y)}{d\mu}(u) = \frac{1}{Z(y)} \exp(-\Phi(u; y)), \quad Z(y) := \int_{\mathcal{U}} \exp(-\Phi(u; y)) d\mu(u), \quad (1.7)$$

with Φ denoting the negative log-likelihood of y given u ; see (1.28) for its expression. The standard approach for sampling $\nu(\cdot \mid y)$ is to use function space MCMC algorithms such as preconditioned-Crank-Nicolson (pCN) or Metropolis Adjusted Langevin (MALA) [7], but these algorithms require long sequential runs for each value of y . However, by solving Problem 1.1 we can obtain an approximation $\widehat{T}_{\mathcal{U}}^*$ to the map $T_{\mathcal{U}}^*$ such that $\widehat{T}_{\mathcal{U}}(y, \cdot) \# \mu \approx \nu(\cdot \mid y)$. We use this approach in Section 1.5.3 and compare the performance of a triangular map approach to that of MCMC; see in particular Figure 1.3. The advantage of this approach is that the map is only approximated once and can then be evaluated for any value of y to generate posterior samples. For comparison, an MCMC chain needs to be run to convergence for every new value of y .

In the rest of this section, we introduce our setup and a summary of our contributions in a simplified setting in Section 1.1.1 followed by a review of relevant literature in Section 1.1.2 and an outline of the article in Section 1.1.3.

1.1.1 Setup and summary of contributions

Our primary focus is the study of Problem 1.1 in the context of OT. In particular, we will consider maps $T_{\mathcal{U}}$ that are obtained as a sub-component of a triangular transport map by solving an appropriately constrained OT problem. Following [57, Sec. 10.10(vii)] and [53] we define the class of triangular maps:

Definition 1.2. Consider Hilbert spaces $\mathcal{V}, \mathcal{Z}, \mathcal{U}, \mathcal{Y}$. A map $T : \mathcal{Z} \times \mathcal{V} \rightarrow \mathcal{Y} \times \mathcal{U}$ is (block) *triangular* if there exist maps $T_{\mathcal{Y}} : \mathcal{Z} \rightarrow \mathcal{Y}$ (the $\mathcal{Y}$ -component) and $T_{\mathcal{U}} : \mathcal{Y} \times \mathcal{V} \rightarrow \mathcal{U}$ (the $\mathcal{U}$ -component) so that $T(z, v) = (T_{\mathcal{Y}}(z), T_{\mathcal{U}}(T_{\mathcal{Y}}(z), v))$ for all $(z, v) \in \mathcal{Z} \times \mathcal{V}$. Henceforth, we write $\mathcal{T}_{\triangle}(\mathcal{Z}, \mathcal{V}; \mathcal{Y}, \mathcal{U})$ to denote the space of such maps.

The significance of the triangular maps lies in the implication of our Proposition 1.11: Consider a reference measure $\eta \in \mathbb{P}(\mathcal{Y} \times \mathcal{V})$ of the form $\eta = \eta_{\mathcal{Z}} \otimes \eta_{\mathcal{V}}$, where $\eta_{\mathcal{Z}} \in \mathbb{P}(\mathcal{Z})$ and $\eta_{\mathcal{V}} \in \mathbb{P}(\mathcal{V})$ are arbitrary, i.e., η is the independent coupling of $\eta_{\mathcal{Z}}$ and $\eta_{\mathcal{V}}$. Then if $T \in \mathcal{T}_{\sqsubseteq}(\mathcal{Z}, \mathcal{V}; \mathcal{Y}, \mathcal{U})$ satisfies $T\#\eta = \nu$, its $\mathcal{U}$ -component $T_{\mathcal{U}}$ solves Problem 1.1. This observation leads to numerical algorithms that are capable of generating conditional samples for arbitrary choices of the conditioning variable y as depicted in Figure 1.1 where we used our Algorithm 1.

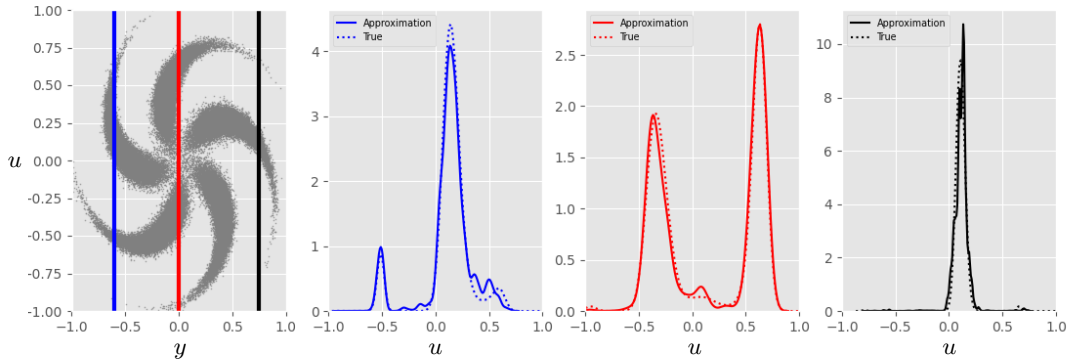


Figure 1.1: A depiction of conditional sampling using triangular maps obtained from Algorithm 1. The left panel shows a scatter plot of samples from ν that were used to solve the conditional OT problem while the blue, red, and black lines denote slices along which conditional samples are generated. The other panels show the conditional histograms with dotted lines showing the true histograms and solid lines showing the numerical approximations.

In most applications, we have freedom in the choice of the reference η , so we will present most of our results in the particular setting where $\mathcal{Z} = \mathcal{Y}$ and $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{V}}$ where $\nu_{\mathcal{Y}}$ denotes the $\mathcal{Y}$ -marginal of ν and $\eta_{\mathcal{V}}$ is an arbitrary measure. This setup is natural for inverse problems since it is often easy for us to simulate $\nu_{\mathcal{Y}}$. For example, consider the setting of the Darcy flow inverse problem described earlier. Then, we can generate samples from ν by the following procedure: (1) draw $u^j \sim \mu$, (2) solve the PDE (1.6) using a numerical solver to obtain the solution $p(u^j)$, and then (3) simulate the measurements $y^j = (p(u^j)(x_1), \dots, p(u^j)(x_m)) + \epsilon^j$. This procedure generates samples $(y^j, u^j) \sim \nu$ where $y^j \sim \nu_{\mathcal{Y}}$.

To this end, we define the following OT problem, which we refer to as the *conditional Monge problem* in the parlance of [58]:

Problem 1.3 (Conditional Monge). Consider $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{V}}$ and a cost function $c : (\mathcal{Y} \times \mathcal{V}) \times (\mathcal{Y} \times \mathcal{U}) \mapsto \mathbb{R}$. Find T^* that solves

$$\begin{cases} \inf_T M(T), & M(T) := \int_{\mathcal{Y} \times \mathcal{V}} c(z, v; T(z, v)) d\eta(z, v), \\ \text{subject to (s.t.) } T(z, v) = (z, T_{\mathcal{U}}(z, v)) \quad \text{and} \quad T\#\eta = \nu. \end{cases} \quad (1.8)$$

Note that we take $T_{\mathcal{Y}} = \text{Id}$ since we chose our reference η to have the same $\mathcal{Y}$ -marginal as ν , which means we do not need to transport any mass in that direction. Since the maps in this problem are clearly triangular, the resulting $\mathcal{U}$ -component of T^* will naturally solve Problem 1.1, i.e., $T_{\mathcal{U}}^*(y, \cdot) \# \eta_{\mathcal{Y}} = \nu(\cdot \mid y)$.

Letting $\Pi(\eta; \nu)$ denote the set of couplings between η and ν , we then define the *conditional Kantorovich problem* that serves as the relaxation of the conditional Monge problem above:

Problem 1.4 (Conditional Kantorovich). Consider $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{V}}$ and a cost function $c : (\mathcal{Y} \times \mathcal{V}) \times (\mathcal{Y} \times \mathcal{U}) \mapsto \mathbb{R}$. Find a coupling π^* that solves

$$\begin{cases} \inf_{\pi} K(\pi) & K(\pi) := \int_{\mathcal{Y} \times \mathcal{V} \times \mathcal{Y} \times \mathcal{U}} c(z, v; y, u) d\pi(z, v; y, u) \\ \text{s. t.} & \pi \in \Pi_{\mathcal{Y}}(\eta, \nu) := \{\gamma \in \Pi(\eta; \nu) \mid (z, v; y, u) \sim \gamma, \text{ such that } z = y\}. \end{cases} \quad (1.9)$$

Simply put, this Kantorovich problem is solved over a constrained set of couplings where the z, y variables are deterministically coupled by the identity map.

With this setup, our first result establishes the well-posedness of the conditional Kantorovich problem under mild assumptions and characterizes its optimal solution in terms of optimal couplings between $\eta_{\mathcal{V}}$ and the conditionals $\nu(\cdot \mid y)$. Indeed, our result states that Problem 1.4 has an optimal solution that can be obtained by coupling z, y using the identity map and then coupling $\eta_{\mathcal{V}}$ and $\nu(\cdot \mid y)$ optimally for a.s. y .

Theorem 1.5. *Consider the setting of Problem 1.4, suppose that the cost c is continuous, $\inf c > -\infty$, and that the infimum in (1.9) is finite. Then it holds that*

- (i) *Problem (1.9) admits a minimizer $\pi^* \in \mathbb{P}(\mathcal{Y} \times \mathcal{V} \times \mathcal{Y} \times \mathcal{U})$.*
- (ii) *It holds that $\pi^*(z, v, u \mid y) = \pi^*(v, u \mid y) \delta(y - z)$ where for $\mathcal{Y}\nu_{\mathcal{Y}}$ -a.s. $y \in \mathcal{Y}$ the conditional measures $\pi^*(v, u \mid y)$ are optimal for*

$$\inf_{\pi \in \Pi(\eta_{\mathcal{V}}, \nu(\cdot \mid y))} \int_{\mathcal{V} \times \mathcal{U}} c(y, v; y, u) d\pi(v, u).$$

This theorem follows directly from Proposition 1.19. We note that the assumption that the infimum of (1.9) is finite is often mild in practice and can be reduced to appropriate moment conditions on η and ν ; see Remark 1.21 for details.

Our second contribution is a strong duality result under slightly different assumptions on the cost c which follows from Proposition 1.22.

Theorem 1.6. *Consider Problem 1.4 with a lower semi-continuous and non-negative cost c . Then it holds that*

$$\begin{aligned} \inf_{\pi \in \Pi_{\mathcal{Y}}(\eta, \nu)} K(\pi) &= \sup_{\psi \in L^1(\eta)} \left(\int_{\mathcal{Y} \times \mathcal{U}} \psi_v^c(z, u) d\nu(z, u) - \int_{\mathcal{Y} \times \mathcal{V}} \psi(z, v) d\eta(z, v) \right), \\ &= \sup_{\phi \in L^1(\nu)} \left(\int_{\mathcal{Y} \times \mathcal{U}} \phi(y, u) d\nu(y, u) - \int_{\mathcal{Y} \times \mathcal{V}} \phi_u^c(y, v) d\eta(y, v) \right), \end{aligned}$$

where $\psi_v^c(z, u) := \inf_v \left(\psi(z, v) + c(z, v; z, u) \right)$ and $\phi_u^c(y, v) := \sup_u \left(\phi(y, u) - c(y, v; y, u) \right)$ $\phi_u^c(y, v) := \sup_u \left(\phi(y, u) - c(y, v; y, u) \right)$ are the partial c -transforms of ψ and ϕ respectively.

For our third contribution, we turn our attention to Problem 1.3 and show that the conditional Kantorovich problem has a solution that is given by a unique transport map. This result is stated in Proposition 1.24 under slightly more general assumptions.

Theorem 1.7. *Suppose $\mathcal{V} = \mathcal{U}$ and consider the cost $c(z, v; y, u) = \|u - v\|_{\mathcal{U}}^p$ for $p \in (1, \infty)$. Suppose that (a) the measures η, ν have finite p -th moments; (b) $\eta_{\mathcal{V}}$ is regular according to Definition 1.14; and (c) the infimum in (1.9) is feasible and finite. Then there exists a map T^* that uniquely solves Problem 1.3 and $\pi^* = (\text{Id} \times T^*)\#\eta$ is the unique solution to Problem 1.4.*

We can formally think of the map T^* as the map that is obtained by considering the parametric OT maps

$$T_y^* = \arg \min_{T: T\#\eta_{\mathcal{V}} = \nu(\cdot | y)} \int_{\mathcal{U}} \|u - T(u)\|_{\mathcal{U}}^p d\eta_{\mathcal{V}}(u)$$

that transport $\eta_{\mathcal{V}}$ to the conditionals $\nu(\cdot | y)$ for each y . One can then concatenate these maps along the y coordinate to define the map $T_{\mathcal{U}}^*(y, v) = T_y^*(v)$ in Problem 1.3. This turns out to be the correct approach, although it leads to delicate issues of measurable selection for $T_{\mathcal{U}}$ as dealt with in [59]. Here, we circumvent these technical details to the assumption that (1.9) has a feasible solution which transfers the issue of measurable selection to the Kantorovich problem where it can be overcome easily.

While analogues of some of the above results appear in [59], we develop and extend the theory of conditional OT from a different perspective. We first study the conditional Kantorovich problem with general costs and characterize its solutions while [59] focuses on the conditional Monge problem and the characterization of conditional transport maps using Brenier potentials, which is naturally limited to quadratic costs. By starting with a Kantorovich problem, our approach naturally requires fewer assumptions at first but allows us to prove Monge solvability as a special case and under stronger assumptions.

Employing the perspective of parameterized OT maps, a natural question arises as to the regularity of the maps $T_{\mathcal{U}}^*(y, \cdot)$ as a function of y . We apply existing quantitative stability estimates [60, Thm. 4.2] in finite dimensions to show that conditional Brenier maps are Hölder continuous with respect to the conditioning variable y if the conditional measures $\nu(\cdot | y)$ are also "smooth" in an appropriate sense. We state this result as the following theorem, which follows from Proposition 1.33. To the best of our knowledge, no such quantitative stability estimates exist for other types of triangular maps.

Theorem 1.8. *Let $\mathcal{U}$ be a compact subset of $\mathbb{R}^d$, $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{U}} \in \mathcal{P}(\mathcal{Y} \times \mathcal{U})$ and set $c(z, v; y, u) = \|u - v\|^2$. Assume further that the support of $\eta_{\mathcal{U}}$ is a convex subset of $\mathcal{U}$ and $\eta_{\mathcal{U}}$ has a density that is bounded above and below. Set T^* to be the solution to Problem 1.3. Then*

$$\|T_{\mathcal{U}}^*(y, \cdot) - T_{\mathcal{U}}^*(y', \cdot)\|_{L^2(\eta_{\mathcal{U}})} \leq CW_1(\nu(\cdot | y), \nu(\cdot | y'))^{1/6},$$

where the constant $C > 0$ depends on $\mathcal{U}$, $\eta_{\mathcal{U}}$, and d .

Our next theoretical contribution concerns the application of our results above to the particular case of BIPs where we show that in the finite-dimensional case $\mathcal{U} = \mathbb{R}^d$, and under regularity assumptions on the likelihood Φ , and the prior measure μ , the conditioning maps $T_{\mathcal{U}}^*$ are regular in y . We also obtain error bounds for certain numerical approximations of such maps. The detailed statement of these results can be found in Proposition 1.33.

We note that optimality in the Brenier sense is not necessarily crucial for solving BIPs using triangular maps. However, the OT characterisation is amenable to analysis and provides desirable theoretical features such as uniqueness and regularity of transport maps; these may be difficult to obtain for other types of maps.

Finally, we present a set of numerical experiments that explore various practical strategies for the solution of Problem 1.3 and approximation of the maps $T_{\mathcal{U}}^*(y, u)$ towards conditional simulation of probability measures. Our algorithms and experiments, including an instance of the Darcy flow inverse problem (1.6) are collected in Section 1.5.

1.1.2 Review of relevant literature

Below, we review the relevant literature to our work with a particular focus on triangular transport and conditional simulation. For a comprehensive review of OT theory see [34, 61–63] and for computational aspects see [35].

Triangular transport and OT

The OT characterization of triangular transport maps goes back to [64] and [65] which characterized the classic triangular rearrangements of Knothe [66] and Rosenblatt [67] as the limit of a sequence of OT problems with a singularly perturbed quadratic cost. We show in Section 1.3.4 that a similar limiting argument is applicable in our broader setting as well. The Knothe-Rosenblatt (KR) map is a classic transport map originally defined in the finite-dimensional setting, which is built by successively transporting one measure into another one by 1D transport of their conditionals using deterministic maps; see [63, Sec. 2.3] for a detailed definition. The KR map is an example of a fully triangular map according to Definition 1.2 and it was further generalized to triangular maps defined on Hilbert spaces in [68, 69] where their abstract theory was developed; also see [57, Sec. 10.10(vii)].

The OT characterization of triangular maps towards conditional simulation is a contemporary topic, and the main results in this direction were presented in the seminal work [59] where an analog for Brenier’s theorem was presented for the characterization of parametric families of transport maps $T_{\mathcal{U}}$ that would characterize the conditionals $\nu(\cdot \mid y)$ of a given target measure ν , i.e., $T_{\mathcal{U}}(y, \cdot) \# \eta_{\mathcal{U}} = \nu(\cdot \mid y)$. However, the main goal of the authors in that work was to introduce a novel characterization of quantiles for vector valued random variables and the theory was not tailored to conditional simulation or triangular transport. This extension was discussed in the recent works [58, 70–75] although these works were focused on finite-dimensional settings with quadratic cost functions and are mainly concerned with the design of algorithms. Our analysis is heavily inspired by these previous works and, in particular [59], but deviates from them in three directions: (a) we develop our theory on

the Hilbert and Polish space settings, which is crucial for PDE inverse problems and high dimensional inference; (b) we study OT problems for arbitrary costs and give refined results for classic choices of the cost, such as the quadratic case; (c) we give a detailed characterization of the corresponding constrained Kantorovich problems that lead to triangular Monge maps.

After the first appearance of this work in preprint form, [76] further developed the theory of conditional OT in the context of sampling. There, the metric properties of conditional OT problems were examined, and a dynamic formulation analogous to that of Benamou and Brenier in standard OT was presented. That work also introduced an algorithm based on flow matching [77–79] for the approximation of conditional OT maps as an alternative to our algorithms in Section 1.5.

Triangular transport and inference

The algorithmic aspects of triangular transport maps for conditional sampling and Bayesian inference developed in [19, 80] rely on the KR constructions of triangular maps. These are used for sampling of target measures or preconditioning MCMC algorithms. The extension to general block-triangular maps was presented in [53, 73] with the former developing algorithms using monotone triangular transport maps and the latter focusing on OT formulations akin to ours with quadratic costs. More recently, these methods have been extended for amortized inference with applications in geoscience [54, 81–83] and filtering and data assimilation [84–86] with [71, 74, 87] focusing on OT formulations that are particular instances of our formulation of the conditional Monge problem. These works further characterize the map $T_{\mathcal{U}}^*$ via a dual formulation using the conditional analog of Brenier’s theorem presented in [59] which is also the basis of the methodologies developed in [58, 70, 75].

Triangular transport and generative modeling

Broadly speaking, modern generative models [20, 21] solve the problem of generating samples from a target measure as a transport problem with a deterministic, stochastic, or dynamic formulation. These models have become very popular in recent years thanks to the success of various families of neural net based techniques such as GANs [22–24], NFs [25–27, 40], and DMs [28–32, 52]. Indeed, GANs and NFs provide deterministic maps, while DMs give a dynamic formulation in the form of a stochastic differential equation with a score function (i.e., drift) that is parameterized by a network. Among these models, NFs are the only ones that rely on triangular transport maps, but the conditioning properties of these maps are rarely utilized or acknowledged in the NF literature. Instead, the triangular structure is mainly utilized for building maps that are easily invertible, and their Jacobian determinants can be computed efficiently.

While traditional generative modeling does not focus on conditional simulation, the methodology can be generalized to conditional simulation; see [70, 88–90] for conditional NFs, [53, 91–93] for conditional GANs, and [94, 95] for conditional DMs. Here one considers pushing a reference $\eta_{\mathcal{U}}$ to the conditionals $\nu(\cdot | y)$ with a map $T_{\mathcal{U}}(y, \cdot)$ that is parameterized by y and in this light our setting can be viewed as the OT analog of such conditional generative models. The works [53, 70, 92, 96–98] are perhaps the closest to our work in this

domain, although they do not present a systematic study of conditional OT at our level of generality.

Other interesting connections

Here, we mention a number of articles in the literature that are related to our work but do not directly study the conditional OT or conditional simulation problems. While our focus here is not on algorithm development, we mention that the applied analysis of transport problems has become a vibrant area of research in recent years. The statistical analysis of OT maps has been an intense research area recently [99–105]. These works often aim to obtain error bounds for empirical estimation of OT maps in terms of the approximation capacity of underlying models and sample complexity.

Alternatively, the article [106] gives a general framework for obtaining quantitative approximation rates for minimum divergence estimators but without necessarily approximating the *optimal* transport map. In parallel with the above, a growing literature has been developed around the analysis of triangular transport maps, and in particular, the KR map: [107, 108] show that the KR map is analytic under some assumptions and can be approximated with neural nets or sparse polynomials; [109, 110] consider minimum KL estimators of KR maps for sampling and density estimation and study their statistical consistency and sample complexities; finally, [111] analyzes the tail behavior of triangular maps, revealing a balance between the tails of the reference and target measures and the expressivity of Lipschitz triangular maps.

We also mention the literature on *sliced Wasserstein distances* [112–115] and *sliced measure transport* [116, 117]. Broadly speaking, the sliced Wasserstein distance is computed by taking random 1D projections of the reference and the target and averaging the 1D Wasserstein distance between these projections. The resulting distance can be computed very efficiently in practice while still imposing a topology similar to that of the Wasserstein distance. The resulting formulation is opposite to that of (1.8) in that the sliced Wasserstein distance marginalizes across values of projections rather than conditioning.

Another related concept is that of pairs of measures with *overlapping marginals* that arise in distributionally robust optimization [118–120]. These are precisely pairs of measures whose marginals on a given subspace match, akin to our assumption that the $\eta_Y = \nu_Y$ in Problem 1.3 and Problem 1.4.

1.1.3 Outline

The rest of the chapter is organized as follows: we review some preliminaries and useful technical results in Section 1.2. Our main theoretical results are presented in Section 1.3 followed by their application to the particular case of BIPs in Section 1.4. Our algorithms and numerical experiments are collected in Section 1.5 followed by our conclusions in Section 1.6.

1.2 Preliminaries

In this section we collect some notations, definitions, and preliminary results that are used throughout the rest of the chapter. We collect some notation in Section 1.2.1. Section 1.2.2

reviews conditional measures and for the most part follows [57, Sec. 10.4] while Section 1.2.3 recalls useful results from OT theory following [34, 62].

1.2.1 Notation

Throughout the chapter, we primarily work with Polish spaces and separable Banach or Hilbert spaces, but the reader may take all of our function spaces to be separable and Hilbert on a first reading. We also reserve upper case calligraphic letters to denote function spaces. For a Polish space $\mathcal{Y}$ we write $\mathcal{B}(\mathcal{Y})$ to denote its Borel σ -algebra and $\mathbb{P}(\mathcal{Y})$ to denote the space of Borel probability measures on $\mathcal{Y}$. For $\varrho \in \mathbb{P}(\mathcal{Y})$ and a function $f : \mathcal{Y} \rightarrow \mathbb{R}$ we write $\int_{\mathcal{Y}} f(y) d\varrho(y)$ to denote the expectation or integral of f with respect to ϱ and sometimes use the shorthand notation $\int f d\varrho$ when no confusion may arise. We often work on product spaces and need to compute integrals with respect to marginals and write $\int_{\mathcal{Y}} f(x, y) \varrho(x, dy)$ to highlight the variable with respect to which the partial integral is computed. For $\varrho \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ we write $\varrho_{\mathcal{Y}}$ and $\varrho_{\mathcal{U}}$ to denote its $\mathcal{Y}$ and $\mathcal{U}$ -marginals respectively.

1.2.2 Conditional measures and triangular maps

Let $\mathcal{Y}, \mathcal{U}$ be Polish spaces and for a set $B \in \mathcal{B}(\mathcal{Y} \times \mathcal{U})$, define the slices $B^y := \{u \mid (y, u) \in B\}$. We recall the following definition of (regular) conditional measures following [57]:

Definition 1.9. A function $(B, y) \mapsto \varrho(B \mid y)$ is a (regular) *conditional measure* of $\varrho \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ if: (a) For every fixed $y \in \mathcal{Y}$ it holds that $\varrho(\cdot \mid y) \in \mathbb{P}(\mathcal{U})$; (b) For every fixed $B \in \mathcal{B}(\mathcal{U})$ the function $y \mapsto \varrho(B \mid y)$ is measurable with respect to $\mathcal{B}(\mathcal{Y})$ and $\varrho_{\mathcal{Y}}$ -integrable; and (c) For all $B \in \mathcal{B}(\mathcal{Y} \times \mathcal{U})$ it holds that $\varrho(B) = \int_{\mathcal{Y}} \varrho(B^y \mid y) \varrho_{\mathcal{Y}}(dy)$.

We further recall the following facts from [57, Lem. 10.4.3 and Cor. 10.4.10]:

Proposition 1.10. Consider $\varrho \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ with $\mathcal{Y}, \mathcal{U}$ Polish spaces. Then the conditional measures $\varrho(\cdot \mid y)$ for ϱ exist and they are essentially unique, i.e., there exists a set $E \in \mathcal{B}(\mathcal{Y})$ so that $\varrho_{\mathcal{Y}}(E) = 0$ and the $\varrho(\cdot \mid y)$ are unique for all $y \in \mathcal{Y} \setminus E$.

We can further characterize conditional measures via triangular transport maps. The proof of the following proposition is nearly identical to [53, Thm. 2.4] and is therefore omitted.

Proposition 1.11. Let $\mathcal{Z}, \mathcal{V}, \mathcal{Y}, \mathcal{U}$ be Polish spaces and consider $\eta \in \mathbb{P}(\mathcal{Z} \times \mathcal{V})$ and $\nu \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$. If $T \# \eta = \nu$ for some $T \in \mathcal{T}_{\triangle}(\mathcal{Z}, \mathcal{V}; \mathcal{Y}, \mathcal{U})$ (recall Definition 1.2) then $T_{\mathcal{U}}(T_{\mathcal{Y}}(z), \cdot) \# \eta(\cdot \mid z) = \nu(\cdot \mid T_{\mathcal{Y}}(z))$, for $\eta_{\mathcal{Y}}$ -a.e. z .

1.2.3 Basics of OT

Here, we collect some preliminary results on OT as they pertain to our analysis in Sections 1.3 and 1.4 and refer the reader to [34, 35, 61–63] for details.

Let $\mathcal{X}, \mathcal{W}$ be Polish spaces and consider Borel probability measures $\eta \in \mathbb{P}(\mathcal{X})$ and $\nu \in \mathbb{P}(\mathcal{W})$. We define $\Pi(\eta, \nu) := \{\pi \in \mathbb{P}(\mathcal{X} \times \mathcal{W}) : \pi_{\mathcal{X}} = \eta, \pi_{\mathcal{W}} = \nu\}$ as the space of

couplings or transference plans between η and ν . Given a cost $c : \mathcal{X} \times \mathcal{W} \rightarrow (-\infty, \infty]^1$ we consider the Kantorovich problem,

$$\inf_{\pi \in \Pi(\eta, \nu)} K(\pi), \quad K(\pi) = \int_{\mathcal{X} \times \mathcal{W}} c(x, w) d\pi(x, w), \quad (1.10)$$

and the corresponding Monge problem,

$$\inf_{T \# \eta = \nu} M(T), \quad M(T) = \int_{\mathcal{X}} c(x, T(x)) d\eta(x). \quad (1.11)$$

We say that a coupling π^* is optimal if it solves (1.10) and similarly say that a map T^* is *optimal* if it solves (1.11). The existence of optimal couplings for the Kantorovich problem can be established under very general conditions:

Proposition 1.12 ([34, Thm. 4.1]). *Let $\mathcal{X}, \mathcal{W}$ be Polish spaces and let c be a lower semi-continuous function such that $\inf c > -\infty$.² Then there exists a minimizer of (1.10).*

Next, we recall a technical result on measurable selections of couplings that will help us in our analysis of conditional Kantorovich problems (1.9), allowing us to build optimal couplings between measures from optimal couplings between their conditionals.

Proposition 1.13 ([34, Cor. 5.22]). *Suppose $\mathcal{X}, \mathcal{W}$ are Polish spaces and c is a continuous cost function such that $\inf c > -\infty$. Let Ω be a measurable space and let $\omega \mapsto (\eta_\omega, \nu_\omega)$ be a measurable function from Ω to $\mathbb{P}(\mathcal{X}) \times \mathbb{P}(\mathcal{W})$. Then there exists a measurable selection $\omega \rightarrow \pi_\omega$ such that $\pi_\omega \in \Pi(\eta_\omega, \nu_\omega)$ is optimal, i.e. π_ω solves (1.10) with η_ω and ν_ω as marginals, for every $\omega \in \Omega$.*

Let us now consider the Monge problem (1.11) and recall a characterization of its solution in the setting where $\mathcal{W} = \mathcal{X}$ are separable Hilbert spaces. First, we recall the definition of a (Gaussian) regular measure.

Definition 1.14. Suppose $\mathcal{X}$ is a separable Hilbert space. We say $B \in \mathcal{B}(\mathcal{X})$ is a Gaussian null set if $\gamma(B) = 0$ for any non-degenerate Gaussian measure $\gamma \in \mathbb{P}(\mathcal{X})$.³ Furthermore, we say that a measure $\varrho \in \mathbb{P}(\mathcal{X})$ is regular if $\varrho(B) = 0$ for all Gaussian null sets B .

With the above notion of a regular measure at hand, we can present the following result regarding the existence of unique optimal solutions to (1.11) in relation to (1.10).

Proposition 1.15 ([62, Thm. 6.2.10]). *Let $\mathcal{X}$ be a separable Hilbert space and consider the cost $c(u, v) = \|u - v\|_{\mathcal{X}}^p$ for $p \in (1, \infty)$. Suppose $\eta, \nu \in \mathbb{P}(\mathcal{X})$ are both measures with finite p -moments, and assume η is regular according to Definition 1.14. Then it holds that:*

¹Throughout the article we, only consider real valued costs.

²This assumption can be slightly generalized to $c(x, w) \geq a(x) + b(w)$ for upper semi-continuous functions $a \in L^1_\eta(\mathcal{X}), b \in L^1_\nu(\mathcal{W})$ as is done in classic texts.

³Recall that γ is said to be a Gaussian measure if $\ell \# \gamma$ is a one-dimensional Gaussian distribution for all $\ell \in \mathcal{X}^*$ the dual of $\mathcal{X}$ [121].

(i) Problem (1.10) has a unique solution $\pi^* \in \Pi(\eta, \nu)$, such that $\pi^* = (I \times T^*) \# \eta$ for an optimal map $T^* \in L_\eta^p(\mathcal{X}; \mathcal{X})$ that solves (1.11).

(ii) If ν is also regular then T^* is injective η -a.s.

Here $L_\eta^p(\mathcal{X}; \mathcal{X})$ denotes the η -weighted L^p space of Bochner integrable maps from $\mathcal{X}$ to itself. Finally, we present a technical lemma about the convergence of OT maps in L^2 spaces, which may be known by experts, but we present the proof for convenience.

Lemma 1.16. *Let $\mathcal{X}$ be a separable Hilbert space, and let $\eta, \nu \subset \mathbb{P}(\mathcal{X})$ have finite second moments. Let $\pi_k = (\text{Id} \times T_k) \# \eta$ be a sequence of deterministic couplings between η and ν , defined by maps T_k . If the couplings π_k converge weakly, that is, $\pi_k \rightharpoonup \pi$, where $\pi = (\text{Id} \times T) \# \eta$, then the maps $T_k \rightarrow T$ in $L_\eta^2(\mathcal{X}; \mathcal{X})$.*

Proof. Note that $\|T_k\|_{L_\eta^2(\mathcal{X}; \mathcal{X})}^2 = \|T\|_{L_\eta^2(\mathcal{X}; \mathcal{X})}^2 = \int_{\mathcal{X}} \|y\|^2 \nu(dy)$ for all k where the second equality follows from the fact that $T \# \eta = \nu$ and similarly for T_k . Since $\pi_k \rightharpoonup \pi$ then $\lim_{k \rightarrow \infty} \int \phi d\pi_k = \int \phi d\pi$ for all continuous functions $\phi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. We then have that

$$\int_{\mathcal{X} \times \mathcal{X}} \phi(x, x') d\pi_k(x, x') = \int_{\mathcal{X} \times \mathcal{X}} \phi(x, x') d(\text{Id} \times T_k) \# \eta(x, x') = \int_{\mathcal{X} \times \mathcal{X}} \phi(x, T_k(x)) d\eta(x).$$

Repeating a similar calculation for π yields that

$$\lim_{k \rightarrow \infty} \int_{\mathcal{X} \times \mathcal{X}} \phi(x, T_k(x)) d\eta(x) = \int_{\mathcal{X} \times \mathcal{X}} \phi(x, T(x)) d\eta(x).$$

Let $\phi(x, y) = \langle \psi(x), y \rangle$ for a continuous function $\psi : \mathcal{X} \rightarrow \mathcal{X}$. Then we see that $\lim_{k \rightarrow \infty} \int_{\mathcal{X}} \langle \psi(x), T_k(x) \rangle d\eta(x) = \int_{\mathcal{X}} \langle \psi(x), T(x) \rangle d\eta(x)$. Thus, since the continuous functions on $\mathcal{X}$ are dense in $L_\eta^2(\mathcal{X}; \mathcal{X})$,⁴ and the sequence (T_k) is bounded, the equality above is extended to all functions $\psi \in L_\eta^2(\mathcal{X}; \mathcal{X})$, and we see that $T_k \rightharpoonup T$. Finally, we apply the Radon-Riesz property in Hilbert spaces to conclude that weak convergence combined with equality (convergence) of the norms implies strong convergence in $L_\eta^2(\mathcal{X}; \mathcal{X})$. $\square$

1.3 General theory of conditional OT

In this section we present our main theoretical results in the form of propositions and theorems that amount to the extended versions of Theorems 1.5 to 1.7. We study the solvability of the conditional Kantorovich problem in Section 1.3.1; present a strong duality result in Section 1.3.2; and investigate the solvability of the conditional Monge problem in Section 1.3.3. We further characterize the conditional Kantorovich problem as the limit of unconstrained problems with a perturbed cost in Section 1.3.4 followed by an extension to broader classes of triangular maps in Section 1.3.5.

⁴This is a consequence of the density of simple functions in $L_\eta^2(\mathcal{X}; \mathcal{X})$ and the fact that simple functions on separable Hilbert spaces can be approximated by continuous functions by Urysohn's theorem [122, Page 7].

Throughout this section (except for Section 1.3.5) we will consider a reference measure $\eta \in \mathbb{P}(\mathcal{Y} \times \mathcal{V})$ and a target $\nu \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ for function spaces $\mathcal{Y}, \mathcal{V}, \mathcal{U}$ which are either taken to be Polish or separable Hilbert spaces. We further make the following assumption regarding the $\mathcal{Y}$ -marginals of these two measures.

Assumption 1.17. The reference $\eta \in \mathbb{P}(\mathcal{Y} \times \mathcal{V})$ and target $\nu \in \mathbb{P}(\mathcal{Y} \times \mathcal{U})$ satisfy $\eta_{\mathcal{Y}} = \nu_{\mathcal{Y}}$.

Under this assumption we can take the map $T_{\mathcal{Y}} = \text{Id}$ in Definition 1.2 and work with the constrained set of couplings $\Pi_{\mathcal{Y}}$ from (1.9) which we recall for convenience:

$$\Pi_{\mathcal{Y}}(\eta; \nu) = \{\pi \in \Pi(\eta; \nu) : (z, v; y, u) \sim \pi \text{ such that } y = z\}^5 \quad (1.12)$$

By definition, any $\pi \in \Pi_{\mathcal{Y}}(\eta; \nu)$ can be written as

$$\pi(z, v, y, u) = \pi(v, u \mid y, z) \pi(y \mid z) \pi(z) = \tilde{\pi}(v, u \mid y) \delta(y - z) \eta_{\mathcal{Y}}(z), \quad (1.13)$$

as these couplings are concentrated on the set $\{y = z\}$. We now verify that the conditional measure $\tilde{\pi}(v, u \mid y)$ is a coupling of $\eta(\cdot \mid y)$ and $\nu(\cdot \mid y)$ a. e. as a particular application of the following results by taking $T_{\mathcal{Y}} = \text{Id}$.

Proposition 1.18. Let $T_{\mathcal{Y}} : \mathcal{Y} \rightarrow \mathcal{Y}$ be an invertible map such that $T_{\mathcal{Y}} \# \eta_{\mathcal{Y}} = \nu_{\mathcal{Y}}$ and consider the class of couplings

$$\Pi(\eta, \nu; T_{\mathcal{Y}}) := \{\pi \in \Pi(\eta, \nu) : (z, v; y, u) \sim \pi \text{ such that } y = T_{\mathcal{Y}}(z)\}.$$

Then every coupling $\pi \in \Pi(\eta, \nu; T_{\mathcal{Y}})$ can be written as $\pi(z, v, y, u) = \tilde{\pi}(v, u \mid y) \delta(y - T_{\mathcal{Y}}(z)) \eta_{\mathcal{Y}}(z)$ where $\tilde{\pi}(v, u \mid y) \in \Pi(\eta(v \mid T_{\mathcal{Y}}^{-1}(y)), \nu(u \mid y))$ for $\nu_{\mathcal{Y}}$ -a.s. y .

Proof. The expression of π in terms of $\tilde{\pi}$ follows from the expansion $\pi(z, v, y, u) = \pi(v, u \mid y, z) \pi(y \mid z) \pi(z)$ and the constraint that $y = T_{\mathcal{Y}}(z)$ almost surely with respect to π . In order to show the $\mathcal{V}$ marginal of $\tilde{\pi}(v, u \mid y)$ is $\eta(v \mid T_{\mathcal{Y}}^{-1}(y))$, note that

$$\int_{\mathcal{Y}} \int_{\mathcal{U}} \pi(z, v, dy, du) = \eta(z, v) = \eta(v \mid z) \eta_{\mathcal{Y}}(z),$$

because π is a coupling of η and ν . Moreover, the expansion of π in terms of $\tilde{\pi}$ leads to

$$\int_{\mathcal{Y}} \int_{\mathcal{U}} \pi(z, v, dy, du) = \int_{\mathcal{Y}} \int_{\mathcal{U}} \tilde{\pi}(v, du \mid y) \delta(dy - T_{\mathcal{Y}}(z)) \eta_{\mathcal{Y}}(z) = \eta_{\mathcal{Y}}(z) \int_{\mathcal{U}} \tilde{\pi}(v, du \mid T_{\mathcal{Y}}(z)).$$

Hence, we have shown that $\int_{\mathcal{U}} \tilde{\pi}(v, du \mid T_{\mathcal{Y}}(z)) = \eta(v \mid z)$ for a.s. z . Replacing z with $T_{\mathcal{Y}}^{-1}(y)$ yields the conclusion $\int_{\mathcal{U}} \tilde{\pi}(v, du \mid y) = \eta(v \mid T_{\mathcal{Y}}^{-1}(y))$ for a.s. y . The proof is completed by following the same procedure to show that the $\mathcal{V}$ marginal of $\tilde{\pi}(v, u \mid y)$ is $\nu(u \mid y)$. In particular, $\int_{\mathcal{Y}} \int_{\mathcal{V}} \pi(dz, dv, y, u) = \nu(y, u) = \nu(u \mid y) \nu_{\mathcal{Y}}(y)$ and

$$\begin{aligned} \int_{\mathcal{Y}} \int_{\mathcal{V}} \pi(dz, dv, y, u) &= \int_{\mathcal{Y}} \int_{\mathcal{V}} \tilde{\pi}(dv, u \mid y) \delta(y - T_{\mathcal{Y}}(z)) \eta_{\mathcal{Y}}(dz) \\ &= \int_{\mathcal{Y}} \int_{\mathcal{V}} \tilde{\pi}(dv, u \mid y) \delta(y - w) T_{\mathcal{Y}} \# \eta_{\mathcal{Y}}(dw) \\ &= \int_{\mathcal{Y}} \int_{\mathcal{V}} \tilde{\pi}(dv, u \mid y) \delta(y - w) \nu_{\mathcal{Y}}(dw) = \nu_{\mathcal{Y}}(y) \int_{\mathcal{U}} \tilde{\pi}(dv, u \mid y) \end{aligned}$$

⁵Note that this set is empty whenever Assumption 1.17 is violated.

concluding $\int_{\mathcal{V}} \tilde{\pi}(dv, u \mid y) = \nu(u \mid y)$ for a.s. y . $\square$

We interpret the above result as the coupling analog of Proposition 1.11, which states that the constraint $y = T_{\mathcal{Y}}(z)$ automatically enforces $\pi \in \Pi_{\mathcal{Y}}$ to couple the conditionals $\eta(\cdot \mid z)$ and $\nu(\cdot \mid y)$ in an analogous manner in which the $\mathcal{U}$ -component $T_{\mathcal{U}}$ of a triangular transport map T that pushes η to ν will satisfy $T_{\mathcal{U}}(y, \cdot) \# \eta(\cdot \mid T_{\mathcal{Y}}^{-1}(y)) = \nu(\cdot \mid y)$. For most of our results in this section we apply Proposition 1.18 with $T_{\mathcal{Y}} = \text{Id}$ under Assumption 1.17 but later use this result to obtain more general triangular OT maps.

1.3.1 Solvability of the conditional Kantorovich Problem

We now consider the conditional Kantorovich problem (1.9).

$$\inf_{\pi \in \Pi_{\mathcal{Y}}(\eta, \nu)} K(\pi). \quad (1.14)$$

It is helpful for us to write an equivalent unconstrained Kantorovich problem with a modified cost function:

$$\begin{cases} \inf_{\pi \in \Pi(\eta, \nu)} \int c_{\chi}(z, v; y, u) \, d\pi(z, v, y, u), \\ c_{\chi}(z, v; y, u) := c(z, v; y, u) + \chi(y, z), \quad \chi(y, z) := \begin{cases} 0 & y = z, \\ +\infty & \text{otherwise.} \end{cases} \end{cases} \quad (1.15)$$

The minimizers and minimum of these problems coincide, and the two problems are completely equivalent under assumption 1.17. We are now ready to present our main result concerning the solvability of (1.14) and the characterization of its minimizers as optimal couplings between the conditionals $\eta(\cdot \mid y)$ and $\nu(\cdot \mid y)$. Our main Theorem 1.5 also follows from this result.

Proposition 1.19. *Let $\mathcal{V}, \mathcal{U}, \mathcal{Y}$ be Polish spaces and consider a cost function $c : (\mathcal{Y} \times \mathcal{V}) \times (\mathcal{Y} \times \mathcal{U}) \rightarrow \mathbb{R}$ that is continuous and $\inf c > -\infty$. Suppose Assumption 1.17 holds and that the infimum in (1.14) is finite. Then it holds that:*

(i) Problem (1.14) admits a minimizer $\pi^* \in \mathbb{P}(\mathcal{Y} \times \mathcal{V} \times \mathcal{Y} \times \mathcal{U})$.

(ii) $\pi^*(z, v, u \mid y) = \tilde{\pi}^*(v, u \mid y) \delta(y - z)$ where for $\eta_{\mathcal{Y}}$ a.s. y , $\tilde{\pi}^*(v, u \mid y)$ is optimal for

$$\inf_{\pi \in \Pi(\eta(v|y), \nu(u|y))} \int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) \, d\pi(v, u), \quad c^y(v, u) = c(y, v; y, u). \quad (1.16)$$

Proof. Our proof strategy is to choose a candidate for π^* , which is optimal for (1.16). Using this candidate, we then define π^* and verify claims (i, ii). By Proposition 1.12, the problem (1.16) has a solution π^y for every y . Furthermore, the map $y \rightarrow (\eta(\cdot \mid y), \nu(\cdot \mid y))$ is measurable by Proposition 1.10 and so by Proposition 1.13 there exists a measurable choice

$y \mapsto \pi^y$ such that π^y is an optimal transference plan between $\eta(\cdot | y)$ and $\nu(\cdot | y)$.⁶ We can therefore define the candidate coupling $\pi^*(z, v, y, u) = \pi^y(v, u)\delta(y - z)\eta_y(y)$ which we will now show is optimal. By construction, $\pi^* \in \Pi_{\mathcal{Y}}(\eta, \nu)$ and $\tilde{\pi}^*(v, u | y) = \pi^y(v, u)$ which solves (1.16). We now have, by definition of c^y , that

$$\begin{aligned} \int_{\mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\pi^y(v, u) \right] d\eta_{\mathcal{Y}}(y) &= \\ \int_{\mathcal{Y} \times \mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c(z, v; y, u) d\pi^y(v, u) \right] d\delta(y - z) d\eta_{\mathcal{Y}}(y) &= K(\pi^*). \end{aligned}$$

Now consider an arbitrary $\pi \in \Pi_{\mathcal{Y}}(\eta, \nu)$ as in (1.13), with associated cost $K(\pi)$, towards showing that $K(\pi) \geq K(\pi^*)$, which would imply that π^* is optimal. Disintegration along y, z fibers, and recalling the notation of Proposition 1.18, gives

$$\begin{aligned} K(\pi) &= \int_{\mathcal{Y} \times \mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c(z, v; y, u) d\pi(v, u | y, z) \right] d\pi(y, z), \\ &= \int_{\mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\tilde{\pi}(v, u | y) \right] d\eta_{\mathcal{Y}}(y), \end{aligned}$$

where the second identity follows because π is concentrated on $\{y = z\}$ and its $\mathcal{Y}$ -marginal is $\eta_{\mathcal{Y}}$ since it is an element of $\Pi_{\mathcal{Y}}(\eta, \nu)$. We can now compute

$$\begin{aligned} K(\pi) - K(\pi^*) &= \\ &= \int_{\mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\tilde{\pi}(v, u | y) \right] d\eta_{\mathcal{Y}}(y) - \int_{\mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\pi^y(v, u) \right] d\eta_{\mathcal{Y}}(y), \\ &= \int_{\mathcal{Y}} \left[\int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\tilde{\pi}(v, u | y) - \int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\pi^y(v, u) \right] d\eta_{\mathcal{Y}}(y) \geq 0, \end{aligned}$$

where the last inequality follows from the fact that $\pi^y \in \Pi(\eta(v | y), \nu(u | y))$ is optimal for each y and so $\int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\tilde{\pi}(v, u | y) - \int_{\mathcal{V} \times \mathcal{U}} c^y(v, u) d\pi^y(v, u) \geq 0$. This establishes the optimality of π^* . We can further take $\tilde{\pi}^* = \pi^y$ to conclude the proof of (i), and (ii) is satisfied by construction. $\square$

Remark 1.20. The optimality of $\tilde{\pi}^*(v, u | y)$ in eq. (1.16) in combination with the essential uniqueness of the conditional measures from Proposition 1.10 means that whenever the cost function $c^y(v, u)$ and the conditional measures $\eta(v | z)$ satisfy conditions for uniqueness of OT solutions for all y , the conditional OT coupling is also unique up to measure zero sets. In particular, this holds in the case that $c^y(v, u) = c(v, u) = \|u - v\|_{\mathcal{U}}^p$, or more generally, $c(v, u) = h(\|u - v\|_{\mathcal{U}})$ for strictly convex h and under the assumption that $\eta(v | z)$ is regular. $\diamond$

⁶Note that the measurable selection argument is crucial here since the Kantorovich problems between the conditional measures $\eta(\cdot | y)$ and $\nu(\cdot | y)$ can have multiple minimizers and so the mapping $y \rightarrow \pi^y$ may be multi-valued.

Remark 1.21. We note that Proposition 1.19 assumes the existence of a finite cost coupling for (1.14). While this assumption may seem stringent at first sight, it is often easy to verify in practice. For example, consider the case where $\mathcal{U}$ is a separable Hilbert space, and $\mathcal{V} = \mathcal{U}$ with $c(z, v; y, u) = \|v - u\|_{\mathcal{U}}^2$, then a simple assumption that ensures the existence of a finite coupling is that $\int \|u\|_{\mathcal{U}}^2 d\eta < \infty$ and $\int \|v\|_{\mathcal{U}}^2 d\nu < \infty$ ⁷. To see this, consider the coupling $\pi(z, v, y, u) = \eta(v | y) \nu(u | y) \delta(y - z) \eta_{\mathcal{Y}}(y)$. We call this the conditional independence coupling since it is the law of the random variable (z, v, y, u) by drawing $y \sim \eta_{\mathcal{Y}}$, setting $z = y$ and then sampling $v \sim \eta(\cdot | y)$ and $u \sim \nu(\cdot | y)$ independently. We can now write, recalling that $\eta_{\mathcal{Y}} = \nu_{\mathcal{Y}}$,

$$\begin{aligned} K(\pi^*) &\leq \int_{\mathcal{Y} \times \mathcal{U} \times \mathcal{Z} \times \mathcal{V}} \|v - u\|_{\mathcal{U}}^2 d\pi(z, v, y, u) \\ &\leq \int_{\mathcal{Y} \times \mathcal{U} \times \mathcal{Y} \times \mathcal{V}} (\|v\|_{\mathcal{U}}^2 + \|u\|_{\mathcal{U}}^2) \eta(dv | y) \nu(du | y) d\delta(z - y) d\eta_{\mathcal{Y}}(y). \\ &= \int_{\mathcal{Y} \times \mathcal{U}} \|v\|_{\mathcal{U}}^2 d\eta(z, v) + \int_{\mathcal{Y} \times \mathcal{U}} \|u\|_{\mathcal{U}}^2 d\nu(y, u) < \infty. \end{aligned}$$

We also note that the above argument can be extended to broader classes of costs that satisfy $c(z, v; y, u) \leq C(c(z, v; z, v') + c(y, u; y, u'))$, for some constant $C \geq 0$ and fixed points $v', u' \in \mathcal{V} \times \mathcal{U}$. Then if $\int c(z, v; z, v') d\eta(z, v) < \infty$ and $\int c(y, u; y, u') d\nu(y, u) < \infty$ then the finite cost assumption can be verified. $\diamond$

1.3.2 The conditional dual problem

We now turn our attention to the dual problem of (1.14) and show that a strong duality result, akin to the standard result [34, Thm. 5.10] holds; in fact, our proof relies heavily on the standard setting after considering the unconstrained formulation (1.15).

Proposition 1.22. *Let $\mathcal{V}, \mathcal{U}, \mathcal{Y}$ be Polish spaces and suppose Assumption 1.17 holds. Consider a cost function $c : (\mathcal{Y} \times \mathcal{V}) \times (\mathcal{Y} \times \mathcal{U}) \rightarrow \mathbb{R}$ that is lower semicontinuous and $\inf c > -\infty$.⁸ Then it holds that*

$$\begin{aligned} \min_{\pi \in \Pi_{\mathcal{Y}}(\eta, \nu)} K(\pi) &= \sup_{\substack{(\psi, \phi) \in L^1(\eta) \times L^1(\nu) \\ \phi - \psi \leq c_{\chi}}} \left(\int_{\mathcal{Y} \times \mathcal{U}} \phi(y, u) d\nu(y, u) - \int_{\mathcal{Y} \times \mathcal{V}} \psi(z, v) d\eta(z, v) \right), \\ &= \sup_{\psi \in L^1(\eta)} \left(\int_{\mathcal{Y} \times \mathcal{U}} \psi_v^c(z, u) d\nu(z, u) - \int_{\mathcal{Y} \times \mathcal{V}} \psi(z, v) d\eta(z, v) \right), \\ &= \sup_{\phi \in L^1(\nu)} \left(\int_{\mathcal{Y} \times \mathcal{U}} \phi(y, u) d\nu(y, u) - \int_{\mathcal{Y} \times \mathcal{V}} \phi_u^c(y, v) d\eta(y, v) \right), \end{aligned}$$

where the partial c -transforms of ψ and ϕ are defined as

$$\psi_v^c(z, u) := \inf_v \left(\psi(z, v) + c(z, v; z, u) \right), \quad \phi_u^c(y, v) := \sup_u \left(\phi(y, u) - c(y, v; y, u) \right).$$

⁷In fact, this is precisely Assumption (M) in [59].

⁸A slightly more general assumption would be the existence of upper semicontinuous functions $a \in L^1_{\eta}(\mathcal{Y} \times \mathcal{V})$ and $b \in L^1_{\nu}(\mathcal{Y} \times \mathcal{U})$ such that $c(z, v; y, u) \geq a(z, v) + b(y, u)$ for all (z, v, y, u) .

Proof. Consider the formulation (1.15). Observe that this problem, with the cost c_χ , satisfies the assumptions of [34, Thm. 5.10] for strong duality and so

$$\begin{aligned} \min_{\pi \in \Pi_{\mathcal{Y}}(\eta, \nu)} K(\pi) &= \sup_{\substack{(\psi, \phi) \in L^1(\eta) \times L^1(\nu); \\ \phi - \psi \leq \tilde{c}}} \left(\int_{\mathcal{Y} \times \mathcal{U}} \phi(z, v) d\nu(y, u) - \int_{\mathcal{Y} \times \mathcal{V}} \psi(y, u) d\eta(z, v) \right), \\ &= \sup_{\psi \in L^1(\eta)} \left(\int_{\mathcal{Y} \times \mathcal{U}} \psi^{c_\chi}(y, u) d\nu(y, u) - \int_{\mathcal{Y} \times \mathcal{V}} \psi(z, v) d\eta(z, v) \right), \\ &= \sup_{\phi \in L^1(\nu)} \left(\int_{\mathcal{Y} \times \mathcal{U}} \phi(y, u) d\nu(y, u) - \int_{\mathcal{Y} \times \mathcal{V}} \phi^{c_\chi}(z, v) d\eta(z, v) \right), \end{aligned} \quad (1.17)$$

where $\psi^{c_\chi}(y, u) := \inf_{(z, v)} (\psi(z, v) + c_\chi(z, v; y, u))$ and similarly, $\phi^{c_\chi}(z, v) := \sup_{(y, u)} (\phi(y, u) - c_\chi(z, v; y, u))$. Since c_χ is real valued when restricted to $\{y = z\}$ and infinite valued when $\{y \neq z\}$, we see that when taking the infimum in the definition of ψ^{c_χ} , we may restrict to $\{y = z\}$ and so

$$\psi^{c_\chi}(y, u) = \inf_{z, v} (\psi(z, v) + c_\chi(z, v; z, u)) = \inf_v (\psi(z, u) + c(z, v; z, u)) = \psi_v^c(z, v)$$

and a similar calculation can be repeated to show $\phi^{c_\chi} = \phi_u^c$. Substituting back into (1.17) completes the proof. $\square$

Remark 1.23. In the case that c is a metric we recover an analogous version of the Kantorovich-Rubinstein duality [34, Particular Case 5.16], since in this case c_χ is also a metric as the sum of the metric c and an ∞ -valued discrete metric on the y and z variables. Therefore,

$$\min_{\pi \in \Pi_{\mathcal{Y}}(\eta, \nu)} K(\pi) = \sup_{\psi \in L^1(\eta) \cap \text{Lip}_{c_\chi}^1} \left(\int \psi(z, v) d\eta(z, v) - \int \psi(y, u) d\nu(y, u) \right),$$

where $\text{Lip}_{c_\chi}^1$ is the space of 1-Lipschitz functions with respect to c_χ . $\diamond$

1.3.3 Solvability of the conditional Monge problem

We dedicate this section to establishing the solvability of the conditional Monge problem 1.3. Our main result in this direction establishes the existence of unique solutions for this problem under structural assumptions on the cost and regularity assumptions on the reference measure η . We state the result for simple families of costs c that are often used in practice (see, for example [59, 71–74]) and provide a remark later on how the result can be generalized to broader choices. Having started with the conditional Kantorovich problem previously, we have already handled issues of measurable selection, which allows us to simplify the proof compared to [59].

Proposition 1.24. *Let $\mathcal{U}$ be a separable Hilbert space, take $\mathcal{V} = \mathcal{U}$, and let $\mathcal{Y}$ be a Polish space. Consider the cost $c(z, v; y, u) = c(v, u) = \|u - v\|_{\mathcal{U}}^p$ for $p \in (1, \infty)$. Suppose that (a) the measures η and ν satisfy Assumption 1.17 and have finite p -th moments and (b)*

the reference conditionals $\eta(\cdot \mid y)$ assign zero measure to Gaussian null sets, i.e., regular according to Definition 1.14, for η_Y -a.s. y . Then $\pi^* = (\text{Id} \times T^*)\#\eta$ is the unique solution of problem 1.4 where T^* uniquely solves problem 1.3.

Proof. Let π^* be a solution to problem 1.4, which exists following Proposition 1.19 and has finite cost by Remark 1.21. Disintegrating $K(\pi^*)$ we can write

$$\begin{aligned} \inf_{\pi} \int_{Y \times U \times Y \times U} c(v, u) d\pi(z, v, y, u) &= \int_{Y \times U \times Y \times U} c(v, u) d\pi^*(z, v, y, u), \\ \int_{Y \times Y} \left[\int_{U \times U} c(v, u) d\pi^*(v, u \mid z, y) \right] d\delta(z - y) d\eta_Y(z) &= \int_Y \left[\int_{U \times U} c(v, u) d\pi^y(v, u) \right] d\eta_Y(y). \end{aligned}$$

where $\pi^y(v, u)$ is an optimal coupling of $\eta(v \mid y)$ and $\nu(u \mid y)$ per the notation in the proof of Proposition 1.19. By proposition 1.15 the coupling π^y is induced by a unique OT map, i.e., $\pi^y = (\text{Id} \times T_y)\#\eta(\cdot \mid y)$. We now wish to argue that the maps T_y can be concatenated together in such a way that the solution to problem 1.3 is given by $T^*(z, v) = (z, T_U^*(z, v))$ where $T_U^*(z, v) = T_z(v)$ η -a.s. In order to see that this procedure results in a measurable map, we may extract a map through the barycentric projection, setting $T_z(v) = \int_U \pi^z(u, v) du = \int_U \tilde{\pi}^*(u, v \mid z) du$. This yields the joint measurability of $T_z(v)$ in z and v by Fubini's theorem, which implies that our candidate map $T^*(z, v)$ is measurable. The optimality of T^* follows from the cost being equivalent to that of π^* , the solution to the Kantorovich problem, which lower bounds the optimal Monge cost, yielding optimality. Furthermore, the maps $T_U^*(z, \cdot)$ are essentially unique due to remark 1.20. This, in turn, implies the essential uniqueness of the map T^* . $\square$

Remark 1.25. Note that our argument for the measurability of T^* is brief and simple (for example, in comparison to [59]). This is due to the fact that we have already dealt with measurable selection issues in defining the solution to the Kantorovich problem. Since the solution to that problem was obtained by concatenation of couplings between the conditionals, we were able to simply integrate those conditionals for fixed z to define the maps T_y . $\diamond$

Remark 1.26. We highlight that our method of proof for Proposition 1.24 mainly relies on Proposition 1.19 and a solvability result for the Monge problems on the conditionals $\eta(\cdot \mid y)$ and $\nu(\cdot \mid y)$. This implies that the result can be generalized to broader choices of c for which we can still establish the solvability of the Monge problems between the conditionals. A simple example in finite dimensions would be a choice of costs of the form $c(z, v; y, u) = h(u - v)$ with h strictly convex. $\diamond$

1.3.4 Relaxing the conditional Kantorovich and Monge problems

In this section we consider a relaxation of the conditional Kantorovich problem 1.4 and Monge problem 1.3 with quadratic costs to an unconstrained problem with a singularly perturbed cost function. Our result can be viewed as the Hilbert space generalization of the main result of [64] where the authors showed that the KR map can be obtained as a limit of OT maps obtained from singularly perturbed costs. Our interest in this relaxation

stems from algorithmic considerations. For example, constructing triangular OT maps by first transporting the conditionals of η, ν is not practical in situations where we only have access to an empirical approximation to ν since in this case we may only have a single u sample per each y sample. This motivates the formulation of an OT problem between η, ν directly, where the solution to the empirical OT problem converges to that of the true solution. Motivated by these applied considerations, we present the following result:

Proposition 1.27. *Let $\mathcal{U}$ and $\mathcal{Y}$ be separable Hilbert spaces and consider the cost $c(z, v, ; y, u) = c(v, u) = \|u - v\|_{\mathcal{U}}^2$. Suppose the conditions of Proposition 1.24 hold for $p = 2$ so that the conditional Kantorovich and Monge problems have unique solutions π^* and T^* respectively. For $\epsilon > 0$, define*

$$K_\epsilon(\pi) := \int c_\epsilon d\pi, \quad c_\epsilon(z, v; y, u) := \|z - y\|_{\mathcal{Y}}^2 + \epsilon \|v - u\|_{\mathcal{U}}^2. \quad (1.18)$$

Let $\pi_\epsilon := \inf_\pi K_\epsilon(\pi)$. Then (i) the π_ϵ are unique and $\pi_\epsilon = (\text{Id} \times T_\epsilon) \# \eta$ for a unique map T_ϵ ; (ii) $\pi_\epsilon \rightharpoonup \pi^$ (weak convergence) as $\epsilon \rightarrow 0$; and (iii) T_ϵ converges in $L_\eta^2(\mathcal{Y} \times \mathcal{U}; \mathcal{Y} \times \mathcal{U})$ to the triangular map T^* defined in proposition 1.24 as $\epsilon \rightarrow 0$.*

Proof. As stated earlier, our result is closely related to the main result in [64] and follows a similar path at the beginning of the proof before deviating towards the end where we rely on Propositions 1.18 and 1.24 instead of more direct arguments. The key differences are that (1) we work with a pair of separable Hilbert spaces $\mathcal{Y}, \mathcal{U}$ while [64] considers a finite-dimensional Euclidean space which decouples into a sequence of one dimensional problems, and (2) we work under Assumption 1.17 which simplifies our proof.

Statement (i) follows directly from Proposition 1.15 by observing that $\sqrt{c_\epsilon}$ is a norm defined on $\mathcal{Y} \times \mathcal{U}$ which is a separable Hilbert space.

We now prove (ii), which consists of the main technical aspects of the proof. Let $c_\epsilon(z, v, y, u) = \|z - y\|_{\mathcal{Y}}^2 + \epsilon \|v - u\|_{\mathcal{U}}^2$ so that $K_\epsilon(\pi) = \int c_\epsilon d\pi$ and set

$$\pi_\epsilon := \arg \min_\pi K_\epsilon(\pi), \quad \pi_0 := \lim_{\epsilon \rightarrow 0} \arg \min_\pi K_\epsilon(\pi),$$

where the limit exists (up to a subsequence) by the tightness of the set of couplings between fixed marginals. [34, Lemma 4.4]. By the optimality of π_ϵ , we have that

$$\int c_\epsilon d\pi_\epsilon \leq \int c_\epsilon d\pi^*. \quad (1.19)$$

Towards taking the limit $\epsilon \rightarrow 0$ on both sides, note that the right hand side is continuous in ϵ by the hypothesis that η and ν have finite second moments (so the integral is finite) and the fact that the measure π^* is fixed. On the other hand, we can verify that $\epsilon \mapsto \min_\pi K_\epsilon(\pi)$ is continuous since,

$$\begin{aligned} \left| \int c_\epsilon d\pi_\epsilon - \int c_{\epsilon'} d\pi_{\epsilon'} \right| &= \left| \int c_\epsilon d\pi_\epsilon - \int c_{\epsilon'} d\pi_\epsilon + \int c_{\epsilon'} d\pi_\epsilon - \int c_{\epsilon'} d\pi_{\epsilon'} \right|, \\ &\leq \left| \int c_\epsilon d\pi_\epsilon - \int c_{\epsilon'} d\pi_\epsilon \right| + \left| \int c_{\epsilon'} d\pi_\epsilon - \int c_{\epsilon'} d\pi_{\epsilon'} \right|. \end{aligned} \quad (1.20)$$

Now let $\epsilon \rightarrow \epsilon'$. Once again, by the compactness of the couplings, we know that $\pi_\epsilon \rightarrow \pi_{\epsilon'}$ up to a subsequence, which means that the second term on the right hand side vanishes since the cost is fixed. Additionally, $|\int c_\epsilon d\pi_\epsilon - \int c_{\epsilon'} d\pi_{\epsilon'}| = |\epsilon - \epsilon'| \int_{\mathcal{Y} \times \mathcal{U} \times \mathcal{Y} \times \mathcal{U}} \|u - v\|_{\mathcal{U}}^2 \pi(dz, dv; dy, du) \leq C|\epsilon - \epsilon'|$ for a uniform constant $C > 0$ depending on the second moments of η, ν , proving that the first term in the right hand side of (1.20) also vanishes as $\epsilon \rightarrow \epsilon'$ allowing us to conclude that $\epsilon \mapsto \min_\pi K_\epsilon(\pi)$ is continuous. We can now take $\epsilon \rightarrow 0$ on both sides of (1.19) to obtain

$$\int c_0 d\pi_0 \leq \int c_0 d\pi^*. \quad (1.21)$$

Now consider the operator $P_{\mathcal{Y} \times \mathcal{Y}} : (z, v; y, u) \rightarrow (z, y)$ and observe that since $c_0(z, v, y, u) = \|z - y\|_{\mathcal{Y}}^2$ is independent of v, u then $P_{\mathcal{Y} \times \mathcal{Y}} \# \pi^*$ is by definition the minimizer of $\int c_0 d\pi$ as it couples z, y by the identity map and achieves zero cost. Thus, by optimality of π^* we have that $\int c_0 d\pi_0 = \int c_0 d\pi^* = 0$. In turn, this implies that $P_{\mathcal{Y} \times \mathcal{Y}} \# \pi_0 = P_{\mathcal{Y} \times \mathcal{Y}} \# \pi^* =: \pi_{\mathcal{Y} \times \mathcal{Y}}^*$ since for any coupling π we have that $\int c_0 d\pi = \int \|z - y\|_{\mathcal{Y}}^2 P_{\mathcal{Y} \times \mathcal{Y}} \# \pi(dz, dy)$ which is precisely the standard OT problem between the $\mathcal{Y}$ marginals of η and ν and has a unique solution.

Let us further define $P_{\mathcal{U} \times \mathcal{U}} : (z, v; y, u) \mapsto (v, u)$. By the above observation and the fact that π_ϵ is optimal for c_ϵ we can write,

$$\begin{aligned} \epsilon \int \|u - v\|_{\mathcal{U}}^2 d\pi_\epsilon(z, v; y, u) &= \int \|z - y\|_{\mathcal{Y}}^2 d\pi_{\mathcal{Y} \times \mathcal{Y}}^*(z, y) + \epsilon \int \|u - v\|_{\mathcal{U}}^2 dP_{\mathcal{U} \times \mathcal{U}} \# \pi_\epsilon(v, u) \\ &\leq \int \|z - y\|_{\mathcal{Y}}^2 + \epsilon \int \|u - v\|_{\mathcal{U}}^2 d\pi_\epsilon(z, v; y, u) \leq \int \|z - y\|_{\mathcal{Y}}^2 + \epsilon \int \|u - v\|_{\mathcal{U}}^2 d\pi^*(z, v; y, u) \\ &= \epsilon \int \|u - v\|_{\mathcal{U}}^2 d\pi^*(z, v; y, u). \end{aligned}$$

Hence, $\int \|u - v\|_{\mathcal{U}}^2 d\pi_\epsilon \leq \int \|u - v\|_{\mathcal{U}}^2 d\pi^*$ for $\epsilon > 0$ and so passing to the limit, we get

$$\int \|u - v\|_{\mathcal{U}}^2 d\pi_0 \leq \int \|u - v\|_{\mathcal{U}}^2 d\pi^*. \quad (1.22)$$

Notice that since $P_{\mathcal{Y} \times \mathcal{Y}} \# \pi_0 = \pi_{\mathcal{Y} \times \mathcal{Y}}^* = (\text{Id} \times \text{Id}) \# \eta_{\mathcal{Y}}$, then $\pi_0 \in \Pi_{\mathcal{Y}}$ and, by Proposition 1.18, can be decomposed as $\pi_0(z, v; y, u) = \tilde{\pi}_0(v, u \mid y) \delta(z - y) \eta_{\mathcal{Y}}(y)$ where $\tilde{\pi}_0 \in \Pi(\eta(\cdot \mid y), \nu(\cdot \mid y))$. This fact together with (1.22) results in the inequality

$$\int_{\mathcal{Y}} \left[\int_{\mathcal{U} \times \mathcal{U}} \|u - v\|_{\mathcal{U}}^2 d\tilde{\pi}_0(u, v \mid y) \right] d\eta_{\mathcal{Y}}(y) \leq \int_{\mathcal{Y}} \left[\int_{\mathcal{U} \times \mathcal{U}} \|u - v\|_{\mathcal{U}}^2 d\tilde{\pi}^*(u, v \mid y) \right] d\eta_{\mathcal{Y}}(y).$$

Furthermore, the $\tilde{\pi}^*(\cdot, \cdot \mid y)$ are the unique optimal couplings of $\eta(\cdot \mid y)$ and $\nu(\cdot \mid y)$ for a.s. y , which implies that $\tilde{\pi}_0(\cdot, \cdot \mid y)$ should also be optimal for a.s. y and so $\tilde{\pi}_0 = \tilde{\pi}^*$ which completes the proof of the fact that $\pi_0 = \pi^*$. Finally, (iii) follows immediately from lemma 1.16. $\square$

Remark 1.28. We note that while Proposition 1.27 is stated for the quadratic cost c_ϵ , the argument can be extended, with adjustment to the proof of Proposition 1.15, to more general costs of the form $c_\epsilon(z, v, y, u) = \|z - y\|_{\mathcal{Y}}^p + \epsilon \|v - u\|_{\mathcal{U}}^q$ for exponents $p, q \in (1, +\infty)$. In fact, our proof of (ii) remains unchanged. One simply needs to verify the existence and uniqueness of the π_ϵ in (i) and generalize Lemma 1.16 to L^p spaces for $p \in (1, +\infty)$ to generalize (iii). $\diamond$

1.3.5 Extension to general triangular couplings and maps

Up to this point, we have worked under assumption 1.17, which has helped simplify many of our arguments. Because the $\mathcal{Y}$ marginals of η and ν match, we can simply take the $T_{\mathcal{Y}}$ component of our triangular maps to be Id . While this assumption is sufficient for many practical applications, such as the BIPs considered in Section 1.4, there are still important problems where Assumption 1.17 does not hold, and we may need to work with the broader family of triangular transport maps. In this section we present a generalization of our results to such a setting where $T_{\mathcal{Y}}$ itself is replaced with an OT map. Our result can be viewed as a generalization of KR rearrangements to the case of Hilbert spaces. We present our result as a corollary since the proof follows directly from our previous results as well as standard results on the uniqueness of Monge maps.

Corollary 1.29. *Let $\mathcal{Y}, \mathcal{U}$ be separable Hilbert spaces. Consider the cost functions Suppose that the reference marginal $\eta_{\mathcal{Y}}$ and the reference conditionals $\eta(\cdot | y)$ are regular and η, ν have finite second moments Then it holds that*

(i) *There exists a unique map $T_{\mathcal{Y}}^* : \mathcal{Y} \rightarrow \mathcal{Y}$ that solves*

$$\arg \min_{T_{\mathcal{Y}}} \int_{\mathcal{Y}} \|z - T_{\mathcal{Y}}(z)\|_{\mathcal{Y}}^2 d\eta(z, v), \quad \text{s. t.} \quad T_{\mathcal{Y}} \# \eta_{\mathcal{Y}} = \nu_{\mathcal{Y}}.$$

(ii) *There exists a unique map $T_{\mathcal{U}}^* : \mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{U}$ that solves*

$$\arg \min_{T_{\mathcal{U}}} \int_{\mathcal{Y}} \|v - T_{\mathcal{U}}(z, v)\|_{\mathcal{U}}^2 d\eta(z, v) \\ \text{s. t.} \quad T(z, v) = (z, T_{\mathcal{U}}(z, v)), \quad T \circ (T_{\mathcal{Y}}^* \times \text{Id}) \# \eta = \nu.$$

(iii) *Define the map $T^* : \mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{Y} \times \mathcal{U}$ as $T^* := (\text{Id} \times T_{\mathcal{U}}^*) \circ (T_{\mathcal{Y}}^* \times \text{Id})$. If $T_{\mathcal{Y}}^*$ is invertible,⁹ then it holds that $T^* = \lim_{\epsilon \rightarrow 0} T_{\epsilon}$ in $L^2(\eta)$ where T_{ϵ} are the maps defined in Proposition 1.27.*

Proof. Statement (i) is a direct consequence of proposition 1.15 under our hypothesis on $\eta_{\mathcal{Y}}$. To prove (ii) consider $\bar{\eta} = (T_{\mathcal{Y}}^* \times \text{Id}) \# \eta$ and note that $\bar{\eta}_{\mathcal{Y}} = \nu_{\mathcal{Y}}$. Then the result follows from Proposition 1.24 as soon as we verify that $\bar{\eta}$ satisfies the conditions of that theorem, in particular that $\bar{\eta}(\cdot | y)$ are regular for all y . To see this, let $A \subset \mathcal{B}(\mathcal{U})$ be a $\eta(\cdot | z)$ null set for all z , that is, $\eta(A | z) = 0$. Then $\bar{\eta}(A | y) = \int_{\mathcal{Y}} \eta(A | z) \delta(T_{\mathcal{Y}}^*(z) - y) d\eta_{\mathcal{Y}}(z) = 0$. To prove (iii), we need to slightly generalize the proof of Proposition 1.27. In fact, the proof is identical up to (1.21) by taking π^* to be the coupling induced by the optimal map T^* , which achieves zero cost for c_0 . Then we repeat the same arguments until the paragraph after (1.22) where we use Proposition 1.18 with a nontrivial and invertible map $T_{\mathcal{Y}}^*$. $\square$

Remark 1.30. We note that this result can be further generalized to the setting where the costs in (i) and (ii) are replaced with $\|z - T_{\mathcal{Y}}(z)\|_{\mathcal{Y}}^p$ and $\|v - T_{\mathcal{U}}(z, v)\|_{\mathcal{U}}^q$ for $p, q \in (1, +\infty)$ respectively. Then we need $\eta_{\mathcal{Y}}$ and $\eta(\cdot | y)$ to have finite p and q moments respectively. $\diamond$

⁹This will hold, for example, if $\nu_{\mathcal{Y}}$ is also regular by Proposition 1.15(ii).

1.4 Application to BIPs

In this section we outline the application of our results from Section 1.3 to the case of BIPs, giving a likelihood-free characterization of posterior measures. Consider separable Hilbert spaces $\mathcal{U}, \mathcal{Y}$ and a forward map $F : \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{Y}$ with a model

$$y = F(u, \xi), \quad \xi \sim \gamma \in \mathbb{P}(\mathcal{Y}),$$

that relates a parameter u to data y with possible observation noise ξ . The goal of Bayesian inference is to estimate the parameter u given an instance of the data y . To achieve this, consider a prior measure $\mu \in \mathbb{P}(\mathcal{U})$, reflecting our prior belief about u , and define the joint measure

$$\nu = \text{Law}\{(y, u) : \text{draw } u \sim \mu, \xi \sim \gamma \text{ and set } y = F(u, \xi)\} \in \mathbb{P}(\mathcal{Y} \times \mathcal{U}). \quad (1.23)$$

Then the conditional $\nu(\cdot \mid y)$ is precisely the Bayesian posterior measure that updates our prior belief about u given the observed data y . The standard BIP framework proceeds to identify the posterior via Bayes' rule (1.7) by its Radon-Nikodym derivative where the likelihood function $\exp(-\Phi(u; y))$ is proportional to the density of $\nu(y \mid u)$ with respect to γ , i.e., $\exp(-\Phi(u; y)) \propto \frac{d\nu(\cdot \mid u)}{d\gamma}(y)$; see [6, Thm. 6.29] for details.

While MCMC approaches [7, 10, 13] for sampling the posterior measure rely on the explicit form of the likelihood potential Φ in conjunction with the choice of the prior μ , here we aim to characterize the posterior $\nu(\cdot \mid y)$ using conditional OT maps that circumvent the explicit form of the likelihood, thereby yielding a likelihood-free inference method that does not even require absolute continuity of the posterior with respect to the prior [53, 123–125]. To this end, we consider the reference measure $\eta = \nu_{\mathcal{Y}} \otimes \mu$, i.e., the independence coupling between the $\mathcal{Y}$ -marginal of ν and the prior μ . This choice of the reference η is natural in the setting where F is known, and samples from μ can be generated cheaply, since we have that $\nu_{\mathcal{Y}} = F\#(\mu \otimes \gamma)$ and so reference samples can be generated at the cost of evaluating the forward map. We can then obtain the following corollary to Proposition 1.24.

Corollary 1.31. *Suppose $\mathcal{U}, \mathcal{Y}$ are separable Hilbert spaces, the prior $\mu \in \mathbb{P}(\mathcal{U})$ is regular according to Definition 1.14, $\gamma \in \mathbb{P}(\mathcal{Y})$, and F is continuous. For $p \in (1, \infty)$ consider the conditional Monge problem*

$$\inf_T \int_{\mathcal{Y} \times \mathcal{U}} \|v - T_{\mathcal{U}}(z, v)\|_{\mathcal{U}}^p d\eta(z, v) \quad \text{s.t.} \quad T\#\eta = \nu \text{ and } T(z, v) = (z, T_{\mathcal{U}}(z, v)). \quad (1.24)$$

If the infimum is finite and feasible, then the solution $T_{\mathcal{U}}^$ satisfies $T_{\mathcal{U}}^*(y, \cdot)\#\mu = \nu(\cdot \mid y)$, $\nu_{\mathcal{Y}}$ -a.s.*

Proof. The continuity of F ensures that ν is a well-defined Radon measure by [57, Thm. 9.1.1]. Then, the corollary follows directly from Proposition 1.24 with the cost $c(z, v; y, u) = \|v - u\|_{\mathcal{U}}^p$ and by observing that $\eta(\cdot \mid y) = \mu$ which is assumed regular. $\square$

We interpret Corollary 1.31 as a characterization of a family of prior-to-posterior maps $T_{\mathcal{U}}^*(y, \cdot)$ parameterized by the data y . We highlight a few important points: first, our choice

of the reference η is non-standard in comparison to other conditional generative models in the literature such as [53, 92] where $\eta = \nu_{\mathcal{Y}} \otimes \eta_{\mathcal{U}}$ with $\eta_{\mathcal{U}}$ taken to be standard Gaussian while here we take $\eta_{\mathcal{U}} = \mu$. Second, μ is completely arbitrary, so the transport-based characterization of posteriors naturally extends to non-Gaussian priors which is contemporary in the context of infinite dimensional MCMC [13, 15, 126].

1.4.1 Finite-dimensional approximations and perturbation analysis

The use of conditional transport maps for practical infinite dimensional inference requires the inevitable discretization of the problem to achieve a finite-dimensional approximation, and so we turn our attention to such approximations of both the parameter u and the data y . Letting $\mathcal{U}^N$ denote an N -dimensional linear subspace of $\mathcal{U}$ we consider projection maps $P_{\mathcal{U}}^N : \mathcal{U} \rightarrow \mathcal{U}^N$ representing an appropriate discretization of the function space $\mathcal{U}$. In our numerical experiments, we project onto finitely many modes from the Karhunen-Lo  ve (KL) expansion of a Gaussian process prior [121, Thm. 3.5.1 and Thm. 3.5.10]; i.e., the eigenfunctions of the covariance operator. Writing $\mu^N := P_{\mathcal{U}}^N \# \mu$ we can now define the approximate target measure

$$\nu^N := \text{Law} \{ (y, u) : \text{draw } u \sim \mu^N, \xi \sim \gamma \text{ and set } y = F(u, \xi) \} \in \mathbb{P}(\mathcal{Y} \times \mathcal{U}),$$

with the associated posterior measures

$$\frac{d\nu^N(\cdot | y)}{d\mu^N}(u) = \frac{1}{Z^N(y)} \exp(-\Phi(u; y)), \quad Z^N(y) = \int_{\mathcal{U}} \exp(-\Phi(u; y)) d\mu^N(u).$$

The above measures are concentrated on the finite-dimensional subspace $\mathcal{U}^N$ and hence amenable to numerical computations. Posing (1.24) for the measure ν^N we obtain the discretized conditional Monge problem

$$\inf_T \int_{\mathcal{Y} \times \mathcal{U}} \|v - T_{\mathcal{U}}(z, v)\|_{\mathcal{U}}^p d\eta(z, v) \quad \text{s.t.} \quad T \# \eta = \nu^N \text{ and } T(z, v) = (z, T_{\mathcal{U}}(z, v)), \quad (1.25)$$

with solution T^N . Since $T_{\mathcal{U}}^N(y, \cdot) \# \mu = \nu^N(\cdot | y)$ then we trivially obtain the identity

$$D(T_{\mathcal{U}}^N(y, \cdot) \# \mu, T_{\mathcal{U}}^*(y, \cdot) \# \mu) = D(\nu^N, \nu),$$

for any statistical divergence or metric D for a.s. y . Controlling the right hand side of this expression in terms of the operator norm of $I - P_{\mathcal{U}}^N$ has been the subject of extensive study in the perturbation analysis of BIPs [127, 128]. We recall a simple version of such a result following [128, Sec. 4.1, Cor. 3.6, and Thm. 3.8]:

Lemma 1.32. *Suppose Φ is positive, bounded, and bi-Lipschitz, and μ has bounded moments of degree 2. Then there exist constants $C_1, C_2 > 0$ so that*

$$\begin{aligned} W_1(\nu^N(\cdot | y), \nu(\cdot | y)) &\leq C_1 W_2(\mu, \mu^N), \quad \forall N \geq 0, \\ W_1(\nu(\cdot | y), \nu(\cdot | y')) &\leq C_2 \|y - y'\|_{\mathcal{Y}}, \quad \forall y, y' \in \mathcal{Y}. \end{aligned}$$

Using the coupling $(\text{Id} \times P_{\mathcal{U}}^N) \# \mu$ we can write $W_2^2(\mu, \mu^N) = \inf_{\pi} \int \|u - v\|_{\mathcal{U}}^2 d\pi(u, v) \leq \|\text{Id} - P_{\mathcal{U}}^N\|_{L^2(\mu)}^2$ leading to the error bound

$$W_1(T_{\mathcal{U}}^N(y, \cdot) \# \mu, T_{\mathcal{U}}^*(y, \cdot) \# \mu) \leq C \|\text{Id} - P_{\mathcal{U}}^N\|_{L^2(\mu)}.$$

This error bound is sufficient from the perspective of sampling since it quantifies the quality of samples coming from $T_{\mathcal{U}}^N$ as compared to the ground truth map $T_{\mathcal{U}}^*$. However, it is also interesting to ask whether $T_{\mathcal{U}}^N$ is also a good approximation to $T_{\mathcal{U}}^*$. Such a result can be obtained in the finite-dimensional setting and for Brenier maps thanks to stability results for Wasserstein distances under perturbations [60, 129, 130]. We state such a result below.

Proposition 1.33. *Take $\mathcal{U} = \mathbb{R}^d$ and $c(z, v; y, u) = \|u - v\|_2^2$ and suppose that $\gamma \in \mathbb{P}(\mathcal{Y})$ and F is continuous. Further, suppose that: (a) the prior μ is supported on a compact and convex set $A \subset \mathbb{R}^d$ and absolutely continuous with respect to the Lebesgue measure with a density that is bounded from above and below on A ; and (b) the potential Φ is positive, bounded, and bi-Lipschitz. Then for $\nu_{\mathcal{Y}}$ -a.s. y, y' and for $N \leq d$, it holds that*

$$\begin{aligned} \|T_{\mathcal{U}}^N(y, \cdot) - T_{\mathcal{U}}^*(y, \cdot)\|_{L^2(\mu)} &\leq C_1 \|\text{Id} - P_{\mathcal{U}}^N\|_{L^2(\mu)}^{1/6}, \\ \|T_{\mathcal{U}}^*(y, \cdot) - T_{\mathcal{U}}^*(y', \cdot)\|_{L^2(\mu)} &\leq C_2 \|y - y'\|_{\mathcal{Y}}^{1/6}. \end{aligned} \tag{1.26}$$

Here the constants $C_1, C_2 \geq 0$ depend on d, A, Φ and the choice of μ .

Proof. Hypothesis (b) ensures that the posterior measures $\nu(\cdot | y)$ and $\nu^N(\cdot | y)$ are well-defined. The choice of c implies that the maps $T_{\mathcal{U}}^*(y, \cdot)$ and $T_{\mathcal{U}}^N(y, \cdot)$ are precisely the Brenier prior-to-posterior maps for each y . Since μ is assumed to be supported on A then the posterior measures $\nu(\cdot | y)$ are also supported on the same set, and the same holds for μ^N and the corresponding $\nu^N(\cdot | y)$. This, together with hypothesis (a), allows us to use [60, Thm. 4.2] to get the bounds

$$\begin{aligned} \|T_{\mathcal{U}}^N(y, \cdot) - T_{\mathcal{U}}^*(y, \cdot)\|_{L^2(\mu)} &\leq C_1 W_1(T_{\mathcal{U}}^N(y, \cdot) \# \mu, T_{\mathcal{U}}^*(y, \cdot) \# \mu)^{1/6} \\ \|T_{\mathcal{U}}^*(y, \cdot) - T_{\mathcal{U}}^*(y', \cdot)\|_{L^2(\mu)} &\leq C_2 W_1(T_{\mathcal{U}}^*(y, \cdot) \# \mu, T_{\mathcal{U}}^*(y', \cdot) \# \mu)^{1/6}. \end{aligned} \tag{1.27}$$

The desired result now follows from Lemma 1.32. $\square$

Before the end of this section, we make a few remarks regarding the implications and possible generalizations of the above result. We emphasize that the main hurdle in obtaining analogous error bounds for $T_{\mathcal{U}}^*$ in the setting of infinite dimensional $\mathcal{U}$ is that, to our knowledge, analogous stability results to (1.27) have not been established in that setting. In fact, any divergence D for which Lemma 1.32 can be established and satisfies a stability result of the form $\|T_1 - T_2\|_{L^2(\mu)} \leq CD(T_1 \# \mu, T_2 \# \mu)^\alpha$ for OT maps T_1, T_2 and constants $C, \alpha \geq 0$ can be used to establish a result similar to Proposition 1.33.

The second display in (1.26) implies the Lipschitz continuity of $T_{\mathcal{U}}^*(z, v)$ along the z coordinate which has important implications when it comes to the numerical approximations of this map where one may choose to parameterize it using polynomials, kernels, or neural nets. Obtaining quantitative error bounds for such parameterizations will depend

on establishing the regularity of $T_{\mathcal{U}}^N$ on the space $\mathcal{Y} \times \mathcal{U}$ inherited from $T_{\mathcal{U}}^*$; see for example [106].

Finally, we note that our error analysis is done entirely in the *population* regime where we assume that the measures μ, ν are known. In likelihood-free inference applications, such as our experiments in Section 1.5, we only have access to an empirical approximation to ν arising from a finite collection of samples $\{y_j, u_j\}_{j=1}^J \sim \nu$ in which case, the statistical consistency and sample complexity of the underlying maps should be established. This is an exciting and contemporary direction for future research that we do not address here.

1.5 Numerical Experiments

We dedicate this section to numerical experiments that demonstrate how our theoretical analysis throughout the paper can lead to algorithms for simulation based inference. We emphasize that our focus is not on algorithmic development and performance but rather on proof of concept. In Section 1.5.1, we outline two algorithms that we used for conditional simulation, one based directly solving empirical OT problems and another using a modified GAN. Section 1.5.2 presents our experiments involving two dimensional benchmark data sets while Section 1.5.3 outlines a set of experiments for the Darcy flow inverse problem introduced as a motivating example in the introduction. Throughout the section, we give a brief account of the setup of our experiments and algorithms, but further details and hyper-parameter choices can be found in our online repository.¹⁰

1.5.1 Summary of algorithms

We focus on the particular setting where we only have access to empirical samples $\{(y_j, u_j)\}_{i=1}^J \sim \nu$ and write ν^J to denote corresponding empirical measures. Similarly, we write η^J to denote the empirical measure of the samples $\{(y_j, v_j)\}_{j=1}^J \sim \eta$ where the y_j are copies of the y -coordinates of the target samples while the $v_j \sim \eta_{\mathcal{U}}$ are drawn from an arbitrary reference. This particular construction of the reference and target measures ensures that η^J and ν^J exactly satisfy Assumption 1.17. In both algorithms the output is a map $\hat{T}_{\mathcal{U}}$ which approximates the OT conditioning map $T_{\mathcal{U}}^*$ from Section 1.3. Given this approximation, we can generate an arbitrary number of conditional samples simply by drawing new reference samples $v \sim \eta_{\mathcal{U}}$ (assumed to be accessible) and evaluating $\hat{T}_{\mathcal{U}}(y, v)$ to get an approximate sample from $\nu(\cdot | y)$.

Our first algorithm follows from proposition 1.27 and the observation that the maps T_{ϵ} approximate the triangular map T^* . Given the empirical measures η^J, ν^J , we formulate the OT problem (1.18) for a small value of ϵ . Since we choose the reference and target measures to have the same number of samples, the solution to the discrete OT problem defines a map on the sample set. We then extend this map to the entire support of η by a simple interpolation strategy. The resulting algorithm is summarized in Algorithm 1. This algorithm is very simple to implement as it only requires solving a linear program and interpolating the result. In our experiments, we used a distance weighted two nearest neighbor interpolant. We observed that this simple approach worked well in low-dimensional problems such as

¹⁰<https://github.com/TADSGroup/ConditionalOT2023>

the toy examples in Section 1.5.2, but it suffers from the curse of dimensionality in high dimensional problems and is unable to take advantage of the smoothness of transport maps.

Algorithm 1 Plugin Conditional OT

- 1: **Input:** Samples $\{(y_j, u_j)\}_{j=1}^N \sim \nu$ and relaxation parameter $\epsilon > 0$
 - 2: **Output:** Conditioning map $\hat{T}_{\mathcal{U}} : \mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{U}$
 - 3: Draw J samples $\{v_j\}_{j=1}^J \sim \eta_{\mathcal{U}}$ and form empirical measures $\eta^J, \nu^J \in \mathbb{P}(\mathcal{U} \times \mathcal{V})$.
 - 4: Solve the empirical OT problem defined by cost (1.18) between η^J and ν^J to obtain a discrete map.
 - 5: Construct $\hat{T}_{\mathcal{U}}$ to satisfy the interpolation constraints $\hat{T}_{\mathcal{U}}(y_j, v_j) = u_j$ for $j = 1, \dots, J$.
-

Our second algorithm follows Proposition 1.24 and aims to approximate the map $T_{\mathcal{U}}^*$ by a relaxation of the conditional Monge problem. The idea here is to parameterize $T_{\mathcal{U}}^*$ using an expressive function approximator, a neural net in our case, and solve a problem of the form

$$\hat{T}_{\mathcal{U}} = \arg \min_{T_{\mathcal{U}} \in \mathcal{T}} \int_{\mathcal{Y} \times \mathcal{U}} \|v - T_{\mathcal{U}}(z, v)\|_{\mathcal{U}}^2 d\eta(z, v) + \frac{1}{\lambda} D(T\#\eta^J, \nu^J), \text{ s.t. } T(z, v) = (z, T_{\mathcal{U}}(z, v)),$$

where $\mathcal{T}$ is a parametric family of functions from $\mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{U}$, D is a divergence or discrepancy that distinguishes pairs of measures in $\mathbb{P}(\mathcal{Y} \times \mathcal{U})$, and $\lambda > 0$ is a regularization parameter. Simply put, the above approach amounts to solving the conditional Monge problem Equation (1.8) for the empirical measures and by relaxing the constraint $T\#\eta = \nu$ to simply minimize a divergence between $T\#\eta$ and ν . We implement an instance of this approach with D taken to be a Wasserstein GAN-GP loss [37, 38] and letting $\mathcal{T}$ be a class of neural operators with linear projection and reconstruction maps in the input and output layers in the setting of function space data following the setup of [53, 131]. Indeed, our algorithm is a modification of the Monotone GANs of [53] with the exception that we minimize the Monge loss while Monotone GANs enforce monotonicity of the transport map $\hat{T}_{\mathcal{U}}$ using a penalty term.

1.5.2 Two dimensional benchmarks

Here we consider a set of two dimensional benchmark examples from [132] outlined in Figure 1.2. Our goal is to sample vertical slices of the target measure $\nu \in \mathbb{P}([-1, 1]^2)$. We use $J = 20000$ target samples and take $\eta_{\mathcal{U}} = N(0, 1)$. The results presented in Figure 1.2 were obtained using Algorithm 1 with $\epsilon = 5 \times 10^{-3}$ and two nearest neighbor interpolation to extend the map to the entire unit square. Another example of these results was presented in Figure 1.1.

We generally observe good agreement between the numerical and true conditional histograms across these examples, with larger errors when the slices pass through a low density regions. We also observe that our maps $\hat{T}_{\mathcal{U}}(y, \cdot)$ are noisy due to the two nearest neighbor interpolation. One can obtain a smoother map by using a different interpolant or using more neighboring points, but this leads to further smoothing of the conditional histograms.

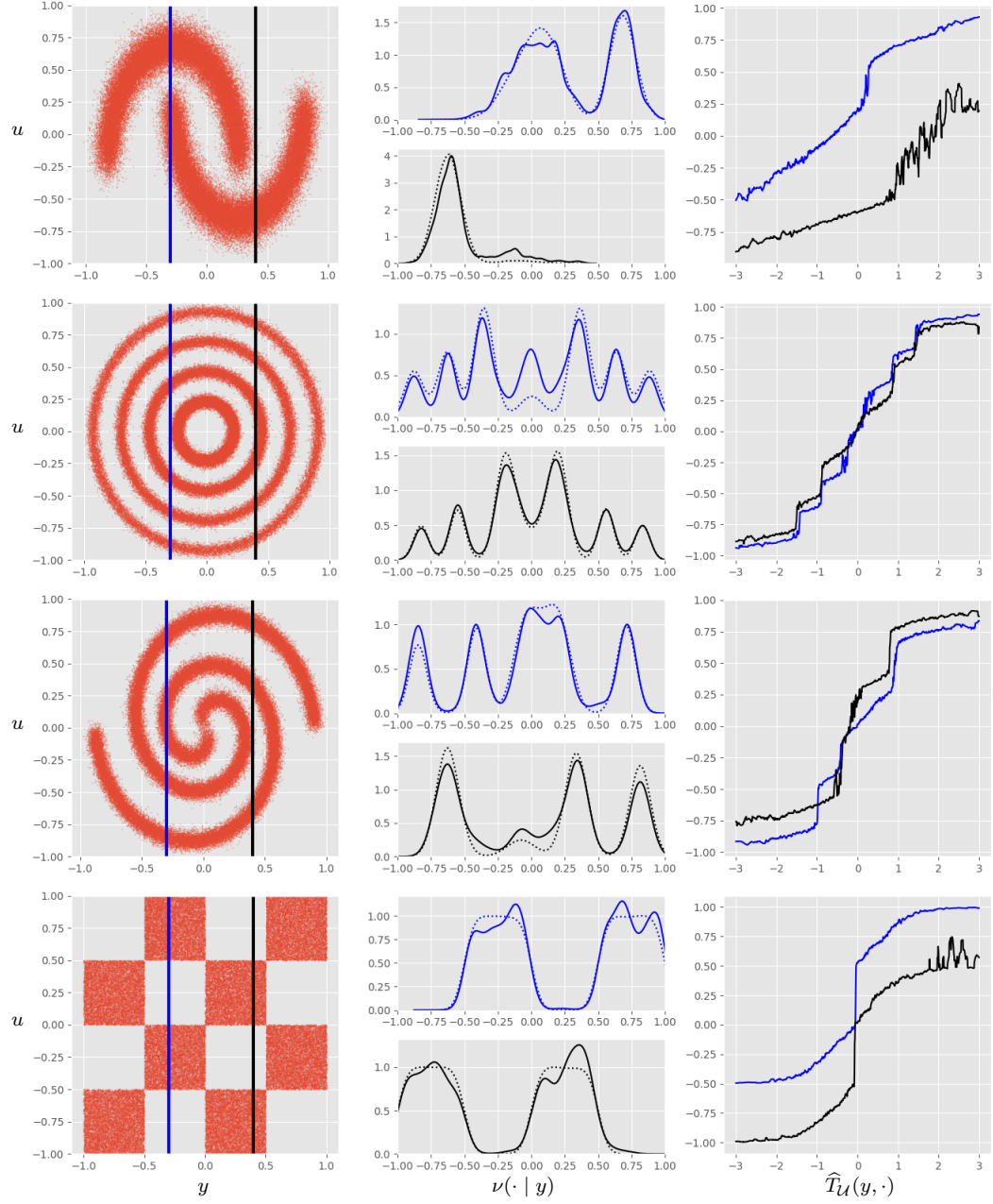


Figure 1.2: Conditional OT experiments on two dimensional benchmarks using Algorithm 1: (Left column) the empirical samples from ν . The blue and black lines indicate slices along which we condition the measures. (Middle column) comparison between the ground truth histogram indicated with dotted lines and the conditional OT histogram indicated with solid lines. (Right column) The conditioning maps $\hat{T}_u(y, \cdot)$ for each slice computed using two nearest neighbor interpolation. The ground truth conditional densities were computed by performing rejection sampling to generate 50,000 samples in the intervals $(y - 0.01, y + 0.01)$, and treating these as conditional samples from $\nu(\cdot | y)$. Kernel density estimation was then used on these samples to plot the conditional distributions.

Algorithm 2 Conditional Wasserstein Monge GAN (WaMGAN)

- 1: **Input:** Samples $\{(y_j, u_j)\}_{j=1}^J \sim \nu$, regularization parameter $\lambda > 0$
- 2: **Output:** Conditioning map $\hat{T}_{\mathcal{U}} : \mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{Y} \times \mathcal{U}$
- 3: Initialize batch size M , critic iterations n_{critic} , learning rate $\alpha > 0$
- 4: Initialize neural operator generator $T_{\mathcal{U}}^{\theta} : \mathcal{Y} \times \mathcal{U} \rightarrow \mathcal{U}$ parameterized by weights θ
- 5: Initialize neural operator critic $C^w : \mathcal{Y} \times \mathcal{V} \rightarrow \mathbb{R}$ parameterized by weights w
- 6: **for** number of training iterations **do**
- 7: Sample a mini-batch $\{(y_j, u_j)\}_{j=1}^M \sim \nu^J$
- 8: Sample a mini-batch $\{v_j\}_{j=1}^M$ from $\eta_{\mathcal{U}}$ and form $\{(y_j, v_j)\}_{j=1}^M$
- 9: Update the critic by ascending its stochastic gradient ¹¹:

$$w \leftarrow w + \alpha \cdot \nabla_w \left[\frac{1}{M} \sum_{j=1}^M C^w(y_j, u_j) - \frac{1}{M} \sum_{j=1}^M C^w(y_j, T_{\theta}(y_j, v_j)) + R(C_w) \right]$$

- 10: Update the generator by descending its stochastic gradient:

$$\theta \leftarrow \theta - \alpha \cdot \nabla_{\theta} \left[\frac{1}{M} \sum_{j=1}^M \|v_j - T_{\mathcal{U}}^{\theta}(y_j, v_j)\|_{\mathcal{U}}^2 - \frac{1}{\lambda M} \sum_{j=1}^M C_w(y_j, T_{\mathcal{U}}^{\theta}(y_j, v_j)) \right]$$

- 11: **end for**
 - 12: $\hat{T}_{\mathcal{U}} \leftarrow T_{\mathcal{U}}^{\theta}$
-

1.5.3 Darcy flow Bayesian inverse problem

For our next example, we consider the Darcy flow PDE in (1.6). Thanks to the Gaussian noise assumption for the data vector y we obtain a quadratic likelihood of the form

$$\Phi(u; y) = \frac{1}{2\sigma^2} \sum_{j=1}^m |p(u)(x_j) - y_j|^2, \quad (1.28)$$

where we recall $\sigma^2 > 0$ is the measurement noise variance and we write $p(u)$ to denote the solution of (1.6) with a diffusion field u .

We take the x_j 's (the measurement locations) to be a uniform 8×8 grid over the unit square as shown in Figure 1.3 where we also show examples of $u^{\dagger}$ and $p(u^{\dagger})$ for our experiments. We consider a Gaussian prior μ on u , in particular, we take $\mu = N(0, C)$ where C is the Matérn covariance kernel $C(x, y) = \left(1 + \frac{\sqrt{3}|x-y|}{\ell}\right) \exp\left(-\frac{\sqrt{3}|x-y|}{\ell}\right)$ with parameter $\ell = 1/2$.

To solve the inverse problem we employed Algorithm 2 and compared its performance to the pCN algorithm [7] with which we take to be the ground truth. To train our map, we used $J = 10^5$ samples that were generated by drawing $u_j \sim \mu$, computing $p(u_j)$ by solving the PDE using the FEniCSx package [133] on a fine mesh and using interpolation to compute $y_j = (p(u_j)(x_1), \dots, p(u_j)(x_{64})) + \epsilon_j$ with $\epsilon_j \sim N(0, \sigma^2 I)$; see our Github repository for further details of parameter choices for the solver and Algorithm 2. In order to obtain mesh invariance, we computed the Karhunen-Loève (KL) expansion of the Gaussian process prior by diagonalizing the associated covariance operator evaluated on a grid, and applied the GAN to the latent KL modes to obtain conditional samples of the u given y in the KL basis

akin to the PCA-Net architecture of [131] and similarly the Monotone GANs of [53]. We tuned the step size of pCN to obtain an average acceptance rate of 0.25 after burn-in and used $5 \cdot 10^6$ iterations to ensure the chains had converged and the mean and variances were accurate and stable. We did not use PCA in our implementation of pCN since we could directly work with the covariance kernel of the prior and used the analytic eigenfunctions.

In Figure 1.3, we present a summary of four instances of the Darcy problem. Here, we drew four samples of $u^\dagger$ from the prior and generated synthetic data y . We solved the inverse problem using pCN for each data, taken to be the ground truth. The WaMGAN map $\hat{T}_U$ was trained only once and was simply evaluated for each value of y to generate posterior samples and compute statistics. This highlights the efficiency of triangular maps since each pCN run required a large number of forward map evaluations while WaMGAN only used 10^5 evaluations. We also took care to present instances of the problem that are representative of the transport approach. In all four examples, we see good agreement between the mean of WaMGAN and pCN; the former often captures the shape and features of the mean but sometimes misses the amplitude. The variances are more smoothed out by WaMGAN, but for the most part, we see good agreement between the two methods in that regard as well. Further insight into the variance can be gained from Figure 1.4 where we plot the pointwise variance of WaMGAN posterior against pCN. Ideally, we would like the data to lie on the diagonal line, but we see a spread of variances, which indicates that WaMGAN can sometimes both over and under estimate variances in each instance of the problem.

In Figure 1.5, we investigate the error (in $L^2(\eta)$) of the learned triangular maps as a function of N the number of KL modes. Here the numerical maps $\hat{T}^N(y, u) = (y, \hat{T}_U^N(y, u))$ are full triangular maps with $\hat{T}_U^N$ taken as empirical approximations to the maps T_U^N defined in Section 1.4.1 with P^N denoting the projection operator onto the first N KL modes of the prior. Since we do not have access to the ground truth map in this example, we numerically computed the high resolution map $\hat{T}^{64}$ and reported the error $\|\hat{T}^N - \hat{T}^{64}\|_{L^2(\eta)}$ using an independent set of reference samples. We observed that the sequence of maps converge as N increases which is expected from our error analysis. This decay also suggests that the problem is of relatively low intrinsic dimension—the noisy subsampled measurements of u are only informative in low frequency modes and posterior variance in higher frequencies is small. We note that the error at $N = 64$ is not zero. We conjecture this accuracy floor is due to the optimization gap of our stochastic optimization algorithm for training the neural nets as indicated by the blue line in Figure 1.5. This line indicates the error between multiple training paths of $\hat{T}^{64}$ by restarting the stochastic gradient descent procedure.

Finally, in Figure 1.6, we compare five independent posterior samples from the two algorithms for one instance of the inverse problem. We see that these samples are qualitatively very similar, indicating that WaMGAN respects the regularity of the samples. We also observed that the Monge penalty allowed WaMGAN to inherit a near monotonicity property in the sense that $\langle \hat{T}_U(y, z_1) - \hat{T}_U(y, z_2), z_1 - z_2 \rangle \geq 0$ for over 99% of the samples in the data, with $\lambda = 0.1$. However, there appears to be a tradeoff between monotonicity and quality of posterior statistics. We found that with $\lambda = 0.05$, we could obtain slightly better posterior samples, mean, and variance, but this resulted in a map that was monotone over a smaller fraction of the training data.

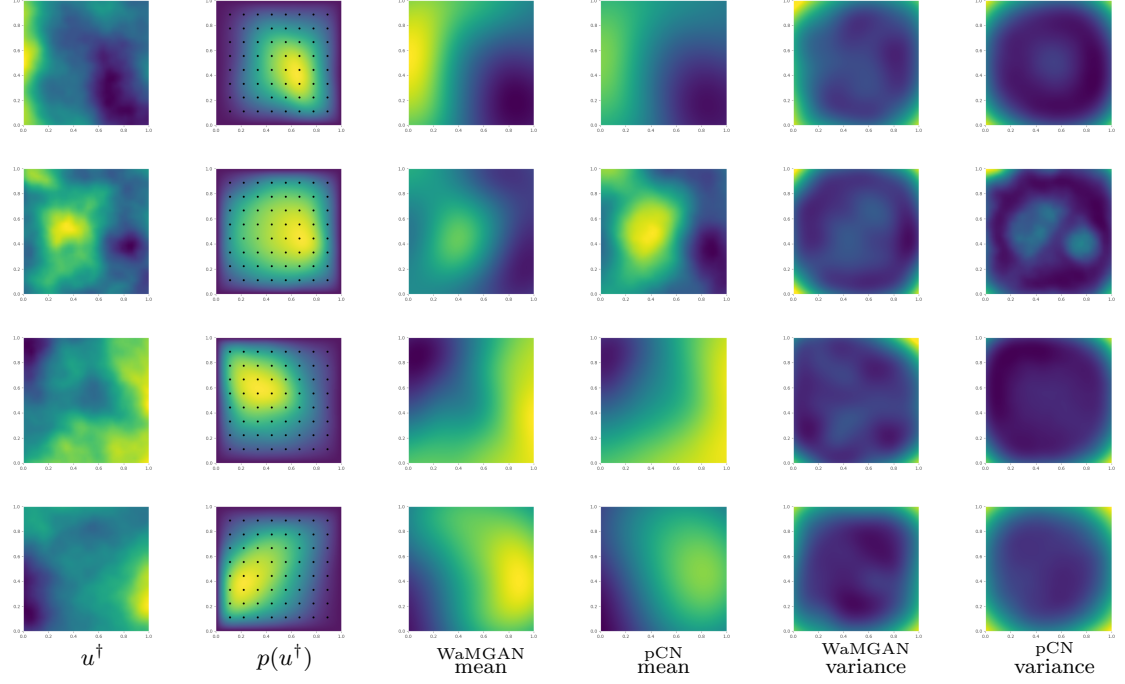


Figure 1.3: Summary of our experiments for Darcy flow inverse problem comparing pCN and WaMGAN. From left, the first column shows the ground truth field $u^\dagger$, second column shows the corresponding $p(u^\dagger)$ and measurement locations x_j denoted as black dots. Third and fourth columns show posterior means for WaMGAN and pCN respectively while the fifth and sixth columns show pointwise variances of the posteriors in the same order. In all cases we used the same color range for pCN and WaMGAN to ensure proper comparison between the two methods.

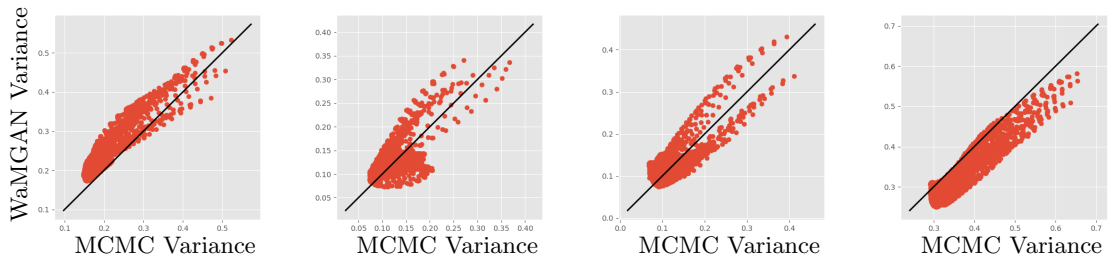


Figure 1.4: Scatterplot comparison of pointwise posterior variances calculated by WaMGAN and pCN for the four examples in Figure 1.3. The diagonal line represents perfect agreement between the two methods.

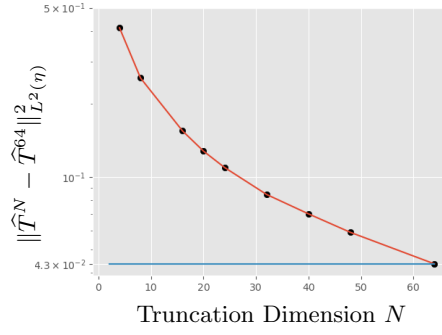


Figure 1.5: Numerical convergence plot of transport maps for the Darcy flow example. We compared models trained on an increasing number of KL modes ($\hat{T}^N$) to a high resolution map trained with 64 modes ($\hat{T}^{64}$), which was taken as the numerical ground truth. We measured the error of the maps with respect to the $L^2(\eta)$ norm which was computed using an independent set of reference samples.

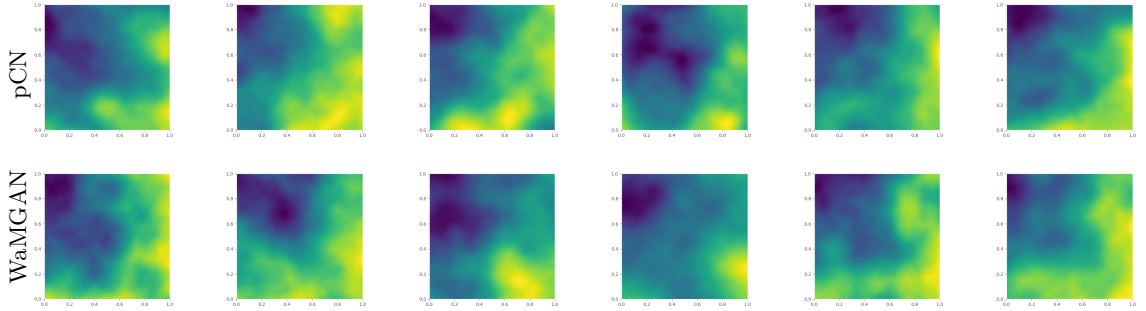


Figure 1.6: A comparison of posterior samples from pCN (top row) vs WaMGAN (bottom row) for one of our experiments.

1.6 Conclusions

We presented a systematic analysis of the conditional OT problem underlying triangular transport problems with a view towards conditional simulation, likelihood-free, and amortized inference problems. We focused our analysis on infinite dimensional spaces, and in particular separable Hilbert spaces, that are useful for PDE inverse problems. We introduced the conditional Kantorovich and Monge problems and studied their solvability and characterized their solutions as a limit of unconstrained OT problems with a singularly perturbed cost and further tailored our theory to the case of BIPs. Finally, we presented two numerical experiments that showed how our analysis could inform the design of algorithms for amortized inference.

While we tackled the foundational theory of conditional OT in this work, there are many open questions and directions that remain. For example, understanding the regularity of conditional OT maps is an interesting and important question in practice, and in particular, the regularity of the $T_{\mathcal{U}}$ map with respect to the y variable. We gave some preliminary results in this direction in Section 1.4.1, showing that the maps are Lipschitz in y , but our results are limited to the finite-dimensional setting and for approximation theory, we may need more regularity. Another related question of interest is the sample complexity of triangular maps, and in particular, the $T_{\mathcal{U}}$ component. The question of interest is *how close is $T_{\mathcal{U}}(y, \cdot) \# \eta_{\mathcal{U}}$ to the target conditional $\nu(\cdot | y)$* ? This is a nontrivial question because having $T \# \eta$ close to ν (on the product space) does not imply closeness of the conditionals. Furthermore, our experiments in Section 1.5.2 indicate that conditioning in low-density areas of ν can lead to higher errors. Finally, a major practical question is that of the design of algorithms for amortized inference based on triangular maps that are efficient and robust. Our experiments indicate that solving OT problems as linear programs yields poor convergence rates in high dimensions. However, working directly with the Monge formulation allowed us to leverage the approximation power of neural nets and obtain good results in a Hilbert space setting. While we used a simplistic relaxation of the Monge problem for this task, the question remains open as to what is the "best" relaxation of the Monge problem that leads to good empirical performance and approximates the conditional OT map; see [53, 70–73] for other works in this direction. Approaches based on flow matching [77, 134] lead to transport maps defined by ODEs whose vector fields are convenient for computation but do not necessarily seek an OT map. However, recent works show that flow matching can still approximate OT maps via minibatch couplings that further improve performance [78, 79].

Acknowledgments

BH and AH are supported by NSF grant DMS-2208535, "*Machine Learning for Bayesian Inverse Problems*". AH was also partially supported by the NSF Infrastructure grant DMS-2133244, "*Pacific Interdisciplinary Hub on Optimal Transport*" made to the PIMS Kantorovich Initiative. AT is supported by NSF grant EPCN-2318977, "*Variational Optimal Transport Methods for Nonlinear Filtering*".

Chapter 2

DATA-EFFICIENT KERNEL METHODS FOR LEARNING DIFFERENTIAL EQUATIONS AND THEIR SOLUTION OPERATORS

Preface

In this chapter, from [2] which was joint work with Yasamin Jalalian, Juan Felipe Osorio Ramirez, Bamdad Hosseini, and Houman Owhadi, we develop a kernel-based framework for learning differential equations (DEs) from data, along with accompanying error analysis, and discuss the connections to operator learning and PDE solvers. The fundamental problem is the following: given measurements of a system state $u(y)$ which is assumed to satisfy some differential equation

$$P(y, u(y), (D_1 u)(y), \dots, (D_j u)(y)) = f(y), \quad y \in \Omega \quad (2.1)$$

for a domain Ω , and differential operators D_k , infer the function P . A fundamental challenge in this identification problem is reconciling that DEs relate states to their derivatives, but we rarely observe all of the variables involved in the equation. For example, given data in the form of point observations, derivatives must be estimated from said data before any learning can occur, inherently introducing errors that compound from any kind of statistical/numerical differentiation process. This work introduces methods based on approximating both the unknown system state $u(y)$, and the unknown governing equations P as functions in appropriate reproducing kernel Hilbert spaces (RKHS) using the observed data.

One straightforward approach to this problem is to formulate a two step procedure, where an estimate $\hat{u}$ is first generated from the measurements, and then an estimate of $\hat{P}$ is produced using the derivatives computed from $\hat{u}$. Since P is a function of y , u , and the derivatives of u *pointwise*, we may form estimates of the form

$$\widehat{\mathcal{P}}(u)(y) = \hat{P}(y, u(y), (D_1 u)(y), \dots, (D_j u)(y)). \quad (2.2)$$

As such, both stages can be reduced to traditional regression problems to infer u and P . However, this leaves out the fact that the governing equations P may inform the estimates of the system state u , and better estimates of the system state u would further improve estimates of the governing equations.

To this end, we propose one-step kernel methods, where estimates $\hat{P}$ and $\hat{u}$ are computed together. We parametrize both $\hat{P}$ and $\hat{u}$ as functions in appropriate reproducing kernel Hilbert spaces, enforce the differential equation via least squares collocation, regularize via the RKHS norms, and apply a Levenberg-Marquardt based optimization procedure to solve the resulting optimization problem.

This work also develops accompanying error analysis, which centers around convergence in sufficiently high order Sobolev spaces so that derivatives of u can be well approximated, and sufficient coverage of the space induced by observing u and its derivatives so as to recover P .

2.1 Introduction

In recent years, machine learning (ML) has revolutionized the way data is combined with mathematical models to infer and predict the behavior of physical systems. This wide adoption of ML in science has given rise to a new area of computational science and engineering often referred to as physics-informed ML (PIML) [135] or more broadly, scientific machine learning (SciML). Broadly speaking, the goal here is to infer and simulate physical processes driven by differential equations (DEs) by combining data and expert knowledge in an automated manner. In this chapter we focus on three core problems of PIML following [136], i.e., the problems of learning or discovering DEs [137, 138]; learning or approximating their solution operators [139, 140]; and finally solving DEs [141, 142]. All three of these problems have a long history in control, scientific computing, and numerical analysis but the wide adoption of ML has given rise to new ideas and techniques for solving them and in particular blurred the lines between these three problems. In this chapter we are primarily focused on solving the above problems in very scarce data regimes and we shall introduce a simple, accurate, robust, and efficient algorithm based on the theory of Reproducing Kernel Hilbert Spaces (RKHSs) and Gaussian Processes (GPs) leading to the unification of the three problems of PIML under the same umbrella.

2.1.1 Problem statement

For a function $u : \mathcal{Y} \rightarrow \mathbb{R}$ which describes the state of a physical system, we consider PDEs of the form

$$\mathcal{P}(u)(y) = f(y) \quad y \in \mathcal{Y}, \quad \mathcal{B}(u)(y) = g(y) \quad y \in \partial\mathcal{Y}, \quad (2.3)$$

where $\mathcal{P} : \mathcal{U} \rightarrow \mathcal{F}$ denotes a differential operator that describes a PDE in the interior of $\mathcal{Y} \subset \mathbb{R}^d$, $\mathcal{B} : \mathcal{U} \rightarrow \mathcal{G}$ denotes the boundary operator, f and g denote the source term and boundary data/initial conditions, and $\mathcal{U}, \mathcal{F}, \mathcal{G}$ denote appropriate function spaces for u, f, g . We emphasize that in the above formulation y is considered as an abstract input variable that may be spatial only (in the case of steady state PDEs) or a space-time variable (in the case of dynamics). An important feature of PIML is the seamless fusion of noisy observations and prior knowledge about the physical process of interest, thereby unifying numerical PDEs with inverse problems. To this end, we introduce observation operators $\varphi_{\mathcal{U}} : \mathcal{U} \rightarrow \mathbb{R}^{d_{\mathcal{U}}}, \varphi_{\mathcal{F}} : \mathcal{F} \rightarrow \mathbb{R}^{d_{\mathcal{F}}}, \varphi_{\mathcal{G}} : \mathcal{G} \rightarrow \mathbb{R}^{d_{\mathcal{G}}}$. These are often taken to be pointwise observations of the functions u^m, f^m , and g^m on a scattered set of points in $\mathcal{Y}$, where $(u^m, f^m, g^m)_{m=1}^M$ are tuples that satisfy (2.3). From this, we assume access to a data set of the form $(\varphi_{\mathcal{U}}(u^m) + \varepsilon_{\mathcal{U}}, \varphi_{\mathcal{F}}(f^m) + \varepsilon_{\mathcal{F}}, \varphi_{\mathcal{G}}(g^m) + \varepsilon_{\mathcal{G}})_{m=1}^M$, where $\varepsilon_{\mathcal{U}}, \varepsilon_{\mathcal{F}}, \varepsilon_{\mathcal{G}}$ represent possible observation noises. With this data set at hand, we then define the following three core problems of PIML:

- (P1) *Operator learning*: Infer the solution map
- (P2) *Equation learning or discovery*: Infer the differential operators $\mathcal{P} : \mathcal{U} \rightarrow \mathcal{F}$ and $\mathcal{B} : \mathcal{U} \rightarrow \mathcal{G}$ from data; $(\mathcal{P}, \mathcal{B})^{-1} : \mathcal{F} \times \mathcal{G} \rightarrow \mathcal{U}$ from data;
- (P3) *PDE solvers*: Suppose $\mathcal{P}, \mathcal{B}$ are known (perhaps partially) and compute u for a given (f, g) .

Note that the above problems are not entirely distinct. For example, (P1) and (P2) are both problems of inferring operators, just between different spaces. At the same time (P1) and (P3) can be viewed as the same since computing u amounts to evaluating $\mathcal{P}^{-1}$. Regardless, these problems are often considered separately due to their different target applications and solution methods in PIML.

2.1.2 Summary of contributions

Our contribution in this chapter is to propose a computational framework based on the theory of RKHSs and GPs for (P2) Equation learning, which we call Kernel Equation Learning (KEqL), and to unify the above PIML problems through our proposed methodology. Our approach offers significant technical advantages over existing methods, including state-of-the-art neural network techniques, in terms of: (a) accuracy, data efficiency, and computational efficiency, achieving superior performance across multiple benchmarks; (b) rigorous theoretical guarantees, providing the first known quantitative worst-case error bounds for equation learning further leading to operator learning errors; (c) robust and efficient training, leveraging specialized optimization algorithms for improved convergence and stability; and (d) new capabilities in PIML, enabling one-shot operator learning and the discovery of variable-coefficient partial differential equations (PDEs) even in severely data-scarce settings (when M or $d_{\mathcal{U}}$ are small).

Our exposition in this chapter is based on two core ideas:

1. *Leveraging the fact that (u^m, f^m, g^m) collectively satisfy the PDE constraint defined by $\mathcal{P}, \mathcal{B}$ is critical to improving the accuracy of equation learning in scarce data regimes.*
2. *Learning $\mathcal{P}$ in scarce data regimes is more efficient and accurate than learning $\mathcal{P}^{-1}$, i.e., operator learning through equation learning is more effective.*

2.1.3 Review of the relevant literature

We now give a brief review of the related works in the context of problems (P1)-(P3) and connect them to more traditional areas of control, scientific computing, and applied mathematics. We will give ad hoc comments placing our contributions in the context of existing works throughout the chapter and so readers primarily interested in the methodology may safely skip to Section 2.2.

Equation learning and system identification

The discovery or learning of DEs from data was brought into prominence by [143, 144] that used symbolic regression to discover physical laws although learning dynamical systems from time-series data has a long history going back to the 70s in *system identification* [145–147]. Modern iterations of equation learning are often focused on learning dynamical systems and DEs under a sparsity prior over a dictionary [137, 138, 148] among which the Sparse Identification of Nonlinear Dynamics (SINDy) [137] is perhaps the most widely known instance although many other examples exist [148, 149]. Since these contributions, many extensions of the sparse regression have been proposed such as weak-SINDy [150], that improves accuracy and robustness by working with weak-form DEs, ensemble methods [151] that combine multiple models to improve accuracy and robustness, and Bayesian methods [152–157] that enable uncertainty quantification; see [158] and references within for a more detailed review.

The majority of the methods mentioned above belong to a class of methods which we refer to as "2-step methods" in Section 2.2. These are methods where the state variables u^m are first estimated from the observations $\varphi(u^m)$ before being used to infer $\mathcal{P}$. As we will show in Section 2.4, these methods are successful only in high data regimes and fail with scarce data. To overcome this hurdle we propose a so-called "1-step method" where the u^m are estimated simultaneously alongside $\mathcal{P}$ and $\mathcal{B}$. 1-step methods for sparse regression of DEs are quite limited and to our knowledge they were only explored in [159] for ODEs and [160] for PDEs, although methods such as weak-SINDy [150] and UQ-SINDy [155] can be formulated as 1-step methods as well. Similarly, joint optimization methods based on least squares collocation formulations using splines [159] and neural networks [160] have been proposed to extend the SINDy framework towards 1-step methods.

Beyond sparse regression models, RKHS or GP methods have been widely developed for solving PDEs [142, 161–168] but they have recently been extended for learning equations, most notably [169, 170] introduced 1-step methods for learning linear DEs and ODEs in that order. The series of papers [171–176] also introduced GPs for inference of ODEs towards filtering and data assimilation. More recently, [177] introduced a kernel analogue of SINDy and weak-SINDy while [178, 179] aimed at learning stochastic differential equations. Finally, [180] introduced a simple 2-step kernel method for learning ODEs and PDEs with unknown coefficients, inspiring our present chapter.

More recently, various neural net models for learning DEs have also been proposed. These methods range from symbolic regression [181–185] to neural net regression [160, 186–188]. The latter class are largely based on the methodology of physics-informed neural nets (PINNs) [141].

Theory of equation learning

The problem of learning dynamical systems has a vast and mature theoretical foundation [189]. The theory of equation learning for DEs, on the other hand, is recent and scarce in comparison. The convergence properties of sparse regression methods (e.g., SINDy) were studied in [148, 149, 190, 191] using compressed sensing and random matrix theory to provide recovery guarantees. The article [192] studies the identifiability of equation learning

for ODEs and PDEs while [193] considers the limitations of learning a PDE from a single trajectory.

Inverse problems and data assimilation

Identifying the parameters of a DE is the primary focus of the field of inverse problems [194, 195] with a rich history of both theoretical analysis [196, 197] and computational methodologies [198]. While traditional inverse problems focus on known PDEs with unknown spatio-temporal coefficients, the methodologies developed for those problems can be extended to equation learning in both 1-step and 2-step methodologies. Many of the ideas that we develop in the current chapter were inspired by the inverse problems literature. Most notably joint inversion [199] and all-at-once inversion [200, 201].

The problem of recovering the functions u^m and in turn predicting the solution of a dynamic DE in future times falls within the fields of data assimilation [202, 203], filtering [204, 205], and control [206, 207]. The problem of filtering the state of a control system while identifying its unknown parameters is also classical in filtering [208, 209] and can be solved using classic techniques such as extended Kalman smoothing [210, Sec. 5], expectation maximization methods [211, 212], dual extended Kalman filters [213], GP dynamical models [214, 215], and sequential Monte Carlo [216].

Operator learning

Operator learning has become very popular in recent years, especially after the seminal papers [131, 217, 218] where neural net models were developed for the approximation of mappings between Banach spaces. Since then, a large body of work has been developed around operator learning focusing on methodologies as well as theory; see [139] and references within for neural operators and [140] for the kernel perspective. While modern operator learning is largely focused on neural net models this problem has appeared previously in the fields of computer model emulation [219], polynomial chaos [220], stochastic Galerkin methods [221], reduced order modeling [222], and model order reduction [223]. We consider a different perspective towards operator learning. Instead of learning $\mathcal{P}^{-1}$ as a high dimensional mapping, we learn $\mathcal{P}$ as a local differential operator and then invert it numerically to evaluate $\mathcal{P}^{-1}$. To our knowledge, this approach towards operator learning is largely unexplored with the exception of [224–226] that use the differential form of the underlying DE to enhance neural operator training.

2.1.4 Outline of the chapter

The rest of the chapter is organized as follows: In Section 2.2 we outline the optimal recovery framework of 2-step and 1-step kernel methods for learning equations and their solution operators. Section 2.3 describes the algorithmic and implementation details of our methods followed by numerical experiments and benchmarks in Section 2.4. Our theoretical analysis and quantitative error bounds are summarized in Section 2.5. We conclude with additional discussions and future extensions in Section 2.6. Finally, additional information

including omitted proofs, algorithmic details, and extra numerical results are collected in the supplementary material (SM) at the end of this chapter, Sections 2.7 to 2.9.

2.2 Kernel equation learning

In this section we outline the Kernel Equation Learning (KEqL) framework for learning DEs from data starting with a 2-step methodology in Section 2.2.1 followed by a 1-step method in Section 2.2.2; these approaches target (P2) Equation Learning. We then introduce our equation-mediated operator learning method in Section 2.2.3, that tackles (P1) Operator learning.

Let us begin by stating some simplifying assumptions and outlining our setup. We consider the problem of learning the differential operator $\mathcal{P}$ only since this is often the problem of interest in practice and note that our methodology can be generalized to learning the boundary operator $\mathcal{B}$ in a similar way. We also assume that u and f are scalar fields on $\mathcal{Y}$ and postpone the learning of systems and vector fields to future work. Throughout the chapter we assume $\mathcal{P}$ has the form

$$\mathcal{P}(u)(y) = P \circ \Phi(u, y), \quad \Phi : (u, y) \mapsto (y, \delta_y \circ L_1 u, \dots, \delta_y \circ L_Q u) \in \mathbb{R}^{Q+d}, \quad (2.4)$$

and write $\mathcal{S} = \mathbb{R}^{Q+d}$ for brevity. In the above $P : \mathcal{S} \rightarrow \mathbb{R}$ is a generic function mapping point observations of u and its derivatives to $\mathbb{R}$, encoding our estimate of the PDE, and includes possible nonlinearities or variable coefficients. δ_y denotes the pointwise evaluation functional at y , and $(L_q)_{q=1}^Q$ are bounded and linear differential operators that are assumed to be known. For ease of presentation we always assume $L_1 = \text{Id}$ so that P takes y and the point values $u(y)$ as input even though it may not depend on these quantities. We shall assume that the L_q operators are known and note that the map Φ is linear in u but it is generally nonlinear in y since u is nonlinear. The function P and the solution u are the main objects of interest that we wish to learn from data in the context of equation learning. This structure imposes that the operator $\mathcal{P}$ is completely local.

Example. The above assumptions on the form of $\mathcal{P}$ encompass most scalar PDEs of interest in physics and engineering. For example consider the nonlinear PDE:

$$\begin{cases} \mathcal{P}(u)(t, x) = \partial_t u(t, x) - \partial_x(a(x)\partial_x u(t, x)) - u^3(t, x) = f(t, x), \\ u(0, x) = u(t, 0) = u(t, 1) = 0, \end{cases} \quad \text{for } (t, x) \in (0, T] \times (0, 1), \quad (2.5)$$

with a smooth coefficient $a : [0, 1] \rightarrow (0, +\infty]$. Writing $y = (t, x)$ and letting $\mathcal{Y} = (0, T] \times (0, 1)$ and introducing the differential operators $L_1 : u \mapsto u$, $L_2 : u \mapsto \partial_t u$, $L_3 : u \mapsto \partial_x u$, and $L_4 : u \mapsto \partial_x^2 u$ we can cast (2.5) in the form (2.4) by writing $\mathcal{P}(u)(y) = P(t, x, u(t, x), L_2 u(t, x), L_3 u(t, x), L_4 u(t, x))$ with the nonlinear map

$$P : \mathbb{R}^6 \rightarrow \mathbb{R}, \quad P(s_1, s_2, s_3, s_4, s_5, s_6) = s_4 - \partial_x a(s_2)s_5 - a(s_2)s_6 - s_3^3.$$

We emphasize that the example PDE above is precisely the type of equation that we have in mind, i.e., nonlinear equations with variable coefficients that may not have sparse representations in a known basis. In this light, we merge equation learning, the problem of learning P , with inverse problems, the problem of inferring unknown coefficients such as a .

Recalling the notation of Section 2.1.1, we further focus our setup by considering equation learning problems where the solutions u^m are observed on a finite set of points, i.e., when $\varphi_{\mathcal{U}}$ is a collection of pointwise evaluation functionals. More precisely, consider sets of points $Y^m = \{y_1^m, \dots, y_N^m\} \subset \mathcal{Y}$, which we refer to as the *observation points*¹, for some integer $N \geq 1$. Further introduce the compact notation $u^m(Y^m) := (u^m(y_1^m), \dots, u^m(y_N^m)) \in \mathbb{R}^N$. Then our goal throughout the chapter is to learn P from the training data $(u^m(Y^m), f^m)_{m=1}^M$, i.e., u^m is only observed on the Y^m while f^m is assumed to be known everywhere; this is a simplifying assumption for us and can be relaxed to having finite information on the f^m under some circumstances but we will leave this as a future direction. Figure 2.1(a) shows a cartoon of our problem where the red elements are the unknowns of interest, the blue elements are the data, and the black elements are assumed to be known.

With the above setup, we propose three approaches for learning P and the u^m : (1) a 2-step kernel method (2-step KEqL), where we first learn the u^m 's from the observations $u^m(Y^m)$ and use the result to approximate P , as a kernel analog to the PDE-FIND algorithm of [138]; and (2) a 1-step kernel method (1-step KEqL), where u^m and P are learned jointly, and can be viewed as a kernel analogue to [160]. Our 2-step and 1-step methods can also be cast as computational graph completion problems in the setting of [227, 228]. We show these computational graphs in Figure 2.1(b,c) where, as before, the red edges denote unknown nonlinearities that we wish to infer from the data is depicted using dashed blue lines and is injected into the nodes which represent our variables. Black edges are assumed to be known.

2.2.1 2-step KEqL: first learn the u^m , then learn P

Let $\mathbf{U} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ denote a positive definite and symmetric (PDS) kernel with its associated RKHS $\mathcal{U}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{U}}$ and norm $\| \cdot \|_{\mathcal{U}}$; for a review of RKHS theory see Section 2.5.1 or the texts [229, 230]. Then the first step in 2-step KEqL approximates u^m via the optimal recovery problems²

$$\hat{u}^m := \arg \min_{v^m \in \mathcal{U}} \|v^m\|_{\mathcal{U}} \quad \text{subject to (s.t.)} \quad v^m(Y^m) = u^m(Y^m). \quad (2.6)$$

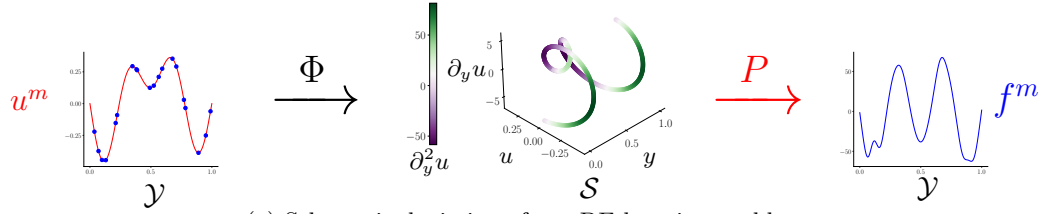
We can also relax the equality constraints using a penalty method leading to a quadratic optimization problem with nugget parameter (or observation noise standard deviation) $\sigma_u^2 > 0$,

$$\hat{u}^m := \arg \min_{v^m \in \mathcal{U}} \|v^m\|_{\mathcal{U}}^2 + \frac{1}{2\sigma_u^2} \|v^m(Y^m) - u^m(Y^m)\|_2^2. \quad (2.7)$$

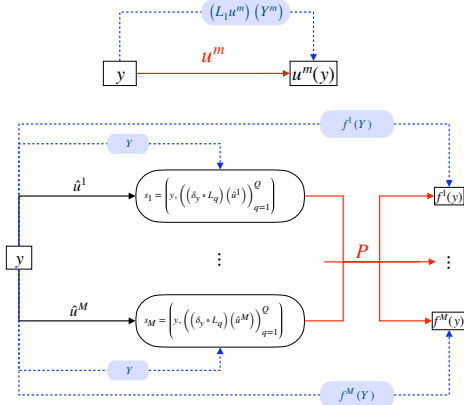
With $\hat{u}^m$ at hand we proceed to step two where we approximate P through a second optimal recovery problem. To do this, let us consider an independent set of points $Y = \{y_1, \dots, y_K\} \subset \mathcal{Y}$ for integer $K \geq 1$ which we call *collocation points*. This set of points may be chosen independent of the individual Y^m but to make our formulation simpler we will

¹One may also let N change with the index m but we keep the size of the mesh fixed to keep our notation light.

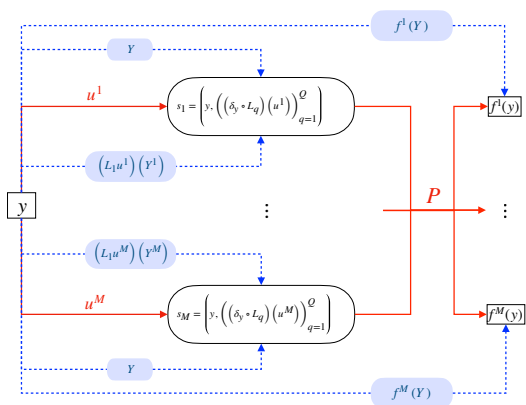
²Note that one could also employ a different kernel $\mathbf{U}^m$ for each instance of the problem but we will not pursue this for brevity.



(a) Schematic depiction of our DE learning problem.



(b) Computational graph of 2-step KEqL



(c) Computational graph of 1-step KEqL

Figure 2.1: (a) Schematic depiction of our equation learning problem. Red elements are the unknowns of interest, blue objects are the given data, while black objects are assumed to be known. The left and right panels show the solution and right-hand side of an example second order PDE depending on $y, u, \partial_y u$, and $\partial_y^2 u$ while the middle panel shows $\Phi(y, u)$; (b) The computational graph of 2-step KEqL. Red edges are unknown nonlinearities of interest, blue boxes denote data that is known for various nodes with dashed lines denoting where the data is injected, black edges are known. Note that the graphs for u^m and P are disconnected, hence the learning of u^m and P is performed sequentially in two steps; (c) The computational graph for 1-step KEqL. Coloring conventions follow panel (b) with the main difference being that the u^m and P are now connected and have to be learned simultaneously.

assume it is chosen such that $\cup_{m=1}^M Y^m \subset Y$, i.e., the collocation points Y contain all of the observation points Y^m .

Now observe that the differential operators L_q can be directly applied to the $\hat{u}^m$, in fact, as we show in Section 2.3 the functions $L_q \hat{u}^m$ can be computed analytically or using automatic differentiation. Then given a PDS kernel $P : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ with RKHS $\mathcal{P}$ we can approximate P via the optimal recovery problem

$$\hat{P} := \arg \min_{G \in \mathcal{P}} \|G\|_{\mathcal{P}} \quad \text{s.t.} \quad G \circ \Phi(\hat{u}^m, Y) = f^m(Y), \quad m = 1, \dots, M, \quad (2.8)$$

where $f^m(Y) := (f^m(y_1), \dots, f^m(y_K)) \in \mathbb{R}^K$ and we used the shorthand notation $\Phi(\hat{u}^m, Y) = (\Phi(\hat{u}^m, y_1), \dots, \Phi(\hat{u}^m, y_K)) \subset \mathcal{S}$. Similar to (2.7) this problem can also be relaxed using a nugget parameter $\sigma_P^2 > 0$,

$$\hat{P} := \arg \min_{G \in \mathcal{P}} \|G\|_{\mathcal{P}}^2 + \frac{1}{2\sigma_P^2 M} \sum_{m=1}^M \|G \circ \Phi(\hat{u}^m, Y) - f^m(Y)\|_2^2. \quad (2.9)$$

In either case $\hat{P}$ also admits an analytic formula akin to the $\hat{u}^m$ which we present in Section 2.3. Note that in (2.8, 2.9) we are using the collocation points Y to impose the infinite PDE constraint $G \circ \Phi(\hat{u}^m, y) = f(y)$ for all $y \in \mathcal{Y}$, on a finite discrete set, which justifies our choice of terminology as this is precisely the role of collocation points in PDE solvers [142]. This also motivates our preference to choose Y to be dense within computational budget constraints.

In many applications we have some prior knowledge about the differential operator $\mathcal{P}$, for example in dynamic problems we know that a time derivative of $\partial_t^h u$ for some integer h is present. In such cases we can simply work with the model $\mathcal{P}(u) = (P + \bar{P}) \circ \Phi(u, \cdot)$ with $\bar{P}$ representing the known part of the differential operator. This modification amounts to a simple reformulation of (2.9) (and similarly (2.8)) as

$$\hat{P} = \arg \min_{G \in \mathcal{P}} \|G\|_{\mathcal{P}}^2 + \frac{1}{2\sigma_P^2 M} \sum_{m=1}^M \|(G + \bar{P}) \circ \Phi(\hat{u}^m, Y) - f^m(Y)\|_2^2. \quad (2.10)$$

Henceforth we will include $\bar{P}$ to account for the known parts of $\mathcal{P}$.

Remark 2.1. We highlight two key observations: (1) Taking $\mathcal{U}, \mathcal{P}$ to be RKHSs is innocuous at this point in our exposition and only becomes important for our theory and algorithms later. Therefore, one can take $\mathcal{U}, \mathcal{P}$ to be abstract Banach spaces to obtain a generalization of our framework so long as the observation functionals $\varphi_{\mathcal{U}}$ are bounded in the appropriate dual space, and functions $\mathcal{P}$ are regular enough for solutions to be defined. For example, taking these to be Barron spaces [231] models a neural net method such as that of [160] while a sequential l_1 or l_0 space over some basis coefficients model SINDy type methods [137, 138]; (2) as mentioned previously, problems (2.6) and (2.8) can be viewed as sequential computational graph completion tasks following the framework of [227] and as shown in Figure 2.1(b). The simplicity of these graphs reflects the convenience of the 2-step method. However, this convenience comes at the cost of accuracy and data efficiency since the 2-step method will only succeed when data is sufficient for the $L_q \hat{u}^m$ to be good approximations to $L_q u^m$ which is difficult for higher order derivatives, and approximation with respect to higher order Sobolev norms from only point observations. $\diamond$

Connection to existing works A slightly different version of 2-step KEqL was first introduced in [180] as the RKHS/GP analogue of SINDy/PDE-FIND [137, 138]. The RKHS formulation is different from these works in three directions: (1) the prior knowledge about P is encoded in the kernel $\mathcal{P}$ as opposed to a dictionary; (2) SINDy looks for $\hat{P}$ that is sparse in the dictionary while KEqL computes a minimum RKHS norm solution that is not necessarily sparse; (3) we readily include y as an input to P to accommodate variable coefficient PDEs. Dealing with such problems is problematic for SINDy [232] since the variable coefficients may depend on y in a complex manner that is not sparse in a particular dictionary.

2.2.2 1-step KEqL: simultaneously learn the u^m and P

The primary shortcoming of 2-step KEqL is the decoupled learning of the u^m and P which limits its performance in scarce data regimes. To remedy this we propose 1-step KEqL that estimates $\hat{u}^m$ and $\hat{P}$ at the same time while imposing the requisite PDE constraint on the collocation points Y . To this end, consider the joint optimal recovery problem

$$\begin{aligned} (\hat{\mathbf{u}}, \hat{P}) = \arg \min_{\mathbf{v} \in \mathcal{U}^M, G \in \mathcal{P}} \quad & \|G\|_{\mathcal{P}}^2 + \frac{\lambda}{M} \sum_{m=1}^M \|v^m\|_{\mathcal{U}}^2 \\ \text{s.t.} \quad & v^m(Y^m) = u^m(Y^m), \\ & (G + \bar{P}) \circ \Phi(v^m, Y) = f^m(Y), \quad \text{for } m = 1, \dots, M, \end{aligned} \quad (2.11)$$

where $\lambda > 0$ is a user defined parameter and we used the notation $\mathbf{v} := (v^1, \dots, v^M) \in \mathcal{U}^M$ to denote the vector of candidate RKHS functions with their optimal values denoted as $\hat{\mathbf{u}} \in \mathcal{U}^M$. In complete analogy with 2-step KEqL we can also relax the equality constraint in (2.11) using appropriate nugget parameters $\sigma_u^2, \sigma_P^2 > 0$ to obtain the formulation,

$$\begin{aligned} (\hat{\mathbf{u}}, \hat{P}) = \arg \min_{\mathbf{v} \in \mathcal{U}^M, G \in \mathcal{P}} \quad & \|G\|_{\mathcal{P}}^2 + \sum_{m=1}^M \frac{\lambda}{M} \|v^m\|_{\mathcal{U}}^2 + \frac{1}{2\sigma_u^2} \|v^m(Y^m) - u^m(Y^m)\|_2^2 \\ & + \frac{1}{2\sigma_P^2} \|(G + \bar{P}) \circ \Phi(v^m, Y) - f^m(Y)\|_2^2. \end{aligned} \quad (2.12)$$

Note that the estimation of the $\hat{\mathbf{u}}$ and $\hat{P}$ is now coupled due to the composition of G and the v^m in the constraint in (2.11) or in the last term in (2.12). This coupling of the estimation of the u^m and P is the source of algorithmic challenges to solving 1-step KEqL. Nonetheless, we solve this problem after reformulation using a representer theorem and using an efficient Levenberg–Marquardt (LM) algorithm, described in Section 2.3.

Remark 2.2. We highlight three key observations: (1) Once again our choice of $\mathcal{U}, \mathcal{P}$ as RKHSs is innocuous at this point and choosing these to be other Banach spaces leads to other 1-step methods but this will have important consequences at the level of algorithms and theory as was discussed in Remark 2.1; (2) the second term in (2.11) resembles the discretization of a PDE using a kernel collocation method akin to the method of [142]; (3) as mentioned previously 1-step KEqL can also be viewed as a computational graph

completion problem with its graph shown in Figure 2.1(c). However, compared to the graph of 2-step KEqL we see that the underlying graph is more complex and in particular edges corresponding to the u^m and P are both unknown, reflecting the compositional structure of the problems (2.11) and (2.12). $\diamond$

Connection to existing methods The concept of optimizing a loss function that jointly matches the given observations and the PDE constraint has been previously explored in the PIML and inverse problem literature. Most notably, in [159, 160] use sparse regression for P but neural net or spline models for the u^m . At the same time weak SINDy [150, 233] and SINDy-UQ [155] can also be formulated as 1-step methods although they were not presented or implemented as such. Our 1-step KEqL formulation is different from these works in two directions: (1) the regularization due to our formulation automatically leads to a stable algorithm that is amenable to second order optimization when using RKHS methods; (2) the kernel formulation accommodates a larger number of features and more flexibility in choosing and tuning the features for both u^m and P including variable coefficients.

2.2.3 Equation-mediated operator learning

We now turn our attention to the operator learning problem of estimating the solution map $\mathcal{P}^{-1} : \mathcal{F} \rightarrow \mathcal{U}$ of the PDE (2.3). The dominant paradigm for operator learning [139] approximates $\mathcal{P}^{-1}$ via function space regression directly from input-output pairs. A key appeal in doing this is learning discretization-independent maps between functions, using grids only through finite measurements. However, the inverse solution operator $\mathcal{P}^{-1}$ is often global and nonlocal, so purely data-driven learning must infer a complex high-dimensional map without directly using the PDE’s local structure. With many solution pairs this works well, but in sparsely measured data regimes, direct operator regression is highly underdetermined. This motivates an alternative, equation-mediated view of operator learning. In our perspective, operator and equation learning problems are two sides of the same coin: where equation learning approximates $\mathcal{P}$, operator learning approximates the inverse $\mathcal{P}^{-1}$. To this end, following [180], let us write $\widehat{\mathcal{P}} := (\widehat{P} + \overline{P}) \circ \Phi$ to denote the differential operator associated with $\widehat{P}$. We then propose to approximate the operator $\mathcal{P}^{-1}$ by computing the *pseudo inverse* operator $\widehat{\mathcal{P}}^\dagger$ defined variationally as

$$\widehat{\mathcal{P}}^\dagger(f) := \arg \min_{v \in \mathcal{U}} \|v\|_{\mathcal{U}} \quad \text{s.t.} \quad (\widehat{P} + \overline{P}) \circ \Phi(v, Y) = f(Y). \quad (2.13)$$

We will solve this problem using the GP-PDE solver of [142, 234] which is effectively a meshless collocation method. It is important to note that the operator $\widehat{\mathcal{P}}$ is not invertible in general since the learned PDE is not guaranteed to be well-posed. Then (2.13) computes a well-defined regularized solution to this ill-posed PDE.

Remark 2.3. We highlight two important observations: (1) solving $\widehat{\mathcal{P}}(u) = f$ using traditional finite volume or finite element solvers is often impossible due to presence of arbitrary nonlinearities and so employing residual based solvers such as GP-PDE [142] or PINNs [141] is crucial; (2) the estimation of $\widehat{\mathcal{P}}^\dagger$ comes at the cost of a PDE solve. This may be expensive for model emulation applications, reflecting an accuracy-complexity tradeoff. However, one

can still train an emulator for $\widehat{\mathcal{P}}^\dagger$ using synthetic data generated by solving (2.13) for a set of f 's. $\diamond$

Connection to existing methods To our knowledge the idea of operator learning through equation learning has been barely explored in the literature since most works in this area focus on function space regression with neural operators [136, 139], although other models such as kernel methods and GPs remain competitive [140, 235]. In most operator learning applications the training data is plentiful and high resolution, which is not often realistic for engineering applications although the recent family of physics-informed operator learning methods [225] take a step towards scarce data problems.

2.3 Implementation and algorithms

In this section we outline a practical implementation of the 2-step and 1-step KEqL problems. We focus our discussion on high-level ideas and refer the reader to the SM as well as our GitHub repository³ for further details. We first present the representer theorem, which forms the backbone of our implementation in Section 2.3.1, then outline the implementation of 2-step KEqL in Section 2.3.2 followed by 1-step KEqL in Section 2.3.3. Finally, we discuss our choice of kernels and aspects of modeling prior information in Section 2.3.4.

2.3.1 Representer theorem

We start by stating the representer theorem for interpolation problems in RKHSs which directly implies the closed form solution formulae for 2-step KEqL and is fundamental to the proof of the solution formula for 1-step KEqL. For the proof the representer theorem, see SM.

Proposition 2.4 (RKHS representer theorem [166, Cor. 17.12]). *Let $\mathcal{H}$ be an RKHS with kernel $\mathbf{H}$ and let $\phi = (\phi_1, \dots, \phi_N) \in (\mathcal{H}^*)^N$ be a vector of bounded linear functionals on $\mathcal{H}$. Consider*

$$\widehat{f} := \arg \min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \quad \text{s.t.} \quad \phi_i(f) = z_i, \quad i = 1, \dots, N,$$

for $z = (z_1, \dots, z_N) \in \mathbb{R}^N$. Then every minimizer $\widehat{f}$ has the form $\widehat{f} = \mathbf{H}(\phi, \cdot)^T \widehat{\alpha}$ where $\widehat{\alpha} = \mathbf{H}(\phi, \phi)^{-1} z$. Here we followed the notation of Theorem 2.5 and wrote $\mathbf{H}(\phi_i, x) = \phi_i(\mathbf{H}(\cdot, x)) \in \mathbb{R}$, $\mathbf{H}(\phi, \cdot) = (\mathbf{H}(\phi_1, \cdot), \dots, \mathbf{H}(\phi_N, \cdot)) \in \mathcal{H}^N$, and $\mathbf{H}(\phi, \phi) \in \mathbb{R}^{N \times N}$ as the matrix with entries $\phi_j(\mathbf{H}(\phi_i, \cdot))$. Whenever $\mathbf{H}(\phi, \phi)$ is not invertible the vector $\widehat{\alpha}$ is defined in the least squares sense.

2.3.2 Implementation of 2-step KEqL

Our implementation of 2-step KEqL is straightforward and utilizes standard representer theorems for kernel regression [142, 230] that are special instance of the more general result

³<https://github.com/TADSGroup/kernelequationlearning>

in Proposition 2.4. Then the solution to (2.7) admits the analytic form

$$\hat{u}^m(y) = \mathbf{U}(Y^m, y)^T \hat{\alpha}^m, \quad \text{where} \quad \hat{\alpha}^m = (\mathbf{U}(Y^m, Y^m) + \sigma_u^2 I)^{-1} u^m(Y^m). \quad (2.14)$$

Here $\mathbf{U}(Y^m, \cdot) := (\mathbf{U}(y_1^m, \cdot), \dots, \mathbf{U}(y_N^m, \cdot)) \in \mathcal{U}^N$ is a column vector field and $\mathbf{U}(Y^m, Y^m) \in \mathbb{R}^{N \times N}$ is a kernel matrix with entries $\mathbf{U}(Y^m, Y^m)_{ij} = \mathbf{U}(y_i^m, y_j^m)$. Setting $\sigma_u = 0$ further characterizes the solution (2.6) assuming $\mathbf{U}(Y^m, Y^m)$ is invertible. Thanks to (2.14) we can directly compute,

$$L_q \hat{u}(y) = L_q \mathbf{U}(Y^m, y)^T \hat{\alpha}^m, \quad \text{where} \quad L_q \mathbf{U}(Y^m, y) = (L_q \mathbf{U}(y_1^m, y), \dots, L_q \mathbf{U}(y_N^m, y)),$$

which requires us to apply the L_q operators to the kernel $\mathbf{U}$ as a function of its second input (the y variable) for fixed values of its first input (the y_n^m values). For typical PDEs this amounts to computing partial derivatives of $\mathbf{U}$ either analytically or using automatic differentiation software.

We can solve (2.10) identically to the above. Writing $S = \cup_{m=1}^M \Phi(\hat{u}^m, Y) \subset \mathcal{S}$ we have the formula

$$\hat{P}(s) = \mathbf{P}(S, s)^T \hat{\beta}, \quad \text{where} \quad \hat{\beta} = (\mathbf{P}(S, S) + \sigma_P^2)^{-1} (f(Y) - \bar{P}(S)). \quad (2.15)$$

Here we used the familiar shorthand notation $f(Y) = (f^1(Y), \dots, f^M(Y)) \in \mathbb{R}^{MK}$, the concatenation of the vectors $f^m(Y)$, and $\bar{P}(S) \in \mathbb{R}^{MK}$, the vector of point values of $\bar{P}$ evaluated on S , both viewed as column vectors. Similar to (2.14), we also defined the vector field $\mathbf{P}(S, \cdot) := (\mathbf{P}(s_1, \cdot), \dots, \mathbf{P}(s_{MK}, \cdot)) \in \mathcal{P}^{MK}$, and the matrix $\mathbf{P}(S, S) = (\mathbf{P}(s_i, s_j))_{i,j=1}^{MK} \in \mathbb{R}^{MK \times MK}$, where the s_j 's denote the elements of the point cloud S (see also Figure 2.1(b)). In further analogy with the first step, setting $\sigma_P = 0$ gives the solution to (2.8) when $\mathbf{P}(S, S)$ is invertible.

The formulae (2.14) and (2.15) highlight the convenience and computational efficacy of 2-step KEqL since each equation requires a single linear solve involving a kernel matrix; this can be done efficiently using sparse or randomized linear algebra techniques for large problems [236–238] or off-the-shelf algorithms for smaller problems.

2.3.3 Implementation of 1-step KEqL

The implementation of 1-step KEqL is more involved and requires the solution of a compositional optimization problem. The first step is to derive an equivalent discrete formulation of (2.11) and (2.12); this amounts to the derivation of a representer theorem which we state for (2.11). Next we design an algorithm that can efficiently solve the aforementioned discrete optimization problem.

Towards stating our representer theorem, we need to introduce some new notation: Assume that the operators $\phi_k^q := \delta_{y_k} \circ L_q$ define elements of $\mathcal{U}^*$ for $q = 1, \dots, Q$ and $k = 1, \dots, K$, recalling our convention that L_1 denotes the identity map so that $\phi_k^1 = \delta_{y_k}$. Further observe that $\Phi(u, y_k) = (y_k, \phi_k^1(u), \dots, \phi_k^Q(u))$ by definition and so the ϕ_k^q denote the linear operators that give the subset of components of $\Phi(u, y_k)$ as a function of u , justifying our choice of notation. Write $\mathbf{U}(\phi_k^q, y) = \phi_k^q(\mathbf{U}(\cdot, y)) \in \mathbb{R}$ where $\mathbf{U}(\phi_k^q, \cdot) \in \mathcal{U}$, i.e., the RKHS function obtained by applying ϕ_k^q to $\mathbf{U}(\cdot, y)$ for every fixed value of y . Further define

$U(\phi_k^q, \phi_j^\ell) := \phi_j^\ell(U(\phi_k^q, \cdot))$ for $\ell = 1, \dots, Q$ and $j = 1, \dots, K$. Next define the vector field $U(\phi, \cdot) := (U(\phi_1^1, \cdot), \dots, U(\phi_K^1, \cdot), \dots, U(\phi_1^Q, \cdot), \dots, U(\phi_K^Q, \cdot)) \in \mathcal{U}^{QK}$, along with the vectors $U(\phi, \phi_k^q) = (U(\phi_1^1, \phi_k^q), \dots, U(\phi_K^1, \phi_k^q), \dots, U(\phi_1^Q, \phi_k^q), \dots, U(\phi_K^Q, \phi_k^q)) \in \mathbb{R}^{QK}$. Further let $U(\phi^q, \phi^\ell) \in \mathbb{R}^{K \times K}$ denote the matrices with entries $U(\phi^q, \phi^\ell)_{k,j} = U(\phi_k^q, \phi_j^\ell)$ and the block-matrix $U(\phi, \phi) \in \mathbb{R}^{QK \times QK}$ with blocks $U(\phi, \phi)_{q,\ell} = U(\phi^q, \phi^\ell) \in \mathbb{R}^{K \times K}$. Note that the above vectors and matrices can be computed offline using either analytical expressions or automatic differentiation akin to the 2-step KEqL. With this new notation we have the following characterization of the minimizers of (2.11).

Theorem 2.5. *Suppose that $\phi_k^q \in \mathcal{U}^*$ for $q = 1, \dots, Q$ and $k = 1, \dots, K$. Then every minimizing tuple $(\hat{\mathbf{u}}, \hat{P})$ of (2.11) can be written in the form*

$$\hat{u}^m(y) = U(\phi, y)^T \hat{\alpha}^m, \quad \hat{P}(s) = P(S(\hat{\alpha}), s)^T \hat{\beta},$$

for a tuple $(\hat{\alpha}, \hat{\beta})$ that solve the equivalent optimization problem

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}) = & \arg \min_{\alpha \in (\mathbb{R}^{QK})^M, \beta \in \mathbb{R}^{MK}} \beta^T P(S(\alpha), S(\alpha)) \beta + \frac{\lambda}{M} \sum_{m=1}^M (\alpha^m)^T U(\phi, \phi) \alpha^m, \\ \text{s.t.} \quad & U(\phi, Y^m)^T \alpha^m = u^m(Y^m), \\ & P(S(\alpha), S^m(\alpha))^T \beta = f^m(Y) - \bar{P}(S^m(\alpha^m)). \end{aligned} \quad (2.16)$$

Here we introduced the block vector $\alpha := (\alpha^1, \dots, \alpha^M) \in (\mathbb{R}^{QK})^M$ for the coefficient vectors $\alpha^m \in \mathbb{R}^{QK}$, the point clouds $S^m(\alpha^m) = \Phi(U(\phi, \cdot)^T \alpha^m, Y) \subset \mathcal{S}$ and in turn $S(\alpha) = \cup_{m=1}^M S^m(\alpha^m)$, as well as the matrices $U(\phi, Y^m) \in \mathbb{R}^{QK \times N}$ with columns $U(\phi, y_n^m)$, and $P(S(\alpha), S^m(\alpha^m)) \in \mathbb{R}^{MK \times K}$ with columns $P(S, s_k^m)$.

The proof technique is based on the observation that conditioned on the values $L_q v^m(y_k)$, the optimization problems for the v^m and G variables in (2.11) will decouple. Then the result is obtained by introducing auxiliary variables representing the $L_q v^m(y_k)$ and applying the representer theorem proposition 2.4; see SM for proof.

The problem (2.16) is the key to the design of our algorithms as it is readily implementable without the need for further discretization. We can further relax the equality constraints in (2.16) to obtain the analogous problem to (2.12):

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}) = & \arg \min_{\alpha \in (\mathbb{R}^{QK})^M, \beta \in \mathbb{R}^{MK}} \beta^T P(S(\alpha), S(\alpha)) \beta + \sum_{m=1}^M \frac{\lambda}{M} (\alpha^m)^T U(\phi, \phi) \alpha^m \\ & + \frac{1}{2\sigma_u^2} \|U(\phi, Y^m)^T \alpha^m - u^m(Y^m)\|_2^2 \\ & + \frac{1}{2\sigma_P^2} \|P(S(\alpha), S^m(\alpha^m))^T \beta + \bar{P}(S^m(\alpha^m)) - f^m(Y)\|_2^2. \end{aligned} \quad (2.17)$$

Remark 2.6. Directly solving the 1-step KEqL objective can be numerically unstable because small changes in the coefficients defining the auxiliary functions v^m may produce large changes in their derivatives and hence in the collocation points $S^m(\alpha^m)$. In practice,

we implement an intermediate method called the reduced 1-step method that interpolates between our 1-step and 2-step methods, and serves to improve computational efficiency and scalability for larger problems and data sets. We solve this reduced 1-step problem using a damped Levenberg-Marquardt scheme based on local quadratic approximations of the residual term. See the SM for details. $\diamond$

2.3.4 Choosing kernels and hyper-parameters

While the 1- and 2-step KEqL are generic and well-defined for any choice of $\mathbf{P}$ and $\mathbf{U}$ (assuming sufficient regularity), their practical performance is closely tied to a good choice of kernels as is typical for RKHS/GP methods. Broadly speaking, the choice of these kernels imposes constraints on our model for the functions $\hat{u}^m$ and the learned PDE $\hat{P}$. Moreover, in many applications we may have access to expert knowledge about the u^m and the type of PDE at hand and we may want to incorporate these into our algorithm. A standard example is periodic boundary conditions or invariance of the solution under certain operators and group actions. For the benchmarks in Section 2.4, we specify the choices of kernels for $\mathbf{U}$ and $\mathbf{P}$ in each experiment. Further details about these kernels are provided in the SM.

2.4 Experiments

We now present a series of numerical experiments that demonstrate the performance of our 1-step and 2-step methodologies for estimating the u^m , learning $\mathcal{P}$, and the solution operator $\mathcal{P}^{-1}$. For benchmarking we compare our algorithms with SINDy [138], as a 2-step dictionary based method, and PINN-SR [160], as a 1-step neural net based method. We will focus on the key results for each experiment and postpone additional results and setup details to SM.

Common structure and performance metrics are discussed in Section 2.4.1. A Duffing oscillator is studied in Section 2.4.2 followed by Burgers' PDE in Section 2.4.3 and Darcy's flow elliptic PDE in Section 2.4.4.

2.4.1 Common structure

We will consider scalar valued DEs of the form (2.4) Φ assumed to be known. For dynamic problems we assume $\mathcal{Y} = (0, T] \times \Omega$ for some $T > 0$ and a spacial domain Ω and take $\bar{P}(\Phi(u, y)) = \bar{P}(\Phi(u, (t, x))) = \partial_t^h u(t, x)$ to be known (the order h will be clarified for each example). We often use RKHSs $\mathcal{U}$ that do not automatically satisfy the boundary conditions of the PDEs so we use additional observation points on the boundaries to impose those conditions. To this end, we write $N_{\mathcal{Y}}$ to denote the number of observation points in the interior of $\mathcal{Y}$ while we use $N_{\partial\mathcal{Y}}$ to denote the number of observation points on the boundary, so that the total number of observation points is always $N = N_{\mathcal{Y}} + N_{\partial\mathcal{Y}}$. For all experiments we consider training data of the form $(u^m(Y^m), f^m)_{m=1}^M$. In addition to the collocation points Y and the observation points Y^m , we also consider an independent test mesh $Y_{\text{test}} \subset \mathcal{Y}$ which will be used for reporting errors.

We will consider three types of test errors:

- The *state reconstruction error* quantifies the relative accuracy of the $\hat{u}^m$ on the training data,

$$\mathcal{R}_{\text{rec}}((u^m)_{m=1}^M) := \frac{1}{M} \sum_{m=1}^M \frac{\|u^m(Y_{\text{test}}) - \hat{u}^m(Y_{\text{test}})\|_2^2}{\|u^m(Y_{\text{test}})\|_2^2}.$$

- The *equation learning error* quantifies the relative accuracy of $\hat{P}$ over a set of functions $\mathcal{W}$

$$\mathcal{R}_{\text{eq}}(\mathcal{W}) := \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \frac{\|\mathcal{P}(w)(Y_{\text{test}}) - \widehat{\mathcal{P}}(w)(Y_{\text{test}})\|_2^2}{\|\mathcal{P}(w)(Y_{\text{test}})\|_2^2},$$

where $|\mathcal{W}|$ denotes the cardinality of $\mathcal{W}$.

- Similarly, the *operator learning error* quantifies the error of the map $\widehat{\mathcal{P}}^\dagger$,

$$\mathcal{R}_{\text{opl}}(\mathcal{W}) := \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \frac{\|\mathcal{P}^{-1}(w)(Y_{\text{test}}) - \widehat{\mathcal{P}}^\dagger(w)(Y_{\text{test}})\|_2^2}{\|\mathcal{P}^{-1}(w)(Y_{\text{test}})\|_2^2}.$$

We highlight that here $\mathcal{W}$ constitutes initial or boundary conditions as well as source terms depending on the problem at hand.

For different experiments we report errors with $\mathcal{W}$ taken to be the training data (training), an in-distribution test set (ID), or out-of-distribution test sets (OOD).

2.4.2 The Duffing oscillator

In this example we compared the performance of 1-step and 2-step KEqL along with SINDy for learning a 1D nonlinear ODE. Our focus is on state reconstruction and forecasting performance.

Problem setup

Consider $\mathcal{Y} = (0, 50)$ and the Duffing ODE

$$\begin{cases} \mathcal{P}(u) = \partial_t^2 u - 3u + 3u^3 + 0.2\partial_t u = \cos(2t), & t \in \mathcal{Y}, \\ u(0) = \partial_t u(0) = 0. \end{cases} \quad (2.18)$$

To generate the training data we solved the ODE using a standard adaptive solver. The numerical solution was then subsampled on the observation points that were also picked on a uniform grid; see Figure 2.2(a). We chose $M = 1$, and so we aim to learn the ODE and its solution from observations of a single trajectory. The test data sets were generated similarly but on a finer uniform mesh with different initial conditions.

Algorithm setup

We chose the collocation points Y to be a fine uniform grid in $\mathcal{Y}$. We further chose the operators $L_1 = \text{Id} : u \mapsto u$, $L_2 : u \mapsto \partial_t u$, and $\bar{P}(u) = \partial_t^2 u$. We chose $\mathbf{U}$ to be a rational quadratic kernel and $\mathbf{P}$ to be RBF. To approximate the derivatives in SINDy we used the same kernel interpolant as in 2-step KEqL. As for the SINDy dictionary, we chose all polynomial terms of up to third degree in the variables $\{1, u, \partial_t u\}$; we empirically found this dictionary to give the best training and test errors for SINDy.

Results

We summarize our key findings for this example in Figure 2.2. In Figure 2.2(a) we visually compare the quality of the reconstructed solution $\hat{u}$ for 1-step and 2-step KEqL. We clearly see that 1-step KEqL is superior in state reconstruction the solution. This visual performance is further confirmed in Figure 2.2(b) where we observe an order of magnitude improvement between 1-step and 2-step methods in terms of relative errors.

For our next test we considered simulating the learned ODE with new initial conditions, this is essentially an extrapolation problem where we try to predict the state of the system outside of the window of observations. We show three examples of random initial conditions in Figure 2.2(c), comparing the solutions computed using 1- and 2-step KEqL and SINDy. We observe that the 1-step method generally tracks the true solution for a longer time interval although all solutions eventually diverge. The superior performance of 1-step KEqL is also evident in panel (b) where we used $\mathcal{R}_{\text{opl}}$ to denote the pointwise error between the extrapolated solutions and the true states. Once again we observe an order of magnitude improvement between 1-step and 2-step with smaller gaps for larger time windows as expected.

2.4.3 The Burgers' PDE

In this example we compared 1-step KEqL with SINDy (as a 2-step competitor) and PINN-SR (as a 1-step competitor) for the 1D Burgers' PDE. Our focus is on equation and operator learning from scarce observations of a single trajectory.

Problem setup

We consider $\mathcal{Y} = (0, 1] \times (0, 1)$ and the Burgers' PDE

$$\begin{cases} \mathcal{P}(u) = \partial_t u + \vartheta u \partial_x u - \nu \partial_x^2 u = 0 \text{ for } (t, x) \in \mathcal{Y}, \\ u(0, x) = u_0(x), \text{ and } u(t, 0) = u(t, 1) = 0. \end{cases} \quad (2.19)$$

In all experiments we took $M = 1$ so that a single solution is observed on a scarce set of observations. These solutions were generated by prescribing the initial conditions u_0 and then solving the PDE using a traditional numerical solver on a fine mesh. The observed data was then subsampled from a set of Chebyshev collocation points that were also used in the implementation of all algorithms as the collocation mesh Y .

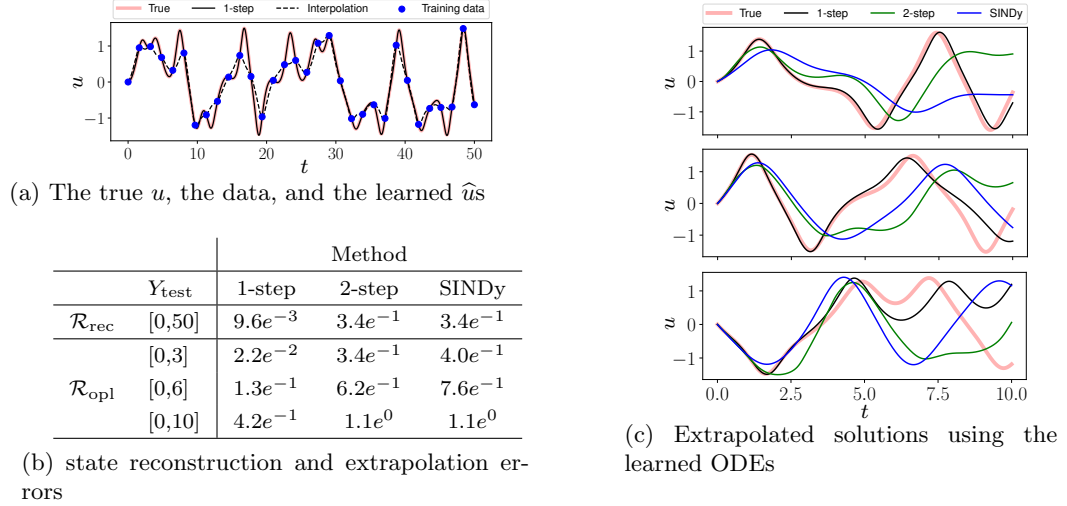


Figure 2.2: Representative numerical results for the Duffing oscillator (2.18): (a) Shows the training data and the ground truth state of the system u in comparison to the reconstructed state $\hat{u}$ using 1-step KEqL and 2-step methods; (b) Quantitative values of relative state reconstruction and operator learning errors, the latter is reported for different time windows. These values were averaged over three random initial conditions; (c) Visualization of three extrapolated dynamics used to compute $\mathcal{R}_{\text{opl}}$.

Algorithm setup

We took $L_1 : u \mapsto u$, $L_2 : u \mapsto \partial_x u$, $L_3 : u \mapsto \partial_x^2 u$ and also $\bar{P}(u) = \partial_t u$. $\mathbf{U}$ was taken to be RBF and $\mathbf{P}$ was a polynomial kernel of degree 2. To approximate the derivatives for SINDy we used the same RBF kernel used in KEqL and the dictionary terms were polynomial features $\{1, u, \partial_x u, \partial_x^2 u\}$; we found that this dictionary gave the best results for SINDy. For the PINN-SR method we used the same dictionary for P along with a feed-forward neural network with 8 layers and width 20 to approximate u .

Results

We begin by considering the state reconstruction and equation learning errors for one-shot learning of Burgers' as shown in Figure 2.3(a) where we are reporting the errors for a fixed smooth initial condition with different number of observation points. Note, these errors were computed over the training data but they are different from the training error of the algorithm as they compare the reconstructed solution and the learned PDE on the test mesh Y_{test} . Since the observation points are random we are reporting the average errors along with the best and worst errors over different runs giving the shaded regions. We found that 1-step KEqL leads to the best state reconstruction errors with a wider performance gap when N_y is small. Interestingly, the performance gap appears to be more pronounced in the equation learning case where we observed an order of magnitude improvement over

SINDy and almost two orders improvement over PINN-SR. We observed significantly larger errors for PINN-SR here which we attribute to difficulties in tuning and optimizing PINN models. In fact, we could often bring these errors down by hand tuning the algorithm in each instance of the experiments but automatic strategies proposed in [160] appeared to be very sensitive to the location of observation points.

This issue was exacerbated further with random initial conditions as depicted in Figure 2.3(b). We report the average errors along with the best and worst errors as shaded areas as before. A similar conclusion can be drawn here as in Figure 2.3(a), where 1-step KEqL continues to achieve the lowest state reconstruction and equation learning errors, with a larger performance gap when N_Y is small. However, we now observe that the KEqL and SINDy methods exhibit greater robustness across different experiments, whereas the performance of the PINN-SR deteriorates significantly. This decline is reflected in larger shaded regions and the larger average error. The main takeaway here is that KEqL and SINDy appear to be very robust to the choice of hyperparameters where as PINN-SR needs tailored hyperparameter tuning for every single run.

In Figure 2.3(c,d) we show two enlightening examples that highlight the performance gap of 1-step and 2-step methods. In (c) we hand picked an initial condition that leads to a solution with multiple (near) shocks. We see that despite scarcity of observations, 1-step KEqL and PINN-SR are able to capture the general shape of the solution while 2-step methods are significantly worse. In (d) we show an instance where we only observed the value of the solution on the boundary of the domain while in row (e) we solved the learned equation for a new initial condition. Once again we see that 1-step methods give better performance for state reconstruction, however, when solving the learned PDE a noticeable performance gap appears across the three methods with 1-step KEqL giving the best solution. This example further reflects the performance gaps observed in panel (a).

2.4.4 Darcy's flow PDE

Here we performed a systematic study of the performance of 1-step and 2-step KEqL for learning an elliptic PDE with a variable diffusion coefficient. In particular, we investigate the ID and OOD performance in terms of state reconstruction, equation learning, and operator learning. Our goal here was to push the algorithm in terms of scalability to a large number of solution and right hand side pairs (large M) and very scarce observations (very small N_Y).

Problem setup

Take $\mathcal{Y} = (0, 1) \times (0, 1)$ and consider the PDE

$$\begin{cases} \mathcal{P}(u) = \operatorname{div}(a \nabla u) = f, & \text{in } \mathcal{Y}, \\ u(y) = g(y), & \text{for } y \in \partial \mathcal{Y}, \end{cases} \quad (2.20)$$

where $a(y) = \exp(\sin(\cos(y_1) + \cos(y_2)))$. The training data was generated by drawing u from a smooth GP and setting $f = \mathcal{P}(u)$ and setting value of $g(y) = u(y)$ on $\partial \mathcal{Y}$ as the boundary condition. Each u was then subsampled on the observation points that were

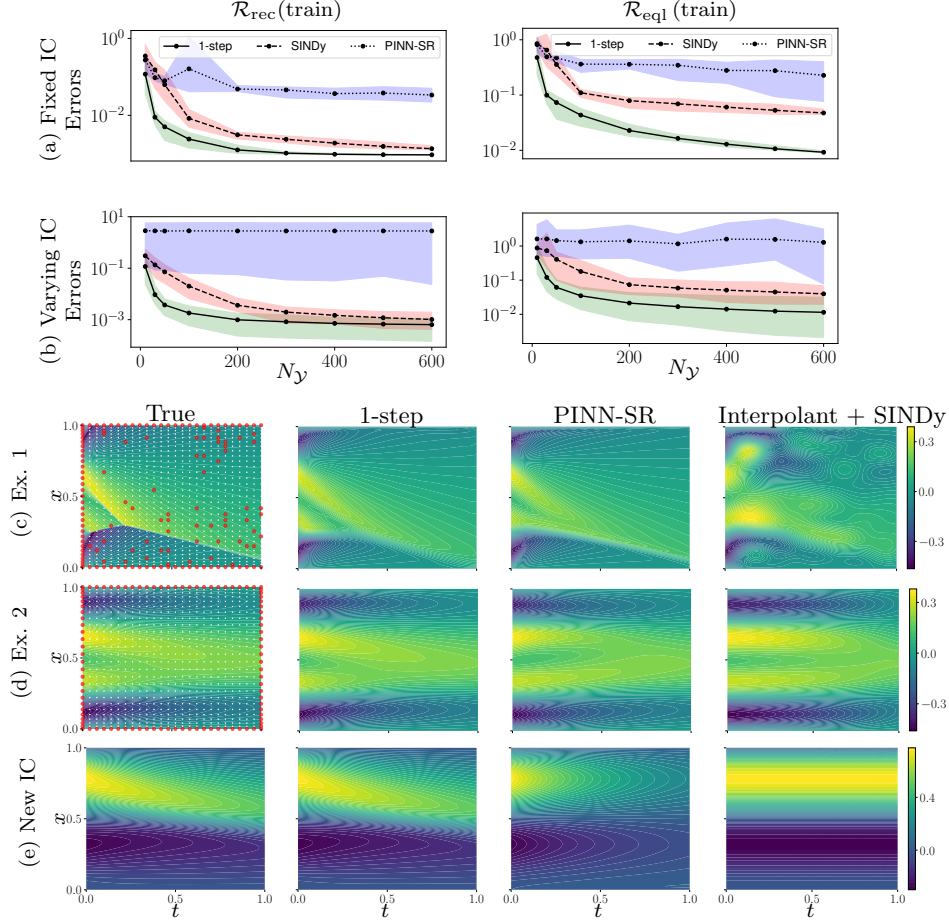


Figure 2.3: Representative numerical results for Burgers' PDE (2.19): (a) State reconstruction and equation learning errors computed on the training functions for 1-step KEqL, kernel interpolant + SINDy, and PINN-SR using $M = 1$ training pairs and varying the number of interior observations N_Y ; (b) Similar experiment as panel (a) but with randomized initial conditions; (c) State reconstruction results from scarce data using parameters that lead to a near shock comparing 1-step KEqL, PINN-SR and a kernel interpolation; (d) Similar setup to row (c) with a smooth solution and observations only on the boundary; (e) Computed solutions to the PDEs identified by 1-step KEqL, PINN-SR, and SINDy from the data in row (d) for a new initial condition.

picked randomly in the interior of a uniform collocation grid Y . The test data sets were generated similarly with the OOD data drawn from a GP with a different length scale.

Algorithm setup

We picked the operators $L_1 : u \mapsto u$, $L_2 : u \mapsto \partial_{y_1} u$, $L_3 : u \mapsto \partial_{y_1}^2 u$, $L_4 : u \mapsto \partial_{y_2} u$, $L_5 : u \mapsto \partial_{y_2}^2 u$, $L_6 : u \mapsto \partial_{y_1 y_2} u$ and $\bar{P} \equiv 0$. We used $\mathbf{U} = \mathbf{U}_{\text{RBF}}$ for learning u and $\mathbf{P}$ was a hybrid polynomial kernel as the product of an RBF for the spatial variables and a polynomial kernel of first degree for the rest of the variables. For all experiments we

employed the 1-step KEqL along with additional sparse numerical linear algebra tricks (see SM) to scale the algorithms to large training data sizes.

Results

Figure 2.4 summarizes various training and test errors for this example focusing on equation learning and operator learning errors. The black lines represent the average errors, while the shaded regions indicate the range of errors, spanning from the worst to the best across multiple runs; see SM for more details.

We observed that the 1-step method consistently outperformed the 2-step method across all tasks both ID and OOD. Most notably, in very scarce observation regime (only $N_Y = 2$ interior observations per function) we see an order of magnitude performance gap between 1-step and 2-step methods across the board. As expected, the gap reduces as we increase the observations N_Y and the size of the training data M . Finally, we note that the operator learning errors follow the same trends as the equation learning errors which suggests that operator learning should inherit similar rates as equation learning.

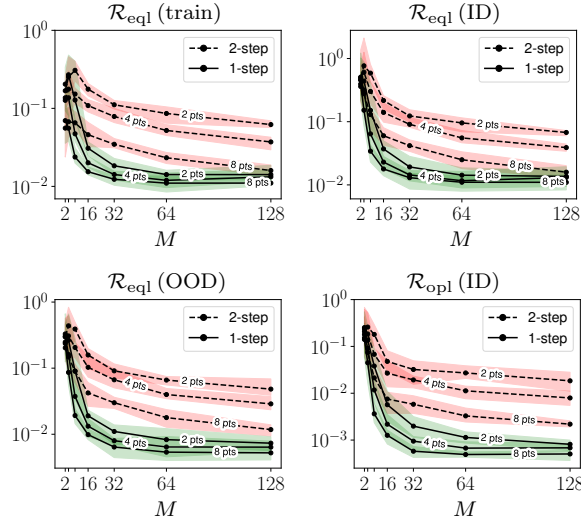


Figure 2.4: Representative numerical results for Darcy’s flow PDE (2.20): The top two figures from the left show the equation learning errors computed over training and ID test sets. The OOD test errors are shown in the next row along with the ID operator learning error. The 1-step KEqL was used and the labels on the graphs denote the number of interior observation points N_Y .

2.5 Theoretical analysis

In this section we turn our attention to the theoretical analysis of KEqL and in particular obtaining quantitative error bounds for the functions $\hat{u}^m$ as well as the operators $\hat{\mathcal{P}}$ and $\hat{\mathcal{P}}^\dagger$. We begin with a brief review of results from RKHS theory and Sobolev spaces needed for our analysis in Section 2.5.1. State reconstruction and Equation learning errors for

KEqL are presented in Section 2.5.2. The subsequent operator learning error analysis is summarized in Section 2.5.3. We omit some repetitive proofs and arguments to save space and collect them in the SM.

2.5.1 Preliminaries

We begin by collecting useful preliminary results relating to the theory of RKHS spaces and Sobolev spaces.

RKHS spaces We assume familiarity with the standard results in the theory of RKHSs; these results can be found in many classic texts such as [166, 229, 230, 239]. We are particularly interested in a scaled characterization of RKHS via Mercer's theorem, which allows us to define the nested ladder of RKHS spaces.

Proposition 2.7. *Suppose $\Omega \subset \mathbb{R}^q$ is bounded and let $\mathbf{H} : \Omega \times \Omega \rightarrow \mathbb{R}$ be a PDS kernel that is continuous in both of its arguments on $\bar{\Omega}$. Then there exists an orthonormal set of continuous eigenfunctions $\{e_i\}_{i=1}^\infty \subset L^2(\Omega)$ and decreasing eigenvalues $\{\lambda_i\}_{i=1}^\infty$, $\lambda_1 \geq \lambda_2 \geq \dots$, such that $\mathbf{H}(x, x') = \sum_{i=1}^\infty \lambda_i e_i(x) e_i(x')$, and for $\gamma > 0$ the scaled RKHS $\mathcal{H}^\gamma$ can be characterized as*

$$\mathcal{H}^\gamma = \left\{ f : \Omega \rightarrow \mathbb{R} \left| f(x) = \sum_{i \in \{i | \lambda_i \neq 0\}} c_i(f) e_i(x), \quad \sum_{i \in \{i | \lambda_i \neq 0\}} \lambda_i^{-\gamma} c_i(f)^2 < +\infty \right. \right\}$$

and for any pair $f, f' \in \mathcal{H}^\gamma$ we have $\langle f, f' \rangle_{\mathcal{H}^\gamma} = \sum_{i \in \{i | \lambda_i \neq 0\}} \lambda_i^{-\gamma} c_i(f) c_i(f')$. In particular, the RKHS $\mathcal{H}$ is recovered as $\mathcal{H} = \mathcal{H}^1$.

These are precisely the RKHSs corresponding to the kernels $\mathbf{H}^\gamma(x, x') := \sum_{i=1}^\infty \lambda_i^\gamma e_i(x) e_i(x')$. Naturally, larger values of γ imply more "smoothness", in particular we have the inclusion $\mathcal{H}^{\gamma_2} \subset \mathcal{H}^{\gamma_1}$ for $1 \leq \gamma_1 < \gamma_2$ following Hölder's inequality. Observe that our definition of the $\mathcal{H}^\gamma$ resembles the spectral definition of Sobolev spaces $H^\gamma(\Omega)$ on compact sets in which case the $\{\lambda_i, e_i\}_{i=1}^\infty$ can be taken as the eigenpairs of the inverse of the Laplacian operator. The following lemma is useful in our proofs later in this section; for the proof of the lemma, see SM.

Lemma 2.8. *Suppose $f \in \mathcal{H}^{2\gamma}$ and $f' \in \mathcal{H}^\gamma$. Then $\langle f', f \rangle_{\mathcal{H}^\gamma} \leq \|f'\|_{L^2(\Omega)} \|f\|_{\mathcal{H}^{2\gamma}}$.*

Proof. For simplicity of notation let us assume that $\lambda_i \neq 0$ so that the kernel $\mathbf{H}$ is strictly PDS (i.e., non-degenerate). By definition of the $\mathcal{H}^\gamma$ inner product, and using Cauchy-Schwartz, we have that $\langle f', f \rangle_{\mathcal{H}^\gamma} = \sum_{i=1}^\infty \lambda_i^{-\gamma} c_i(f') c_i(f) \leq (\sum_{i=1}^\infty c_i(f')^2)^{1/2} (\sum_{i=1}^\infty \lambda_i^{-2\gamma} c_i(f)^2)^{1/2} = \|f'\|_{L^2(\Omega)} \|f\|_{\mathcal{H}^{2\gamma}}$. $\square$

Throughout the remainder of the theoretical analysis section, whenever we write $\mathcal{H}^\gamma$ for $\gamma \in \mathbb{N}$, we mean the nested RKHS, not the product space.

Sobolev spaces We present some essential background on Sobolev spaces, including the version of the Sobolev embedding theorem used in our proofs in SM. In addition to the embedding theorem we will heavily rely on the Sobolev sampling inequality which gives control over the Sobolev norm of a function that is small or zero on a scattered set of points. For an introduction to Sobolev sampling inequalities and related results on scattered data approximation see [240]. The following theorem is a distillation of [241, Prop. 2.4] in the form that we need in our proofs; see also [242–244].

Proposition 2.9 (Sobolev sampling inequality). *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded set with Lipschitz boundary and consider a set of points $X = \{x_1, \dots, x_N\} \subset \overline{\Omega}$ with fill distance $h_X := \sup_{x \in \Omega} \inf_{x' \in X} \|x - x'\|_2$. Let $u|_X$ denote the restriction of u to the set X . Further consider $\gamma > d/2$ and $0 \leq \eta \leq \gamma$ and let $u \in H^\gamma(\Omega)$.*

- (a) (Noiseless) *Suppose $u|_X = 0$. Then there exists $h_0 > 0$ so that whenever $h_X \leq h_0$ we have $\|u\|_{H^\eta(\Omega)} \leq C_\Omega h_X^{\gamma-\eta} \|u\|_{H^\gamma(\Omega)}$, where $C_\Omega > 0$ is a constant that depends only on Ω .*
- (b) (Noisy) *Suppose $u|_X \neq 0$. Then there exists $h_0 > 0$ so that whenever $h_X \leq h_0$ we have the inequality $\|u\|_{L^\infty(\Omega)} \leq C_\Omega h_X^{\gamma-d/2} \|u\|_{H^\gamma(\Omega)} + 2\|u|_X\|_\infty$, where $C_\Omega > 0$ is a constant that depends only on Ω .*

2.5.2 Error analysis for equation learning

We now turn our attention to error bounds for $\hat{u}^m$ and $\hat{P}$, i.e., the state reconstruction and equation learning errors. We will present our proofs for 1-step KEqL assuming that there is no noise in the data or nugget terms in the formulation. We give further theoretical results in the SM including a simpler convergence proof for 1-step KEqL that can generalize beyond the RKHS formulation as well as an analogous error bound for 2-step KEqL.

Let us begin by presenting our main assumptions that will be used in the rest of this section. First we have a standard assumption on the set $\mathcal{Y}$ on which the PDE is defined. This assumption allows us to use Sobolev embedding theorem and sampling inequalities.

Assumption 2.10. The set $\mathcal{Y} \subset \mathbb{R}^d$ is bounded and has Lipschitz boundary.

Next we will need assumptions on the kernels $\mathbf{U}$ and $\mathbf{P}$ to ensure sufficient regularity and continuous embedding in appropriate Sobolev spaces:

Assumption 2.11. The kernel $\mathbf{U} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and its corresponding RKHS satisfy:

- (i) $\mathbf{U}$ is PDS and continuous in its arguments.
- (ii) $\mathcal{U}$ is continuously embedded in $H^\gamma(\mathcal{Y})$ for some $\gamma > d/2 + \text{order}(\mathcal{P})$. In particular, there exists $C_{\mathcal{Y}} > 0$ such that $\|u\|_{H^\gamma(\mathcal{Y})} \leq C_{\mathcal{Y}} \|u\|_{\mathcal{U}}$ for all $u \in \mathcal{U}$.
- (iii) $\mathcal{B}(u) = g$ on $\partial\mathcal{Y}$ for all $u \in \mathcal{U}$, i.e., the boundary conditions of (2.3) hold.

Remark 2.12. We highlight that Assumption 2.11(iii) allows us to simplify our theoretical arguments significantly since we do not need to approximate values of the functions on the boundary. However, this assumption can be removed by adding a separate approximation result for the value of estimated functions $\hat{u}^m$ near the boundary under sufficient regularity assumptions. $\diamond$

Assumption 2.13. The kernel $P : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ and its corresponding RKHS satisfy:

- (i) P is PDS.
- (ii) $\mathcal{P}$ is continuously embedded in $H^\eta(\mathcal{S})$ for some $\eta > \frac{Q+d}{2}$.
- (iii) Elements of $\mathcal{P}$ are locally Lipschitz, i.e., for any compact set $B \subset \mathcal{S}$ there exists a constant $C(B) > 0$ so that $|P(s) - P(s')| \leq C(B)\|P\|_{\mathcal{P}}\|s - s'\|_2$, for all $s, s' \in B$.

Since we will be taking the limit of N (number of observation points) and M (size of training data) to infinity, we will supplement our notations in this section by subscripting pertinent objects with M, N indices. In particular we will write $\hat{u}_{M,N}^m$ to denote the KEqL estimators for u^m with N observations on the points Y_N^m and M training pairs as well as $\hat{P}_{M,N}$ for the KEqL estimator of P . We further write $Y_{M,N}$ to denote the collocation mesh that was used to compute these estimators. With this notation, we further define the fill distances

$$\rho_{m,N} = \sup_{y \in \mathcal{Y}} \inf_{y' \in Y_N^m} \|y - y'\|_2, \quad \varrho_{M,N}(B) = \sup_{s \in B} \inf_{s' \in S \cap B} \|s - s'\|_2,$$

where $S := \bigcup_{m=1}^M \bigcup_{y \in Y_N^m} \Phi(u^m, y)$. We additionally define $\rho_{M,N} := \sup_{1 \leq m \leq M} \rho_{m,N}$.

Error analysis for 1-step KEqL

The following theorem gives a quantitative error bound for 1-step KEqL and constitutes our main theoretical contribution.

Theorem 2.14. *Consider the problem (2.11) with $M, N \in \mathbb{N}$. Suppose Assumptions 2.10, 2.11, and 2.13 hold and $P, \bar{P} \in \mathcal{P}$. Consider pairs $\{u^m, f^m\}_{m=1}^M$ satisfying (2.3), and a bounded set $B \subset \mathcal{S}$ with Lipschitz boundary. Then there exist constants $\rho_0, \varrho_0(B) \in (0, 1)$ so that whenever $\rho_{m,N} < \rho_0$ and $\varrho_{M,N}(B) < \varrho_0(B)$ it holds that:*

- (a) *If $u^m \in \mathcal{U}$ then it holds for $0 \leq \gamma' < \gamma$ and a constant $C > 0$, depending only on $\mathcal{Y}$, that*

$$\frac{1}{M} \sum_{m=1}^M \|\hat{u}_{M,N}^m - u^m\|_{H^{\gamma'}(\mathcal{Y})}^2 \leq C \rho_{M,N}^{2(\gamma - \gamma')} \left(\|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}}^2 \right).$$

- (b) *If $u^m \in \mathcal{U}^2$ then it holds for $d/2 + \text{order}(\mathcal{P}) < \gamma' < \gamma$ and a constant $C > 0$, depending only on $\mathcal{Y}$ and B , that*

$$\begin{aligned} & \|P - \hat{P}_{M,N}\|_{L^\infty(B)} \\ & \leq C \left(\varrho_{M,N}(B)^{\eta - \frac{Q+d}{2}} + \rho_{M,N}^{\gamma - \gamma'} \right) \left(\|P\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2}. \end{aligned}$$

We observe that both of our bounds closely resemble the Sobolev sampling inequalities from Proposition 2.9. Indeed, statement (a) is natural in that light since we have access to pointwise observations of the u^m on the sets Y_N^m . The only technicality in the proof of this result is that the $\hat{u}_{M,N}^m$ are not obtained by solving individual optimal recovery problems but rather a joint problem for all of the u^m as well as P . This is why the sum of those RKHS norms appears on the right hand side. The proof of statement (b) is more complex since our recovery problem for P is polluted by two sources of error: (1) error due to limited observations of the PDE on a collocation mesh and (2) errors due to approximation of u^m with $\hat{u}_{M,N}^m$ that are then plugged into P via the operator Φ . Essentially, we are dealing with a regression problem where the input and output of the function are noisy. This is why we need to assume that elements of $\mathcal{P}$ are Lipschitz in Assumption 2.13.

Proof. For simplicity we will present the proof with the regularization parameter $\lambda = 1$. Other choices will only affect the constant C in the bounds.

(a) Observe that the equality constraints $\hat{u}_{M,N}^m(Y_N^m) = u^m(Y_N^m)$ in (2.11) (with our new supplemented notation) simply state that the $\hat{u}_{M,N}^m$ are interpolating the u^m and so we can apply Proposition 2.9(a). Indeed, directly applying that result followed by Assumption 2.11(ii) we obtain, for $\rho_{m,N} < \rho_0$, the bounds

$$\|\hat{u}_{M,N}^m - u^m\|_{H^{\gamma'}(\mathcal{Y})} \leq C \rho_{m,N}^{\gamma-\gamma'} \|\hat{u}_{M,N}^m - u^m\|_{H^\gamma(\mathcal{Y})} \leq C \rho_{m,N}^{\gamma-\gamma'} \|\hat{u}_{M,N}^m - u^m\|_{\mathcal{U}},$$

with the constant $C > 0$ changing from one inequality to the next and depending only on the domain $\mathcal{Y}$ and the choice of $\mathcal{U}$ and γ . Using the inequality above, the identity $(a+b)^2 \leq 2(a^2 + b^2)$, and the optimality of $\hat{u}_{M,N}^m$, we can write

$$\begin{aligned} \frac{1}{M} \sum_{m=1}^M \|\hat{u}_{M,N}^m - u^m\|_{H^{\gamma'}(\mathcal{Y})}^2 &\leq \frac{C}{M} \rho_{M,N}^{2(\gamma-\gamma')} \left(\sum_{m=1}^M \|\hat{u}_{M,N}^m\|_{\mathcal{U}}^2 + \sum_{m=1}^M \|u^m\|_{\mathcal{U}}^2 \right), \\ &\leq C \rho_{M,N}^{2(\gamma-\gamma')} \left(\|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}}^2 \right). \end{aligned}$$

This concludes the proof of the first inequality.

(b) Fix M and for each u^m define the optimal interpolant

$$\bar{u}_N^m := \arg \min_{v \in \mathcal{U}} \|v\|_{\mathcal{U}} \quad \text{s.t.} \quad v(Y_N^m) = u^m(Y_N^m). \quad (2.21)$$

By optimality, we have the bound $\|\hat{u}_{M,N}^m\|_{\mathcal{U}} \geq \|\bar{u}_N^m\|_{\mathcal{U}}$. At the same time, Proposition 2.4 gives $\bar{u}_N^m = \mathbf{U}(Y_N^m, \cdot)^T \mathbf{U}(Y_N^m, Y_N^m)^{-1} u^m(Y_N^m)$ and a direct calculation using the RKHS reproducing property, i.e. $\langle u, \mathbf{U}(x, \cdot) \rangle_{\mathcal{U}} = u(x)$ for all $u \in \mathcal{U}$, implies that $\langle u^m - \bar{u}_N^m, \bar{u}_N^m \rangle_{\mathcal{U}} = 0$, which in turn gives the identity $\|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2 = \|u^m\|_{\mathcal{U}}^2 - \|\bar{u}_N^m\|_{\mathcal{U}}^2$.

Thus we obtain the lower bound

$$\|\hat{u}_{M,N}^m\|_{\mathcal{U}}^2 \geq \|u^m\|_{\mathcal{U}}^2 - \|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2. \quad (2.22)$$

On the other hand the optimality of $\hat{P}_{M,N}$ and the $\hat{u}_{M,N}^m$ gives

$$\|\hat{P}_{M,N}\|_{\mathcal{P}}^2 \leq \|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M (\|u^m\|_{\mathcal{U}}^2 - \|\hat{u}_{M,N}^m\|_{\mathcal{U}}^2),$$

which together with (2.22) yields

$$\|\hat{P}_{M,N}\|_{\mathcal{P}}^2 \leq \|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2. \quad (2.23)$$

Applying Lemma 2.8 (thanks to the assumption that $u^m \in \mathcal{U}^2$) then gives the bound $\|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2 \leq \|u^m - \bar{u}_N^m\|_{L^2(\mathcal{Y})} \|u^m\|_{\mathcal{U}^2}$. Further applying Proposition 2.9(a) (since $\bar{u}_N^m$ interpolates u^m) to control $\|u^m - \bar{u}_N^m\|_{L^2(\mathcal{Y})}$ we obtain the bound $\|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2 \leq C \rho_{m,N}^\gamma \|u^m\|_{\mathcal{U}} \|u^m\|_{\mathcal{U}^2}$. Substituting into (2.23) and using the embedding $\mathcal{U}^2 \subset \mathcal{U}$ gives

$$\|\hat{P}_{M,N}\|_{\mathcal{P}}^2 \leq \|P\|_{\mathcal{P}}^2 + \frac{C}{M} \sum_{m=1}^M \rho_{m,N}^\gamma \|u^m\|_{\mathcal{U}} \|u^m\|_{\mathcal{U}^2} \leq \frac{C}{M} \sum_{m=1}^M \rho_{m,N}^\gamma \|u^m\|_{\mathcal{U}^2}^2. \quad (2.24)$$

Now we turn our attention to controlling the error between $\hat{P}_{M,N}$ and P . Observe that the PDE constraint in (2.11) implies that, for all k and m ,

$$(\hat{P}_{M,N} + \bar{P})(s_k(\hat{u}_{M,N}^m)) = (P + \bar{P})(s_k(u^m)),$$

where we recall $s_k(v) = \Phi(v, y_k)$ is linear in v . Thus, subtracting $(\hat{P}_{M,N} + \bar{P})(s_k(u^m))$ from both sides of the above equation we obtain the bound

$$\begin{aligned} & \left| (P - \hat{P}_{M,N})(s_k(u^m)) \right| \\ &= \left| \hat{P}_{M,N}(s_k(\hat{u}_{M,N}^m)) - \hat{P}_{M,N}(s_k(u^m)) + \bar{P}(s_k(\hat{u}_{M,N}^m)) - \bar{P}(s_k(u^m)) \right| \\ &\leq \left| \hat{P}_{M,N}(s_k(\hat{u}_{M,N}^m)) - \hat{P}_{M,N}(s_k(u^m)) \right| + \left| \bar{P}(s_k(\hat{u}_{M,N}^m)) - \bar{P}(s_k(u^m)) \right|. \end{aligned}$$

Now consider indices k, m such that $s_k(u^m)$ and $s_k(\hat{u}_{M,N}^m)$ belong to B . Then by the Lipschitz assumption on $\hat{P}_{M,N}$ and $\bar{P}$ (i.e., Assumption 2.13(iii)) we have that

$$\left| (P - \hat{P}_{M,N})(s_k(u^m)) \right|^2 \leq C(B) \left(\|\hat{P}_{M,N}\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 \right) \|s_k(u^m - \hat{u}_{M,N}^m)\|_2^2.$$

Applying the bound (2.24) to control $\|\hat{P}_{M,N}\|_{\mathcal{P}}^2$ under the assumption that $\rho_{m,N} < \rho_0$, we can write

$$\begin{aligned} & \left| (P - \hat{P}_{M,N})(s_k(u^m)) \right|^2 \\ &\leq C \left(\|P\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \rho_{m,N}^\gamma \|u^m\|_{\mathcal{U}^2}^2 \right) \|s_k(u^m - \hat{u}_{M,N}^m)\|_2^2. \end{aligned}$$

It remains for us to control the error $\|s_k(u^m - \hat{u}_{M,N}^m)\|_2^2$. Thanks to Assumption 2.11(ii) and Sobolev embedding theorem, we have for any γ' satisfying $d/2 + \text{order}(\mathcal{P}) < \gamma' \leq \gamma$

that $\|s_k(v)\|_2 \leq C\|v\|_{H^{\gamma'}}$. This, together with statement (a) gives the bound

$$\begin{aligned} \left| (P - \hat{P}_{M,N})(s_k(u^m)) \right|^2 &\leq C \left(\|P\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \rho_{m,N}^{\gamma} \|u^m\|_{\mathcal{U}^2}^2 \right) \\ &\quad \times \rho_{M,N}^{2(\gamma-\gamma')} \left(\|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}}^2 \right) \\ &\leq C \rho_{M,N}^{2(\gamma-\gamma')} \left(\|P\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^2, \end{aligned} \quad (2.25)$$

where we assumed $\rho_{m,N} < 1$ for the second display.

At this point, we have shown that P and $\hat{P}_{M,N}$ are close on the discrete set S . A direct application of the noisy Sobolev sampling inequality Proposition 2.9(b) under the assumption that $\varrho_{M,N}(B) \leq \varrho_0$ gives

$$\|P - \hat{P}_{M,N}\|_{L^\infty(B)} \leq C \varrho_{M,N}(B)^{\eta - \frac{Q+d}{2}} \|P - \hat{P}_{M,N}\|_{H^\eta(B)} + 2 \|(P - \hat{P}_{M,N})|_S\|_\infty. \quad (2.26)$$

Thanks to Assumption 2.13(ii) and the bound (2.24) we can further bound the Sobolev norm in the first term on the right-hand side,

$$\begin{aligned} \|P - \hat{P}_{M,N}\|_{H^\eta(B)} &\leq \|P - \hat{P}_{M,N}\|_{H^\eta(S)} \leq \|P\|_{\mathcal{P}} + \|\hat{P}_{M,N}\|_{\mathcal{P}} \\ &\leq C \left(\|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \rho_{m,N}^{\gamma} \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2}. \end{aligned}$$

Taking the supremum over the index m under the sum, substituting this bound along with (2.25) into (2.26), and using the assumption that $\rho_{m,N} \leq 1$, we obtain the desired inequality

$$\begin{aligned} \|P - \hat{P}_{M,N}\|_{L^\infty(B)} &\leq C \left[\varrho_{M,N}(B)^{\eta - \frac{Q+d}{2}} \left(\|P\|_{\mathcal{P}}^2 + \frac{\rho_{M,N}^{\gamma}}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2} \right. \\ &\quad \left. + \rho_{M,N}^{\gamma-\gamma'} \left(\|P\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2} \right] \\ &\leq C \left(\varrho_{M,N}(B)^{\eta - \frac{Q+d}{2}} + \rho_{M,N}^{\gamma-\gamma'} \right) \left(\|P\|_{\mathcal{P}}^2 + \|\bar{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2}. \end{aligned}$$

□

Remark 2.15. We emphasize that the above error bounds are quite strong, and give pointwise control over the errors of the $\hat{u}_{M,N}^m$ and $\hat{P}_{M,N}$ over subsets of their domains, provided that the training data is sufficiently space-filling. This is in line with the theory of scattered data approximation and suggests various avenues for experimental design of equation learning algorithms. Furthermore, the bound for $\hat{P}_{M,N}$ implies that the learned PDE is accurate for new input-output pairs (u, f) that are close to the training data in an appropriate sense, i.e., if $\Phi(u, \cdot) \in B$. ◇

Remark 2.16. We note that our error analysis above can be extended in various directions to incorporate other types of problems that may be encountered in practice. For example, the terms concerning $\sup_m \rho_{m,N}$ in Theorem 2.14(b) can be too pessimistic in situations where a small portion of the training data are outliers with very large norm or large fill-distances. Then it is natural for us to consider a probabilistic model, i.e., the GP regression approach with a nugget term which allows us to obtain similar error bounds in expectation or high-probability. Such error bounds already exist in the literature and we refer the interested reader to [245, 246] for further reading. We also highlight that our proof technique can be applied beyond the particular instance of equation learning to any compositional model of multiple RKHS functions such as those encountered in computational graph completion [227]. Finally, we can view our equation learning problem as a nonlinear version of the total least squares problem [247, 248] connecting our analysis to a broader set of approximation problems with noise in the input and output of a function. $\diamond$

2.5.3 Error analysis for equation-mediated operator learning

We now extend our equation learning error bound to an operator learning bound. Classical operator-learning error rates are typically stated as convergence guarantees for learned operators between function spaces, rather than bounds for a single fixed input. They often control the worst-case error over a compact input set, as in neural-operator universal approximation results [249], or the average error under an input-function distribution, as in DeepONet analyses of approximation and generalization error [250]. In our notation, this would amount to obtaining an average or supremum error bound for $\mathcal{P}^{-1}$ over a function space containing $\mathcal{P}$. In contrast, we leverage the learned $\mathcal{P}$ as a local differential operator obtained through equation learning, then numerically invert it to evaluate $\mathcal{P}^{-1}$, thereby obtaining the resulting pointwise operator-learning bound.

Remark 2.17 (Order of operations). The error analysis of this subsection assumes that an equation-learning error bound for $\widehat{\mathcal{P}}$ has already been established, for instance the bound for 1-step KEqL in Section 2.5.2. The operator-learning error analysed here measures the additional error incurred when inverting the learned $\widehat{\mathcal{P}}$ via the variational problem (2.13) of Section 2.2.3 to evaluate $\widehat{\mathcal{P}}^\dagger$ at a new input f . The pointwise bound stated below therefore decomposes into two contributions: the equation-learning error already quantified upstream, and a stability term that controls how perturbations in $\mathcal{P}$ propagate through the inversion. $\diamond$

We begin by giving an elementary lemma that allows us to control the error between the solution of a true PDE and that of an approximate equation under sufficient regularity assumptions on the forward and inverse differential operators. For this lemma, we will consider $\mathcal{P} : \mathcal{U} \rightarrow \mathcal{F}$ and its inverse $\mathcal{P}^{-1} : \mathcal{F} \rightarrow \mathcal{U}$ for generic Banach spaces $\mathcal{U}, \mathcal{F}$.

Lemma 2.18. *Consider $A \subset \mathcal{U}$ and $B \subset \mathcal{F}$ for which it holds that:*

1. $\mathcal{P}^{-1}$ is locally Lipschitz on B , i.e., for any pair $f, f' \in B \subset \mathcal{F}$, there exists a constant $L(B)$ so that

$$\|\mathcal{P}^{-1}(f) - \mathcal{P}^{-1}(f')\|_{\mathcal{U}} \leq L(B)\|f - f'\|_{\mathcal{F}}.$$

2. $\widehat{\mathcal{P}}$ approximates $\mathcal{P}$ on A in the sense that

$$\|\widehat{\mathcal{P}}(u) - \mathcal{P}(u)\|_{\mathcal{F}} \leq \epsilon(A) \|u\|_{\mathcal{U}},$$

for some constant $\epsilon(A) > 0$.

Fix an $f \in B$ for which $\mathcal{P}^{-1}(f) \in A$, and let $\widehat{u} \in A$ be any element that solves $\widehat{\mathcal{P}}(\widehat{u}) = f$. Then we have the error bound

$$\|\widehat{u} - \mathcal{P}^{-1}(f)\|_{\mathcal{U}} \leq L(B) \epsilon(A) \|\widehat{u}\|_{\mathcal{U}}.$$

Proof. Write $u = \mathcal{P}^{-1}(f)$. Since $\widehat{u}$ solves the approximate problem we can write $\widehat{\mathcal{P}}(\widehat{u}) + \mathcal{P}(\widehat{u}) - \mathcal{P}(\widehat{u}) = f = \mathcal{P}(u)$. Combining this with the Lipschitz assumption on $\mathcal{P}^{-1}$ and the assumed error bound for $\widehat{\mathcal{P}}$ gives the chain of inequalities

$$\|\widehat{u} - u\|_{\mathcal{U}} \leq L(B) \|\mathcal{P}(\widehat{u}) - \mathcal{P}(u)\|_{\mathcal{F}} = L(B) \|\widehat{\mathcal{P}}(\widehat{u}) - \mathcal{P}(\widehat{u})\|_{\mathcal{F}} \leq L(B) \epsilon(A) \|\widehat{u}\|_{\mathcal{U}}.$$

□

Let us now apply this lemma to obtain an error bound for the learned equation $\widehat{\mathcal{P}}_{M,N}(u) = (\widehat{P}_{M,N} + \overline{P}) \circ \Phi(u, \cdot) = f$ via 1-step KEqL; a similar result can be shown for the 2-step method by a straightforward modification of the proof using our error bound for 2-step KEqL in SM.

Proposition 2.19. *Suppose Theorem 2.14 is satisfied with some set B . Let $f \in C(\mathcal{Y})$ be a right-hand side function such that*

1. $\mathcal{P}^{-1}$ is locally Lipschitz in a neighborhood of f , i.e., $\|\mathcal{P}^{-1}(f) - \mathcal{P}^{-1}(f')\|_{\mathcal{U}} \leq L\|f - f'\|_{C(\mathcal{Y})}$ for f' in some neighborhood of f .
2. $\Phi(\mathcal{P}^{-1}(f), y) \in B$ for all $y \in \mathcal{Y}$.

Then, for sufficiently large M, N , and any $\widehat{u}$ that satisfies $\widehat{\mathcal{P}}_{M,N}(\widehat{u}) = f$, we have

$$\begin{aligned} & \|\widehat{u} - \mathcal{P}^{-1}(f)\|_{\mathcal{U}} \\ & \leq C \|\widehat{u}\|_{\mathcal{U}} \left(\varrho_{M,N}^{\eta - \frac{Q+d}{2}} + \rho_{M,N}^{\gamma - \gamma'} \right) \left(\|P\|_{\mathcal{P}}^2 + \|\overline{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2}. \end{aligned}$$

Proof. Let us write $u = \mathcal{P}^{-1}(f)$, and observe that thanks to hypothesis (2) of the theorem and Theorem 2.14, we have the bound

$$\begin{aligned} & \|\widehat{\mathcal{P}}_{M,N}(u) - \mathcal{P}(u)\|_{C(\mathcal{Y})} \\ & \leq C \left(\varrho_{M,N}^{\eta - \frac{Q+d}{2}} + \rho_{M,N}^{\gamma - \gamma'} \right) \left(\|P\|_{\mathcal{P}}^2 + \|\overline{P}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}^2}^2 \right)^{1/2}. \end{aligned}$$

The result follows by a straightforward application of Lemma 2.18.

□

Remark 2.20. The above proposition tells us that the operator learning problem associated to "solving" $\widehat{\mathcal{P}}(u) = f$ essentially inherits the same rate of convergence as the equation learning problem so long as the new right-hand side function f is not too different from those encountered in the training data. This is inline also with the local nature of the type of error bounds we have derived in this section for $\widehat{P}$ from the theory of scattered data approximation. We also note that our result can naturally accommodate SINDy type methods or other abstract optimal recovery approaches to equation learning once an error bound for $\widehat{P}$ is available. $\diamond$

Remark 2.21. In general, our methodology, convergence results, and experiments hold for non-pointwise settings. However, the error estimates both for equation learning and operator learning are pointwise. Although we did not prove non-pointwise error estimates, techniques for proving such estimates can be found in Theorem 4.11. and Chapter 14 of [166]. $\diamond$

2.5.4 Connection to existing results

The theoretical analysis of equation learning has been largely focused on the case of sparse recovery methods [190] while more recent articles have also considered fundamental limitations of equation learning such as identifiability [192] and PDE learning from a single trajectory [193]. More precise results have also been obtained for particular types of PDEs such as elliptic equations [251, 252] as well as for ODEs in [253]. Our analysis is different from these works in a number of directions: (1) to our knowledge previous works do not cover the case of the 1-step methods or the RKHS formulation; (2) quantitative worst-case error bounds such as ours are first of their kind; (3) some of the stronger guarantees in the literature are tailored to specific PDEs while our results apply to generic nonlinear PDEs under sufficient smoothness assumptions. In connection to operator learning we note that most of the works in this direction are focused on the function space regression framework [254] and, to our knowledge, our results above are a first of their kind in this direction.

2.6 Discussion

We presented the 1- and 2-step formulations of KEqL as an algorithm for learning nonlinear PDEs as well as their solution operators and state reconstruction of observed states in scarce data regimes. While our algorithms, experiments, and theoretical analyses were focused on an RKHS formulation our abstract methodology can be extended to other types of models including neural nets or sparse regression methods such as SINDy. Additionally, our exposition unified the problems of equation learning, operator learning, and solving PDEs under the umbrella of optimal recovery.

Our theoretical results provided quantitative error bounds for our algorithms while our numerical experiments demonstrated significant gains in accuracy and robustness in comparison to existing methods in the literature. Most notably, our methods appear to be a lot less sensitive to choice of hyperparameters and demonstrate more consistent performance compared to equivalent neural net methods.

Various avenues of future research and extensions of the KEqL framework can be identified: (a) we outlined KEqL for a single PDE but its extension to systems of PDEs is an

obvious next step since many physical processes of interest are governed by systems of equations; (b) our approach to operator learning, after the deployment of KEqL, relies on a PDE solver for each evaluation of $\widehat{\mathcal{P}}^\dagger$ which can be expensive. Therefore it may be interesting to investigate the emulation of this process to obtain a cheap solver that can be deployed for real time predictions; (c) In many of our experimental results we observed very competitive accuracy for 1-step KEqL, however, we still observed a relative error barrier (around 10^{-2} in Figure 2.4) which is not in line with our theoretical guarantees, we suspect these issues may arise due to ill-conditioning of the problems and various approximations made in the algorithms in order to scale the computations; (d) While our theoretical analysis addresses asymptotic convergence rates it does not apply to scarce data regimes where 1-step KEqL is most useful and so a non-asymptotic analysis in this regime would be of great interest.

Acknowledgments

JFOR, AH and BH were supported by the National Science Foundations grants DMS-2208535 “Machine Learning for Inverse Problems” and DMS-2337678 “Gaussian Processes for Scientific Machine Learning: Theoretical Analysis and Computational Algorithms”. YJ and HO acknowledge support from the Air Force Office of Scientific Research under MURI awards number FA9550-20-1-0358 (Machine Learning and Physics-Based Modeling and Simulation), FOA-AFRL-AFOSR-2023-0004 (Mathematics of Digital Twins), 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), the National Science Foundations under award number 2425909 (Discovering the Law of Stress Transfer and Earthquake Dynamics in a Fault Network using a Computational Graph Discovery Approach), and the Jet Propulsion Laboratory PDRDF 24AW0133 (Jupiter’s Radiation Environment: Assimilating Data with ML-driven Approaches). Additionally, HO acknowledges support from the DoD Vannevar Bush Faculty Fellowship Program.

This document contains additional material that supplement our main manuscript. Section 2.7 gives additional theoretical results including an error bound for 2-step KEqL and a simple proof of convergence for the 1-step method. Section 2.8 contains further implementation details for our algorithms while Section 2.9 contains additional results from our numerical experiments in the main body.

2.7 Details of theoretical analysis

In this section we outline additional theoretical details such as proofs of convergence of the 2-step and 1-step KEqL methods in Section 2.7.2 and Section 2.7.3 respectively, as well as their representer theorems in Section 2.7.4.

2.7.1 Preliminaries

Here we briefly recall the Sobolev embedding theorem that is central to our error analysis. We will not present a review of Sobolev spaces and instead refer the reader to the classic

texts [255, 256].

For an open set $\Omega \subseteq \mathbb{R}^q$ and $\gamma \in \mathbb{N}$ we write $H^\gamma(\Omega)$ to denote the $L^2(\Omega)$ based Sobolev space of index γ equipped with the norm $\|u\|_{H^\gamma(\Omega)}^2 = \sum_{|\mathbf{a}| \leq \gamma} \|D^{\mathbf{a}}u\|_{L^2(\Omega)}^2$, where $\mathbf{a}$ is a multi-index. We also employ the convention $H^0(\Omega) = L^2(\Omega)$. We then have the following Sobolev embedding theorem:

Proposition 2.22 (Sobolev embedding theorem [255, Thm. 4.12]). *Suppose $\Omega \subset \mathbb{R}^q$ is a bounded set with Lipschitz boundary and that for $p \in \mathbb{N}$ it holds that $\gamma > q/2 + p$. Then $H^\gamma(\Omega)$ is continuously embedded in $C^p(\Omega)$ and it holds that $\|u\|_{C^p(\Omega)} \leq C_\Omega \|u\|_{H^\gamma(\Omega)}$ for an embedding constant $C_\Omega \geq 0$ that depends only on Ω .*

2.7.2 Error analysis for 2-step KEqL

We present an error estimate for the 2-step KEqL. We follow the same notation as the main theorem, and we follow similar steps.

Theorem 2.23. *Consider the problem in (2.6) and (2.8) with $M, N \in \mathbb{N}$. Suppose Assumptions 2.10, 2.11, and 2.13 hold, $P, \bar{P} \in \mathcal{P}$, and $u^m \in \mathcal{U}$. Consider pairs $\{u^m, f^m\}_{m=1}^M$ satisfying (2.3), and a bounded set $B \subset \mathcal{S}$ with Lipschitz boundary. Then there exist constants $\rho_0, \varrho_0(B) \in (0, 1)$ so that whenever $\rho_{m,N} < \rho_0$ and $\varrho_{M,N}(B) < \varrho_0(B)$ it holds that:*

$$\|\hat{u}_N^m - u^m\|_{H^{\gamma'}(\mathcal{Y})} \leq C \rho_{m,N}^{(\gamma-\gamma')} \|u^m\|_{\mathcal{U}},$$

for $0 \leq \gamma' < \gamma$ and a constant $C > 0$ that depends only on $\mathcal{Y}$. Furthermore,

$$\begin{aligned} \|P - \hat{P}_{M,N}\|_{L^\infty(B)} &\leq C \max\{1, 2^{\eta - \frac{Q+d}{2}}\} \times \\ &\left[\varrho_{M,N}^{\eta - \frac{Q+d}{2}} + \max \left\{ \left(\rho_{M,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}} \right)^{\eta - \frac{Q+d}{2}}, \rho_{M,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}} \right\} (\|P\|_{\mathcal{P}} + \|\bar{P}\|_{\mathcal{P}}) \right], \end{aligned}$$

for $d/2 + \text{order}(\mathcal{P}) < \gamma' < \gamma$ and a constant $C > 0$ that depends on $\mathcal{Y}$ and B .

Proof. (a) The first statement is a direct application of the Sobolev sampling inequality Proposition 2.9(a) since the $\hat{u}_N^m$ are simply the kernel interpolants of the u^m .

(b) Consider the definition of P and observe that the interpolation constraint for $\hat{P}_{M,N}$ can be written as

$$\begin{aligned} (\hat{P}_{M,N} + \bar{P})(s_k(\hat{u}_N^m)) &= (P + \bar{P})(s_k(u_N^m)) \\ &= (P + \bar{P})(s_k(\hat{u}_N^m)) + (P + \bar{P})(s_k(u^m)) - (P + \bar{P})(s_k(\hat{u}_N^m)) \\ &= (P + \bar{P})(s_k(\hat{u}_N^m)) + (P + \bar{P})(s_k(u^m - \hat{u}_N^m)), \end{aligned}$$

where once again we used the fact that the $s_k(v)$ is linear in v . Assuming $s_k(u^m)$ and $s_k(u_N^m)$ belong to the set B we can use the local Lipschitz property of P and $\bar{P}$ to infer that

$$|\hat{P}_{M,N}(s_k(\hat{u}_N^m)) - P(s_k(\hat{u}_N^m))| \leq C(B) (\|P\|_{\mathcal{P}} + \|\bar{P}\|_{\mathcal{P}}) \|s_k(u^m - \hat{u}_N^m)\|_2$$

By the Sobolev sampling inequality Proposition 2.9(a) and Assumption 2.11(ii) we can further bound

$$\|s_k(u^m - \widehat{u}_N^m)\|_2 \leq C \rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}}, \quad (2.27)$$

where $\gamma > \gamma' > d/2 + \text{order}(\mathcal{P})$. This further leads to an error bound on the difference between $\widehat{P}_{M,N}$ and P on the set of points $s_k(\widehat{u}_N^m)$:

$$|\widehat{P}_{M,N}(s_k(\widehat{u}_N^m)) - P(s_k(\widehat{u}_N^m))| \leq C \rho_{m,N}^{\gamma-\gamma'} (\|P\|_{\mathcal{P}} + \|\overline{P}\|_{\mathcal{P}}) \|u^m\|_{\mathcal{U}}$$

We can now apply the noisy Sobolev sampling inequality Proposition 2.9(b), viewing $\widehat{P}_{M,N}$ as the noisy interpolant on P on the set of points $s_k(\widehat{u}_N^m)$, to get the error bound

$$\|P - \widehat{P}_{M,N}\|_{L^\infty(B)} \leq C \left[\widehat{\varrho}_{M,N}^{\eta-\frac{Q+d}{2}} \|P\|_{\mathcal{P}} + \rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}} (\|P\|_{\mathcal{P}} + \|\overline{P}\|_{\mathcal{P}}) \right],$$

where we introduced the notation $\widehat{\varrho}_{M,N} := \sup_{s' \in B} \inf_{k,m} \|s' - s_k(\widehat{u}_N^m)\|$. By (2.27) we have that $\|s' - s_k(\widehat{u}_N^m)\| \leq \|s' - s_k(u^m)\| + C \sup_m (\rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}})$ which in turn implies $\widehat{\varrho}_{M,N} \leq \varrho_{M,N} + C \sup_m (\rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}})$. Substituting back into the bound above yields

$$\begin{aligned} \|P - \widehat{P}_{M,N}\|_{L^\infty(B)} &\leq C \left[\left(\varrho_{M,N} + \sup_m (\rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}}) \right)^{\eta-\frac{Q+d}{2}} \|P\|_{\mathcal{P}} \right. \\ &\quad \left. + \rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}} (\|P\|_{\mathcal{P}} + \|\overline{P}\|_{\mathcal{P}}) \right]. \end{aligned}$$

Using the inequality $(a+b)^p \leq \max\{1, 2^{p-1}\}(a^p + b^p)$ for $p \in (0, +\infty]$ we can simplify this bound to take the desired form

$$\begin{aligned} \|P - \widehat{P}_{M,N}\|_{L^\infty(B)} &\leq C \max\{1, 2^{\eta-\frac{Q+d}{2}}\} \times \\ &\quad \left[\varrho_{M,N}^{\eta-\frac{Q+d}{2}} + \max \left\{ \left(\rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}} \right)^{\eta-\frac{Q+d}{2}}, \rho_{m,N}^{\gamma-\gamma'} \|u^m\|_{\mathcal{U}} \right\} (\|P\|_{\mathcal{P}} + \|\overline{P}\|_{\mathcal{P}}) \right]. \end{aligned}$$

□

2.7.3 Convergence guarantees for 1-step KEqL

In this section we present the proof for the convergence of the 1-step KEqL which requires less stringent conditions such as Lipschitzness of elements in the RKHS $\mathcal{P}$. Additionally, this proof is enlightening since it has potential to be extended to non-RKHS optimal recovery models.

Theorem 2.24. *Consider the problem (2.11). Suppose Assumptions 2.10 and 2.11 and Assumption 2.13(i-ii) hold. In addition, suppose that the continuous embeddings stated in Assumption 2.11(ii) and Assumption 2.13(ii) are compact. Finally, let $P \in \mathcal{P}$, and assume that $\overline{P}$ is continuous. Consider pairs $\{u^m, f^m\}_{m=1}^\infty$ satisfying (2.3) and a bounded set $B \subset \mathcal{S}$ with Lipschitz boundary. Finally, suppose $\rho_{m,N} \rightarrow 0$ as $N \rightarrow \infty$ for all m and $\varrho_{M,N}(B) \rightarrow 0$ as $M, N \rightarrow \infty$. Then the following holds:*

- (a) Fix M . If $u^m \in \mathcal{U}$ then $\lim_{N \rightarrow \infty} \hat{u}_{M,N}^m = u^m$ pointwise in $\mathcal{Y}$ and in $H^{\gamma'}(\mathcal{Y})$ for all $\gamma' < \gamma$ and $m \in \{1, \dots, M\}$.
- (b) If $u^m \in \mathcal{U}^2$ then $\lim_{M \rightarrow \infty} \lim_{N \rightarrow \infty} \hat{P}_{M,N} = P$ pointwise in B and in $H^{\eta'}(B)$ for all $\eta' < \eta$.

Proof. Let us verify that our assumptions are sufficient for the problem to be well defined. Since $\mathcal{U}$ is an RKHS then the problem for $\hat{u}_{M,N}^m$ is readily well defined. Moreover, since U is continuous the space $\mathcal{U}^2$ is well defined as in Section 2.5.1. Moreover, Assumption 2.10 along with Assumption 2.11(ii) and Proposition 2.22 ensure that Φ is continuous and bounded, in fact, $\Phi(\cdot, y)$ is a bounded and linear operator on $\mathcal{U}$; this is simply the statement that $\mathcal{P}(u)$ is defined pointwise. Finally, since $\mathcal{P}$ is an RKHS and $\bar{P}$ is continuous we ensure that the problem for $\hat{P}_{M,N}$ is also well defined.

Proof of (a): Observe that the pair $(\mathbf{u}, P)$ (notation: $\mathbf{u} = (u^1, u^2, \dots, u^M)$) are feasible for (2.11) for all values of M, N . Then the optimality of $(\hat{\mathbf{u}}_{M,N}, \hat{P}_{M,N})$ implies that

$$\|\hat{P}_{M,N}\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|\hat{u}_{M,N}^m\|_{\mathcal{U}}^2 \leq \|P\|_{\mathcal{P}}^2 + \frac{1}{M} \sum_{m=1}^M \|u^m\|_{\mathcal{U}}^2. \quad (2.28)$$

Thus, for fixed M we have that $\|\hat{u}_{M,N}^m\|_{\mathcal{U}} \leq C(M)$ for all m , i.e., $\{\hat{u}_{M,N}^m\}_{N=1}^{\infty}$ is bounded in $\mathcal{U}$. Then Assumption 2.11(ii) implies that $\hat{u}_{M,N}^m$ has a convergent subsequence in $H^{\gamma}(\mathcal{Y})$. On the other hand, since $\rho_{m,N} \rightarrow 0$ and thanks to the assumption that u^m are continuous (since they belong to $\mathcal{U}$) we infer that all accumulation points of $\hat{u}_{M,N}^m$ coincide with u^m . This implies that for fixed M and for any $m \in \{1, \dots, M\}$ we have $\lim_{N \rightarrow \infty} \hat{u}_{M,N}^m = u^m$ pointwise and in $H^{\gamma'}(\mathcal{Y})$ for $\gamma' < \gamma$.

Proof of (b): Fix M and for each u^m define the optimal interpolant

$$\bar{u}_N^m := \arg \min_{v \in \mathcal{U}} \|v\|_{\mathcal{U}} \quad \text{s.t.} \quad v(Y_N^m) = u^m(Y_N^m). \quad (2.29)$$

By optimality, we have the bound

$$\|\hat{u}_{M,N}^m\|_{\mathcal{U}} \geq \|\bar{u}_N^m\|_{\mathcal{U}}. \quad (2.30)$$

At the same time, the representer theorem gives

$$\bar{u}_N^m = U(Y_N^m, \cdot)^T U(Y_N^m, Y_N^m)^{-1} u^m(Y_N^m)$$

and a direct calculation using the reproducing property implies that $\langle u^m - \bar{u}_N^m, \bar{u}_N^m \rangle_{\mathcal{U}} = 0$ which in turn gives the identity

$$\|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2 = \|u^m\|_{\mathcal{U}}^2 - \|\bar{u}_N^m\|_{\mathcal{U}}^2. \quad (2.31)$$

At the same time, since we assumed $u^m \in \mathcal{U}^2$ we also have, using Lemma 2.8,

$$\|u^m - \bar{u}_N^m\|_{\mathcal{U}}^2 = \langle u^m - \bar{u}_N^m, u^m \rangle_{\mathcal{U}} \leq \|u^m - \bar{u}_N^m\|_{L^2(\mathcal{Y})} \|u^m\|_{\mathcal{U}^2}.$$

But, an identical argument to part (a) shows that $\|u^m - \bar{u}_N^m\|_{L^2(\mathcal{Y})} \rightarrow 0$ as $N \rightarrow \infty$ and so we infer that $\lim_{N \rightarrow \infty} \|u^m - \bar{u}_N^m\|_{\mathcal{U}} = 0$. This fact, together with (2.30) and (2.31) implies that

$$\limsup_{N \rightarrow \infty} \|\hat{u}_N^m\|_{\mathcal{U}} \geq \|u^m\|_{\mathcal{U}}.$$

Then it follows from (2.28) that

$$\limsup_{N \rightarrow \infty} \|\hat{P}_{M,N}\|_{\mathcal{P}} \leq \|P\|_{\mathcal{P}}.$$

This implies that the sequence $\{\hat{P}_{M,N}\}_{N=1}^{\infty}$ is bounded and so has a convergent subsequence due to Assumption 2.13(ii). From part (a) we also have that $\lim_{N \rightarrow \infty} \Phi(\hat{u}_{M,N}^m, y) = \Phi(u^m, y)$ for all $y \in \mathcal{Y}$ and $m \in \{1, \dots, M\}$. $\square$

2.7.4 Representer theorems for KEqL

Below we give the proofs of representer theorems for both versions of the KEqL algorithm as presented in the main body.

Representer theorems for 2-step KEqL

We will give a brief justification for the expressions (2.14) and (2.15) as direct consequences of Proposition 2.4: Consider (2.6) and apply Proposition 2.4 with $\mathcal{H} = \mathcal{U}$, $\phi_i = \delta_{y_i^m}$, and $z_i = u^m(y_i^m)$ for all $i = 1, \dots, N$. This gives (2.14). Next consider (2.8) and apply Proposition 2.4 with $\mathcal{H} = \mathcal{P}$, $\phi_i = \delta_{s_i}$ for $s_i \in S$, and $z_i = f(y_i)$ for $y_i \in Y$ to obtain (2.15).

Representer theorem for 1-step KEqL

Here, we prove Theorem 2.5). To simplify notation, we first prove the theorem with $M = 1$, i.e., with training data $(u(Y^1), f)$, and $\bar{P} = 0$. At the end of the proof we comment on how the argument can be extended to the general case. Let us recall the corresponding optimal recovery problem for convenience:

$$\begin{aligned} (\hat{u}, \hat{P}) &= \arg \min_{v \in \mathcal{U}, G \in \mathcal{P}} \|G\|_{\mathcal{P}}^2 + \lambda_1 \|v\|_{\mathcal{U}}^2 \\ \text{s.t.} \quad & v(Y^1) = u(Y^1), \quad \text{and} \quad G(S) = f(Y), \\ \text{where} \quad & S = \Phi(v, Y). \end{aligned} \tag{2.32}$$

Recall that Y is the dense set of collocation points that we use to impose the PDE constraint.

Then our goal is to show that under the assumptions of Theorem 2.5, every minimizing tuple $(\hat{u}, \hat{P})$ of (2.32) can be written in the form

$$\hat{u}(y) = \mathbf{U}(\phi, y)^T \hat{\alpha}, \quad \hat{P}(s) = \mathbf{P}(S(\hat{\alpha}), s)^T \hat{\beta},$$

for a tuple $(\hat{\alpha}, \hat{\beta})$ that solves the equivalent optimization problem

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}) &= \arg \min_{\alpha \in \mathbb{R}^{QK}, \beta \in \mathbb{R}^K} \beta^T \mathbf{P}(S(\alpha), S(\alpha)) \beta + \lambda_1 \alpha^T \mathbf{U}(\phi, \phi) \alpha, \\ \text{s.t.} \quad & \mathbf{U}(\phi, Y^1)^T \alpha = u(Y^1), \quad \text{and} \quad \mathbf{P}(S(\alpha), S(\alpha))^T \beta = f(Y), \\ \text{where} \quad & S(\alpha) = \Phi(\mathbf{U}(\phi, \cdot)^T \alpha, Y). \end{aligned} \tag{2.33}$$

Recall our notation $\mathbf{U}(\phi, Y^m) \in \mathbb{R}^{QK \times N}$ for the matrix with columns $\mathbf{U}(\phi, y_n)$, and $\mathbf{P}(S(\alpha), S(\alpha)) \in \mathbb{R}^{K \times K}$ for the matrix with columns $\mathbf{P}(S(\alpha), s_k)$ with $s_k = \Phi(\mathbf{U}(\phi, \cdot)^T \alpha, y_k)$.

Proof. Let us write (2.32) in the equivalent form

$$\begin{cases} \arg \min_{Z \in \mathbb{R}^{K \times Q}} & \begin{cases} \arg \min_{v \in \mathcal{U}, G \in \mathcal{P}} \|G\|_{\mathcal{P}}^2 + \lambda_1 \|v\|_{\mathcal{U}}^2 \\ \text{s.t. } \phi_k^q(v) = Z_{k,q} \text{ and } G(s_k) = f(y_k), \\ \text{for all } k = 1, \dots, K, \text{ and } q = 1, \dots, Q \\ \text{where } s_k := (y_k, Z_{k,1}, \dots, Z_{k,Q}) \end{cases} \\ \text{s.t. } & Z_{j,1} = u(y_j), \quad \text{for } j \in \{k \in \{1, \dots, K\} : y_k \in Y^1\}. \end{cases} \quad (2.34)$$

Observe that the Z variable matrix that we introduce acts as a slack variable for the inner problem and the outer constraint is enforcing the observation locations of v at $Y^1 \subset Y$.

For a fixed $Z \in \mathbb{R}^{K \times Q}$, we can solve the inner optimization problem for v and G explicitly using Proposition 2.4, which leads to

$$v(y) = \mathbf{U}(\phi, y)^T \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z), \quad \text{and} \quad G(s) = \mathbf{P}(S, s)^T \mathbf{P}(S, S)^{-1} f(Y).$$

The RKHS norms of these solutions can be computed explicitly as

$$\|v\|_{\mathcal{U}}^2 = \text{vec}(Z)^T \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z), \quad \text{and} \quad \|G\|_{\mathcal{P}}^2 = f(Y)^T \mathbf{P}(S, S)^{-1} f(Y).$$

Thus, we equivalently write (2.34) as

$$\begin{aligned} & \arg \min_{Z \in \mathbb{R}^{K \times Q}} f(Y)^T \mathbf{P}(S, S)^{-1} f(Y) + \lambda_1 \text{vec}(Z)^T \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z) \\ & \text{s.t. } \mathbf{U}(\phi, Y^1)^T \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z) = u(Y^1), \end{aligned} \quad (2.35)$$

where we now modified the s_k to be of the form

$$s_k = (y_k, \mathbf{U}(\phi, \phi_k^1)^T \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z), \dots, \mathbf{U}(\phi, \phi_k^Q)^T \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z)).$$

Similarly, the constraint in (2.35) is valid since $\phi_k^q(v) = Z_{k,q}$ which particularly includes $\phi_j^1(v) = Z_{j,1}$ for $j \in \{\ell \in \{1, \dots, K\} : y_\ell \in Y^1\}$.

Finally, if we define $\alpha = \mathbf{U}(\phi, \phi)^{-1} \text{vec}(Z)$ and $\beta = \mathbf{P}(S, S)^{-1} f(Y)$ (where in case $\mathbf{U}(\phi, \phi)$ and $\mathbf{P}(S, S)$ are not invertible, understand α and β in the least squares sense) then the problem (2.35) is equivalent to the desired system

$$\begin{aligned} & \arg \min_{\alpha \in \mathbb{R}^{QK}, \beta \in \mathbb{R}^K} \beta^T \mathbf{P}(S, S) \beta + \lambda_1 \alpha^T \mathbf{U}(\phi, \phi) \alpha \\ & \text{s.t. } \mathbf{U}(\phi, Y^1)^T \alpha = u(Y^1), \quad \text{and} \quad \mathbf{P}(S, S)^T \beta = f(Y) \end{aligned}$$

where $S = \{s_1, \dots, s_K\}$ with each $s_k = (y_k, \mathbf{U}(\phi, \phi_k^1)^T \alpha, \dots, \mathbf{U}(\phi, \phi_k^Q)^T \alpha)$. Notice that the constraint on β is trivially satisfied since $G(s_k) = f(y_k)$ for all $k \in \{1, \dots, K\}$, however, it becomes explicit due to the change of variables in terms of β in the formulation of G .

For the general proof for an arbitrary number of functions M the same idea follows since the equivalent form (2.34) now contains the slack tensor $Z \in \mathbb{R}^{K \times Q \times M}$ instead of the matrix Z above since each v^m will be represented by a matrix Z^m in our proof. Additionally S will then contain not only K points but MK points corresponding to each of the v^m and their requisite partial derivatives evaluated at the collocation points. To obtain the result when $\bar{P} \neq 0$ the proof simply follows by replacing the data $f(Y)$ in (2.32) with $\bar{P}(S)$ and similarly in subsequent arguments in the proof. $\square$

2.8 Details of algorithms

In this section, we describe additional algorithmic details relating to the 1-step KEqL method, i.e., (2.17). As it was mentioned in the main body of the chapter, implementation of our LM formulation already leads to a convergent algorithm with good empirical performance. However, dealing with large kernel matrices often limits the scalability of that method. To address these issues we outline various strategies in this section that are implemented in our numerical experiments. These optional strategies include: a Nyström approximation to our model for P using a reduced basis; applying a change of variables in the optimization algorithms using Cholesky factors of kernel matrices to improve conditioning; and fast matrix calculations by leveraging block arrow structures.

2.8.1 Implementation of 1-step KEqL: the reduced version

An LM algorithm for 1-step KEqL

Solving (2.17) is difficult in practice as a small perturbation in the α^m can lead to a large perturbation in the derivatives of the corresponding functions $v^m = \mathbf{U}(\phi, \cdot)^T \alpha^m$. This large variation further translates into a sharp change in the $S^m(\alpha^m)$ points leading to instabilities. To mitigate this issue, we propose a Levenberg-Marquardt algorithm that will approximate the objective function of (2.17) with a sequence of quadratic minimization problems.

Define $\mathcal{J}^m(\alpha, \beta) := \mathbf{P}(S(\alpha), S^m(\alpha^m))^T \beta + \bar{P}(S^m(\alpha^m)) - f^m(Y)$ so that the last term in (2.17) is simply $\frac{1}{2\sigma_P^2} \|\mathcal{J}^m(\alpha, \beta)\|_2^2$. Then we compute a minimizing sequence $(\alpha_{(j)}, \beta_{(j)})_{j=1}^\infty$ given by the scheme

$$\begin{aligned}
 (\alpha_{(j+1)}, \beta_{(j+1)}) = & \arg \min_{\alpha \in (\mathbb{R}^{QK})^M, \beta \in \mathbb{R}^{MK}} \beta^T \mathbf{P}(S(\alpha_{(j)}), S(\alpha_{(j)})) \beta \\
 & + \sum_{m=1}^M \left[\frac{\lambda}{M} (\alpha^m)^T \mathbf{U}(\phi, \phi) \alpha^m + \frac{1}{2\sigma_u^2} \|\mathbf{U}(\phi, Y^m)^T \alpha^m - u^m(Y^m)\|_2^2 \right. \\
 & \quad \left. + \frac{1}{2\sigma_P^2} \left\| \mathcal{J}^m(\alpha_{(j)}, \beta_{(j)}) + \nabla \mathcal{J}^m(\alpha_{(j)}, \beta_{(j)}) \begin{pmatrix} \alpha - \alpha_{(j)} \\ \beta - \beta_{(j)} \end{pmatrix} \right\|_2^2 \right] \\
 & + \lambda_{(j)} \left[(\beta - \beta_{(j)})^T \mathbf{P}(S(\alpha_{(j)}), S(\alpha_{(j)})) (\beta - \beta_{(j)}) \right. \\
 & \quad \left. + \sum_{m=1}^M (\alpha^m - \alpha_{(j)}^m)^T \mathbf{U}(\phi, \phi) (\alpha^m - \alpha_{(j)}^m) \right].
 \end{aligned} \tag{2.36}$$

The first two terms are identical to (2.17) with α fixed at the previous value $\alpha_{(j)}$, the third term is unchanged, and the forth term is a local quadratic approximation obtained by linearizing $\mathcal{J}^m$ with $\nabla \mathcal{J}^m$ denoting the Jacobian. The last term acts as a damping term that ensures that our next estimate $(\alpha_{(j+1)}, \beta_{(j+1)})$ does not deviate too far from the current values. The $\lambda_{(j)} > 0$ is inversely proportional to a step-size parameter which is updated using a standard gain ratio heuristic that compares the decrease produced in the true objective to the decrease observed in the quadratic approximation; see SM for more details.

We highlight that naive implementations of (2.36) lead to effective algorithms, but the computational costs due to matrices $P(S, S)$ and $U(\phi, \phi)$ become prohibitive when N, M are large. To address these bottlenecks we implement different computation techniques such as inducing point approximations and block matrix inversion that can be found in SM as well as our Github repository.

Reduced 1-step KEqL: an efficient approximation

We briefly introduce an approximate 1-step KEqL that, at a small cost to accuracy, leads to a better conditioned and more efficient formulation. Let us approximate (2.17) by restricting the v^m 's to the subspace $\text{span}\{U(y_1, \cdot), \dots, U(y_K, \cdot)\} \subset \mathcal{U}$, i.e., we write $v^m = U(Y, \cdot)^T \alpha^m$ for $\alpha^m \in \mathbb{R}^K$ as opposed to $\mathbb{R}^{QK}$ for 1-step KEqL. Thus this reformulation effectively constitutes a reduced basis/feature map formulation leading to the following analogous problem:

$$\begin{aligned}
 (\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha \in (\mathbb{R}^K)^M, \beta \in \mathbb{R}^{MK}} & \beta^T P(S(\alpha), S(\alpha)) \beta + \sum_{m=1}^M \frac{\lambda}{M} (\alpha^m)^T U(Y, Y) \alpha^m \\
 & + \frac{1}{2\sigma_u^2} \|U(Y, Y^m)^T \alpha^m - u^m(Y^m)\|_2^2 \\
 & + \frac{1}{2\sigma_P^2} \|P(S(\alpha), S^m(\alpha^m))^T \beta + \bar{P}(S^m(\alpha^m) - f^m(Y))\|_2^2,
 \end{aligned} \tag{2.37}$$

with $S(\alpha) = \cup_{m=1}^M S^m(\alpha^m)$ as before but with $S^m(\alpha^m) = \Phi(U(Y, \cdot)^T \alpha^m, Y)$.

Observe that if we take $Y^m = Y$ for all m then the first constraint above completely identifies the α^m and the reduced 1-step KEqL coincides with 2-step KEqL. Therefore this method interpolates between the 1-step and 2-step methods. Furthermore, we can apply the same idea to the $P(S, s)$ feature maps and choose a reduced basis for representing the learned equation $\hat{P}$ by, for example, subsampling the points in S in (2.37). Regardless, the LM algorithm of Section 2.8.1 remains applicable.

2.8.2 Nyström approximations and reduced bases for P

It is often possible to replace the canonical representer basis $P(S(\alpha), s)$ for P with a fixed set of $I \ll MK$ inducing points S_I [257]. These inducing points can be chosen by randomly sub-sampling the elements of $S(\alpha)$ at every iteration of the LM algorithm, but this can lead to technical difficulties since the point cloud $S(\alpha)$ changes from one iteration to the next, meaning that our reduced basis for P also needs to change. Instead, we simply choose S_I by

subsampling the point clouds $\Phi(\hat{u}^m, Y)$ where the $\hat{u}^m$ denote our estimates of the u^m from 2-step KEqL. Hence, our reduced basis for P is pre-computed before the LM algorithm is implemented. Note, that this approximation becomes exact whenever the number of inducing points $I \geq \dim(\mathcal{P})$, for example, in the case of polynomial kernels which have finite rank, choosing a sufficiently large number of inducing points ensures exact solution of the problem.

By fixing S_I , our finite dimensional optimization problem becomes

$$\begin{aligned}
 (\hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha \in (\mathbb{R}^{QK})^M, \beta \in \mathbb{R}^I} & \beta^T P(S_I, S_I) \beta + \frac{\lambda_1}{M} \sum_{m=1}^M (\alpha^m)^T U(\phi, \phi) \alpha^m \\
 & + \sum_{m=1}^M \frac{1}{2\sigma_u^2} \|U(\phi, Y^m)^T \alpha^m - u^m(Y^m)\|_2^2 \\
 & + \frac{1}{2\sigma_P^2} \|P(S_I, S^m(\alpha^m))^T \beta - f^m(Y)\|_2^2,
 \end{aligned} \tag{2.38}$$

Notice that the matrix $P(S_I, S_I)$ is now fixed and of much smaller size than the full matrix $P(S(\alpha), S(\alpha))$.

2.8.3 Cholesky change of variables

Taking the Cholesky factorizations $P(S_I, S_I) = C_P C_P^T$ and $U(\phi, \phi) = C_U C_U^T$, we can define the transformed variables

$$w^m = C_U^T \alpha^m, \quad z = C_P^T \beta.$$

Under this transformation, we can re-write (2.38) as,

$$\begin{aligned}
 (\hat{w}, \hat{z}) = \arg \min_{w \in (\mathbb{R}^{QK})^M, z \in \mathbb{R}^I} & \|z\|^2 + \frac{\lambda_1}{M} \sum_{m=1}^M \|w^m\|^2 \\
 & + \sum_{m=1}^M \frac{1}{2\sigma_u^2} \|U(\phi, Y^m)^T C_U^{-T} w^m - u^m(Y^m)\|_2^2 \\
 & + \frac{1}{2\sigma_P^2} \|P(S_I, S^m(w^m))^T C_P^{-T} z - f^m(Y)\|_2^2.
 \end{aligned} \tag{2.39}$$

While the LM algorithm is generally invariant under such changes of variables this greatly improves the conditioning of the optimization problems in practice. Indeed, we found that this change of variables enables first order methods, such as gradient descent, to also produce reasonable solutions when otherwise they would have failed.

To further clarify our implementation of the LM algorithms we abstract the objective

into a general nonlinear least squares problem by rewriting (2.8.3) as

$$\begin{aligned}
(\hat{\mathbf{w}}, \hat{z}) &= \arg \min_{\mathbf{w} \in (\mathbb{R}^{QK})^M, z \in \mathbb{R}^I} \|z\|^2 + \frac{1}{M} \sum_{m=1}^M \|w^m\|^2 + \|F(z, \mathbf{w})\|^2 \\
\text{where } F(z, \mathbf{w}) &= \begin{bmatrix} F_1(z, w^1) \\ F_2(z, w^2) \\ \vdots \\ F_M(z, w^M) \end{bmatrix} \in \mathbb{R}^{M(N+K)}, \text{ and} \\
F_m(z, w_m) &= \begin{bmatrix} \sqrt{\frac{1}{2\sigma_u^2}} \left(\mathbf{U}(\phi, Y^m)^T C_U^{-T} w^m - u^m(Y^m) \right) \\ \sqrt{\frac{1}{2\sigma_P^2}} \left(\mathbf{P}(S_I, S^m(w^m))^T C_P^{-T} z - f^m(Y) \right) \end{bmatrix}.
\end{aligned} \tag{2.40}$$

For this problem we obtain the LM updates

$$\begin{aligned}
(\mathbf{w}_{(j+1)}, z_{(j+1)}) &= \arg \min_{\mathbf{w} \in (\mathbb{R}^{QK})^M, z \in \mathbb{R}^I} \|z\|^2 + \frac{1}{M} \sum_{m=1}^M \|w^m\|^2 + \|F_{(j)}(z, \mathbf{w})\|^2 \\
&\quad + \lambda_{(j)} \left(\|z - z_{(j)}\|^2 + \frac{1}{M} \sum_{m=1}^M \|w^m - w_{(j)}^m\|^2 \right) \\
\text{where } F_{(j)}(z, \mathbf{w}) &= \nabla F(z_{(j)}, \mathbf{w}_{(j)}) \begin{bmatrix} z - z_{(j)} \\ \mathbf{w} - \mathbf{w}_{(j)} \end{bmatrix} + F(z_{(j)}, \mathbf{w}_{(j)}),
\end{aligned} \tag{2.41}$$

and $\lambda_{(j)}$ is a proximal regularization parameter to ensure global convergence. In many cases, applying this iteration with a reasonable update rule for λ_j and solving the least squares subproblems by computing Cholesky factorizations of the normal equations is sufficient to achieve accurate solutions. When issues with the conditioning of the normal equations arise, we use an SVD solver which is more accurate but has a higher computational cost.

A heuristic for choosing $\lambda_{(j)}$

We now discuss the adaptation of the damping parameter $\lambda_{(j)}$, which is crucial for both ensuring stability of the optimization algorithm far from a minimum, and providing fast convergence near minima. Following [258, 259], define the *gain ratio* $\rho_{(j)}$ to be the ratio of

the decrease of the true objective value to the decrease predicted by the linearized objective,

$$\begin{aligned}\rho_{(j)} &:= \frac{\theta_{(j)}}{\theta'_{(j)}}, \\ \theta_{(j)} &:= \left(\|z_{(j+1)}\|^2 + \frac{1}{M} \sum_{m=1}^M \|w_{(j+1)}^m\|^2 + \|F(z_{(j+1)}, \mathbf{w}_{(j+1)})\|^2 \right) \\ &\quad - \left(\|z_{(j)}\|^2 + \frac{1}{M} \sum_{m=1}^M \|w_{(j)}^m\|^2 + \|F(z_{(j)}, \mathbf{w}_{(j)})\|^2 \right), \\ \theta'_{(j)} &:= \left(\|z_{(j+1)}\|^2 + \frac{1}{M} \sum_{m=1}^M \|w_{(j+1)}^m\|^2 + \|F_{(j)}(z_{(j+1)}, \mathbf{w}_{(j+1)})\|^2 \right) \\ &\quad - \left(\|z_{(j)}\|^2 + \frac{1}{M} \sum_{m=1}^M \|w_{(j)}^m\|^2 + \|F(z_{(j)}, \mathbf{w}_{(j)})\|^2 \right),\end{aligned}$$

For constants $0 < c_0 < c_1 < c_2 < 1$ and an adaptation multiplier $b > 1$, we set

$$\lambda_{(j+1)} = \begin{cases} b\lambda_{(j)} & \rho_{(j)} < c_1 \\ \lambda_{(j)} & \rho_{(j)} \in (c_1, c_2) \\ \frac{1}{b}\lambda_{(j)} & \rho_{(j)} > c_2. \end{cases}$$

In the case that $\rho_{(j)} < c_0$, we reject the step, set $(z_{(j+1)}, \mathbf{w}_{(j+1)}) = (z_{(j)}, \mathbf{w}_{(j)})$, and attempt to compute the next step with the increased damping parameter.

2.8.4 Leveraging block-arrow sparsity

Instantiating and computing solutions to the regularized linear least squares problems within the LM iterations can be difficult in large scale problems where N, M are large. In such cases, the solution to (2.41) can be computed directly in $O((QK + N_1)^3 M)$ operations, by leveraging sparsity of the gradients ∇F , rather than the naive $O((QK + N_1)^3 M^3)$. Observe that

$$\nabla F(z_{(j)}, \mathbf{w}_{(j)}) = \begin{bmatrix} \nabla_z F_1 & \nabla_{w_1} F_1 & 0 & \cdots & 0 \\ \nabla_z F_2 & 0 & \nabla_{w_2} F_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & 0 & & & 0 \\ \nabla_z F_M & 0 & 0 & \cdots & \nabla_{w_M} F_M \end{bmatrix}$$

This form of ∇F induces a block arrowhead structure for the regularized normal equations for the least squares problems within the LM update,

$$\nabla F^T \nabla F + (1 + \lambda_{(j)})I = \begin{bmatrix} * & * & \cdots & * & * \\ * & * & & & \\ \vdots & & * & & \\ * & & & \ddots & \\ * & & & & * \end{bmatrix} = \begin{bmatrix} A_P & B_{UP}^T \\ B_{UP} & D_U \end{bmatrix}$$

where $A_P \in \mathbb{R}^{N_I \times N_I}$ and $D_U \in \mathbb{R}^{QKM \times QKM}$ is a block diagonal matrix with blocks of size QK . From here, block elimination can be applied to solve the normal equations

$$\begin{bmatrix} A_P & B_{UP}^T \\ B_{UP} & D_U \end{bmatrix} \begin{bmatrix} \delta z \\ \delta w \end{bmatrix} = \begin{bmatrix} F_P \\ F_U \end{bmatrix}$$

given by the formulae

$$\begin{aligned} C &= (A_P - B_{UP}^T D_U^{-1} B_{UP}), \\ \delta w &= C^{-1} (F_P - B_{UP}^T D_U^{-1} F_U), \\ \delta z &= D_U^{-1} (F_U - B_{UP} \delta w), \end{aligned}$$

where we note that the Schur complement C and the matrix A_P are of the moderate size $I \times I$, and the inverse of D_U is readily computable due to the block diagonal structure.

2.8.5 Choosing kernels and hyper-parameters

Choosing U

Since Sobolev spaces are a natural solution space for many PDEs we choose the Matérn kernel family for U. In particular, we consider the anisotropic Matérn kernels $U_{\text{Matérn}}(y, y') := \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu} \|y - y'\|_{\Sigma})^{\nu} K_{\nu}(\sqrt{2\nu} \|y - y'\|_{\Sigma})$, where $\nu > 0$ is a parameter, Γ is the standard gamma function, and K_{ν} is the modified Bessel function of the second kind, and $\|y - y'\|_{\Sigma}^2 := (y - y')^T \Sigma^{-1} (y - y')$ for a PDS matrix $\Sigma \in \mathbb{R}^{d \times d}$. Under mild conditions (see [260, Ex. 2.6, 2.8]) the RKHS of the Matérn kernel is norm equivalent to the Sobolev space $H^{\nu+d/2}$. Thus it is crucial that the regularity parameter ν is chosen carefully in light of the order of the PDE. In the limit $\nu \rightarrow \infty$ the Matérn kernel converges to the radial basis function (RBF) kernel $U_{\text{RBF}}(y, y') := \exp(-\frac{1}{2} \|y - y'\|_{\Sigma}^2)$ which has an infinitely smooth RKHS. In addition to this kernel we also use the first order rational quadratic (RQ) kernel, defined as $U_{\text{RQ}}(y, y') := (1 + \|y - y'\|_{\Sigma}^2)^{-1}$. We often take Σ to be a diagonal matrix $\Sigma = \text{diag}(\gamma)$ for a vector of lengthscales $\gamma \in \mathbb{R}_{>0}^d$ to be chosen via maximum likelihood estimation (MLE), cross validation, or hand tuning.

Choosing P

Since many PDEs in physical sciences have polynomial nonlinearities [256, Sec. 1.2], a natural choice for P would be a polynomial kernel; this is also the dominant model for construction of dictionaries in SINDy-type algorithms. Here we consider $P_{\text{poly}}(s, s') = ((s - c)^T B (s' - c) + 1)^{\text{deg}}$, for $s, s' \in \mathbb{R}^{Q+d}$, where $B \in \mathbb{R}^{Q+d \times Q+d}$ is a matrix akin to Σ in the previous section that allows us to scale different input coordinates or to introduce correlations, and $c \in \mathbb{R}^{Q+d}$ is a fixed vector introducing bias. We treat B, c as hyper-parameters for this kernel. In extreme scenarios, if no a priori knowledge about P exists, then one can take P to be a Matérn or RBF kernel as in the case of U above.

We are particularly interested in learning PDEs with variable coefficients, such as the variable coefficient diffusion model (2.5). In such cases we may have a priori knowledge that P has polynomial dependence on a subset of inputs (the $L_1 u(y), \dots, L_Q u(y)$ values) but

general nonlinear dependence on other parameters (the y values). We encode such structures using hybrid product kernels of the form $P_{\text{hybrid}}(s, s') = P_{\text{RBF}}(s_{:d}, s'_{:d})P_{\text{poly}}(s_{d:}, s'_{d:})$ where we used $s_{:d} \in \mathbb{R}^d$ to denote the first d coordinates of s and $s_{d:} \in \mathbb{R}^Q$ to denote the remaining coordinates.

2.9 Details of numerical experiments

We now present details of the numerical experiments summarized in the main body of the chapter, covering various details such as data generation processes, selection of observation and collocation points, hyperparameter, and additional results and observations. Further implementation details can be found in our GitHub repository⁴ where we collect code for regenerating our numerical results and figures.

2.9.1 The Duffing oscillator

In this example we compared the performance of 1-step and 2-step KEqL along with SINDy for learning a 1D nonlinear ODE. Our focus is on the performance of the methods in filtering and extrapolation/forecasting of the dynamics.

Let us recall our problem setup where we took $\mathcal{Y} = (0, 50)$ and considered the Duffing oscillator

$$\begin{cases} \mathcal{P}(u) = \partial_t^2 u - 3u + 3u^3 + 0.2\partial_t u = \cos(2t), & t \in \mathcal{Y}, \\ u(0) = \partial_t u(0) = 0. \end{cases} \quad (2.42)$$

To generate the training data for this example we solved the ODE numerically using a Dopri5 solver in Python with adaptive step size and initial value $1e - 3$. The numerical solution was then subsampled on $N = 32$ observation points that were also picked on a uniform collocation grid Y of size 1000 in $\mathcal{Y}$. The test data set for filtering the solution was generated similarly but on a finer uniform grid of 5000 observation points. To generate the test datasets for forecasting and extrapolation errors, we used the same numerical solver, but, we introduced three new initial conditions, each satisfying $u(0) = 0$, while the initial time derivative $\partial_t u(0)$ was set to 0.5, 1, and -1 , respectively.

For the kernels, we selected U as a RQ kernel and P as RBF, both with $\Sigma = \text{Id}$. We employed a fixed basis with $I = 500$ inducing points for learning the equation (see Section 2.8.2).

For the loss function Section 2.8.2, we used

$$\lambda = 1, \quad \sigma_u^2 = \sqrt{\frac{NM}{\theta_u}}, \quad \text{and} \quad \sigma_P^2 = \sqrt{\frac{MI}{\theta_P}}, \quad (2.43)$$

with $\theta_u = 5e - 8$ and $\theta_P = 1e - 9$, selected manually. The loss history for the optimization of 1-step KEqL can be seen in Figure 2.5 for learning the Duffing ODE. We see that the algorithm has converged in about 200 iterations.

⁴<https://github.com/TADSGroup/kernelequationlearning>

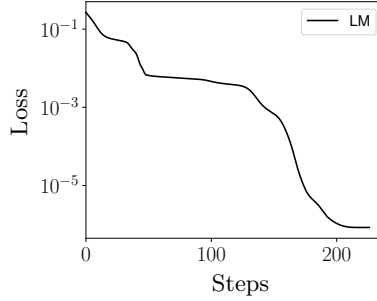


Figure 2.5: Convergence history of LM for 1-step KEqL for the Duffing ODE (2.18).

2.9.2 The Burgers' PDE

In this example we compared 1-step KEqL with SINDy and the PINN-SR algorithm. Here we describe the choice of initial conditions, generation of training and test samples, hyper-parameters, and optimization details used for this example.

Let us recall our problem setting. We took $\mathcal{Y} = (0, 1] \times (0, 1)$ and considered the Burgers' PDE

$$\begin{cases} \mathcal{P}(u) = \partial_t u + \vartheta u \partial_x u - \nu \partial_x^2 u = 0 \text{ for } (t, x) \in \mathcal{Y}, \\ u(0, x) = u_0(x), \text{ and } u(t, 0) = u(t, 1) = 0. \end{cases} \quad (2.44)$$

We remind the reader that whenever we state in any of the subsequent experiments that (2.19) was solved, we mean that a second-order Strang splitting method with a small step size was used, followed by spline interpolation to obtain a solution that can be evaluated at any point of its domain. Additionally, when using the 1-step KEqL method, we always assumed the use of the standard implementation of the algorithm described in (2.39).

Experiments with increasing number of observations

Here, we used the PDE (2.19) with coefficients $\vartheta = 1$ and $\nu = 0.01$ and prescribed initial conditions (IC) u_0 from a Gaussian process with Karhunen–Loève expansion

$$u_0(x) = \sum_{j=1}^{50} \frac{1}{j^2} \sin(j\pi x) Z_j, \quad \text{with } Z_j \sim N(0, 1), \quad x \in [0, 1]. \quad (2.45)$$

The condition used for the fixed IC case is depicted in Figure 2.6(A) while some samples of ICs are shown in Figure 2.6(B-D) for the case when ICs vary also in the experiment.

We then solved each PDE and subsampled the observed data for values $N_{\mathcal{Y}} = 10, 30, 50, 100, 200, 300, 400, 500, 600$ in the interior and $N_{\partial\mathcal{Y}} = 131$ at the boundary from a set of 26×31 Chebyshev collocation points in space-time. In Figure 2.7, we present example training data overlaid on the contour plots of the solutions for the fixed initial condition, as shown in Figure 2.6(A), while varying the number of observations $N_{\mathcal{Y}} = 20, 50, 200$. Additionally, we include instances of the training data for the case where the initial condition varies, corresponding to the ICs introduced earlier in Figure 2.6(B-D).

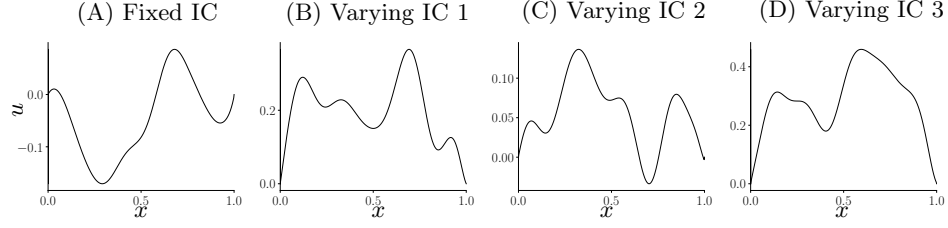


Figure 2.6: Initial conditions sampled from the process (2.45) for examples of Burgers' PDE (2.19) increasing number of observations: (A) Condition used in the fixed IC case; (B-D) Sample conditions for the varying IC case.

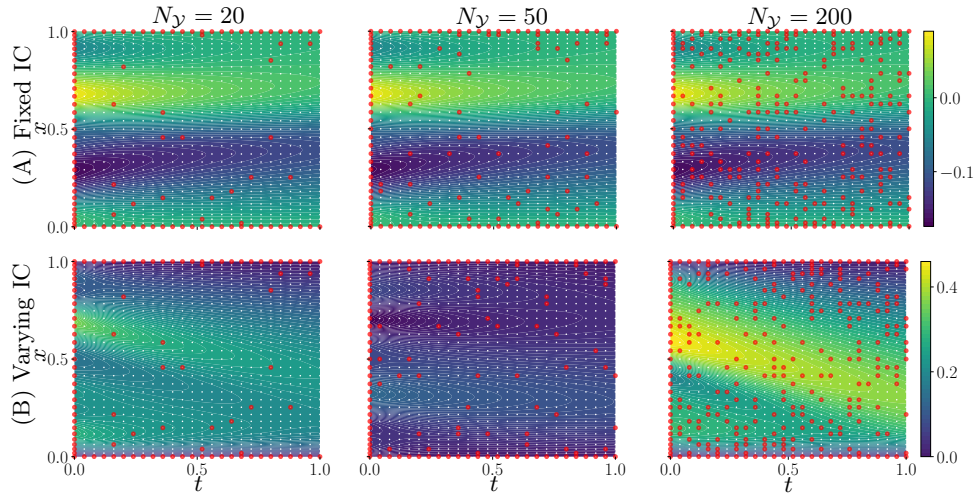


Figure 2.7: Samples of training solutions along with their N observation points (red) and the collocation points (white) for Burgers' PDE (2.19) with ICs sampled from the GP (2.45): (A) Solution obtained using a fixed IC depicted in Figure 2.6(A); (B) Three solutions obtained from solving the PDE with different ICs depicted in Figure 2.6(B-D).

We took $\mathbf{U}$, to be RBF with a diagonal matrix Σ , where its entries were determined using maximum likelihood estimation (MLE) based on the available observation points $u(Y^1)$ for each number of observations case, basically by a first application of the 2-step method. In both cases for this experiment, we chose $\mathbf{P}$ to be a polynomial kernel of degree 2, with the shift parameter set to $c = \text{mean}(S_{2,:})$ and the scaling matrix given by $B = \text{diag}(\text{cov}(S_{2,:}))^{-1}$, where $\text{cov}(A)$ represents the sample covariance of A . Specifically,

$$S = \Phi(\hat{u}_2, Y^1) := \begin{bmatrix} \Phi(\hat{u}_2, y_1) \\ \vdots \\ \Phi(\hat{u}_2, y_N) \end{bmatrix} \in \mathbb{R}^{N \times 5}, \quad (2.46)$$

where $\hat{u}_2$ is the learned solution obtained from the first step of the 2-step KEqL method, using the MLE-fitted hyperparameters mentioned above.

Recall that for SINDy, we used the same kernel U for derivative approximation and employed a polynomial library for the dictionary terms, which matched the features of the degree-2 polynomial used in KEqL. These same features were also used by PINN-SR. For PINN-SR, we fixed the hyperparameters across different experiments. For example, in the fixed IC case, where we aim to learn a PDE by varying the number of observation points, we selected a set of hyperparameters—such as network size, loss function weights, and iteration steps—and kept them the same for every experiment with different numbers of observation points. The same approach was applied when varying the initial condition. This strategy is consistent with the kernel method and SINDy, as we also maintained the same hyperparameters across different experiments for those methods.

For the loss function (2.39), we followed the definitions in (2.43), with the exception that $I = 26 \times 31$ and we set $\theta_u = 1e - 9$ and $\theta_P = 1e - 11$ as manually chosen hyperparameters.

Convergence history of the optimization algorithm for some sample cases of learning the equation is shown in Figure 2.12(A-B).

An experiment with a smoothed shock

Here, we considered the PDE (2.19) with coefficients $\vartheta = 5$ and $\nu = 0.001$, using the prescribed IC

$$u_0(x) = \frac{1}{4} (x \sin(\pi x) - \sin(5\pi x) - \sin(3\pi x)). \quad (2.47)$$

We then solved the PDE, and the solution's behavior at the initial and final times is shown in Figure 2.10(A). We then subsampled $N_{\mathcal{Y}} = 60$ interior observation points and $N_{\partial\mathcal{Y}} = 131$ boundary points from a selected 26×31 Chebyshev collocation grid in space-time, as illustrated in Figure 2.8.

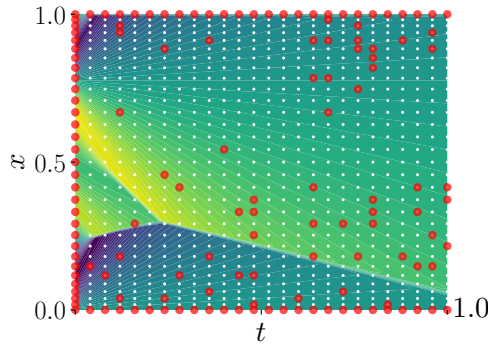


Figure 2.8: Solution of the Burgers' PDE (2.19) with initial condition (2.47). Sampled observation points are shown in red, while collocation points are overlaid in white.

We took U , to be RBF with a diagonal matrix Σ , where its entries were determined using MLE based on the N observation points. We chose P to be a polynomial kernel of degree 2, with the shift parameter and the scaling matrix constructed using S as in (2.46). Thus, once again we used the hyperparameter information from the 2-step KEqL methodology to

guide the selection of hyperparameters for the 1-step KEqL. For the loss function (2.17), we followed the definitions in (2.43), with the exception that $I = 26 \times 31$ and we set $\theta_u = 1e - 9$ and $\theta_P = 1e - 11$ as manually chosen hyperparameters. For SINDy we used the same kernel $\mathbf{U}$ for derivative approximation and employed a polynomial library for the dictionary terms, which matched the features of the degree-2 polynomial used in KEqL. These same features were also used by the PINN-SR method.

We clearly observe that the 1-step KEqL method remains competitive in shock filtering. This is evident from the relative filtering errors presented in Figure 2.11, where the 1-step KEqL achieves comparable results in shock recovery to those of SINDy and PINN-SR. Convergence history of the optimization algorithm for learning the equation is shown in Figure 2.12(C).

An ablation experiment with a smoothed shock

Here, we consider once more the PDE (2.19) with coefficients $\vartheta = 5$ and $\nu = 0.001$, subject to the initial condition (2.47), with same observations subsampled from the 26×31 Chebyshev collocation grid in space-time.

To learn the dynamics, we adopt the same kernels and hyperparameters for each method as in Section 2.9.2, but now test two variants per method: one in which the degree-2 polynomial features are restricted to $\{1, u, \partial_x u\}$ (excluding $\partial_x^2 u$), and one in which the feature set is expanded to $\{1, u, \partial_x u, \partial_x^2 u, \partial_{tx} u, \partial_t^2 u\}$. Results are reported in Figure 2.11.

Overall, the results across all methods share the same order of error, with the expanded feature set yielding more accurate results than the reduced one, in which the diffusion term $\partial_x^2 u$ is omitted. We note that, upon removal of the diffusion feature, the PINN-SR method reduces to predicting an equation containing only the u term, in contrast to the 1-step kernel method, which retain a linear combination over the full feature set. The kernel interpolant is unchanged as it is independent of the step of inferring the dynamics.

A one-shot learning experiment

Here, we considered the PDE (2.19) with coefficients $\vartheta = 0.5$ and $\nu = 0.01$, using the prescribed IC given by (2.47). We then solved the PDE, whose solution behavior at initial and final times are shown in Figure 2.10(B). Here we subsampled $N = 100$ points placed on the boundary, see first column in Figure 2.3(A). The placement of the points amounts to observing the initial and end conditions along with the boundary values in the physical domain. To generate data for operator learning, we additionally solved the same PDE with a new initial condition, $u_0(x) = -x \sin(2\pi x)$. The solution behavior at the initial and final times is shown in Figure 2.10(C).

In this experiment we took $\mathbf{U}$ to be RBF with the diagonal matrix $\Sigma = 0.0125I$. To learn the equation, we used $\tilde{\mathbf{P}}$ as a polynomial kernel of degree 2, where the shift parameter and scaling matrix were constructed using S , as described in (2.46). Here again, we used the hyperparameter information from the 2-step KEqL methodology to guide the selection of hyperparameters for the kernel $\mathbf{P}$ of the 1-step KEqL. Recall that a similar rationale was used for selecting hyperparameters in SINDy and PINN-SR, including the choice of dictionary terms, as in the previous experiment discussed in Section 2.9.2. For the loss

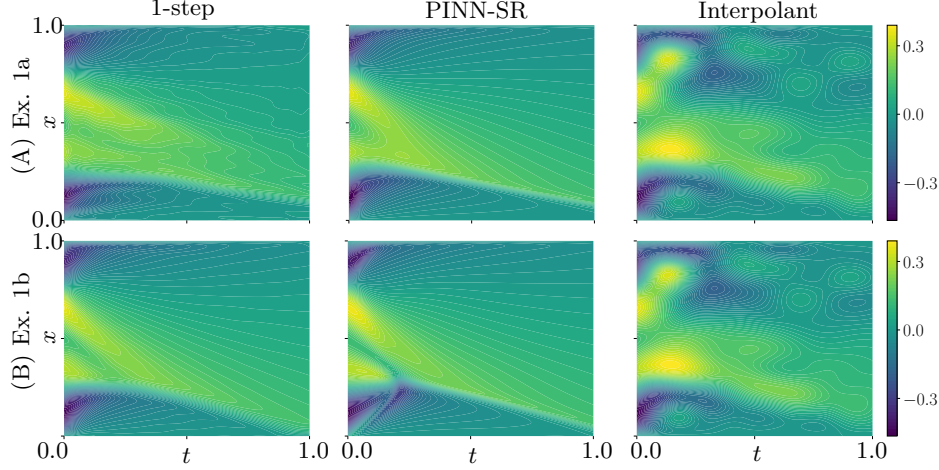


Figure 2.9: Representative numerical results for an ablation study on Burgers' PDE (2.19) with initial condition (2.47), examining the effect of removing the diffusion feature across the 1-step KEqL, PINN-SR, and kernel interpolation methods: (A) solutions recovered using the reduced feature set $\{1, u, \partial_x u\}$; (B) solutions recovered using the expanded feature set $\{1, u, \partial_x u, \partial_x^2 u, \partial_{tx} u, \partial_t^2 u\}$.

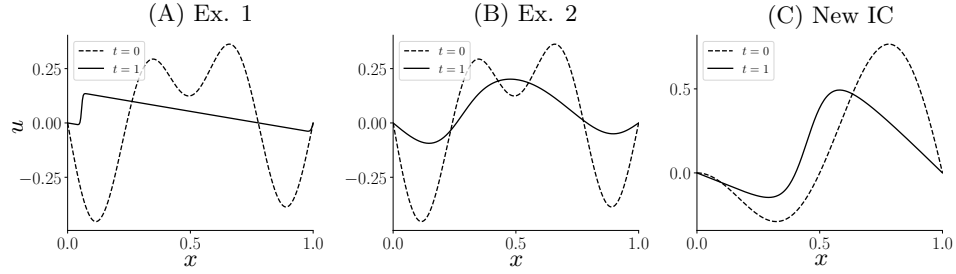


Figure 2.10: Behavior of solutions at initial and final time for one-shot learning examples of Burgers' PDE (2.19): (A) Shock development (Section 2.9.2); (B) Smooth transition (Section 2.9.2); (C) Solution behavior to the PDE in (B) for the depicted new initial condition.

function (2.17), we omitted the terms associated with the RKHS norms, divided all the terms by 10 and selected values for σ_u^2 and σ_P^2 as defined in (2.43). For which we set $I = 26 \times 26$. Additionally, we manually chose $\theta_u = 100$ and $\theta_P = 1$ as hyperparameters.

We clearly observe that the 1-step KEqL method outperforms the other methods in filtering the smooth solution. This is evident from results presented in Figure 2.11, where the one-step KEqL achieves a two-order-of-magnitude improvement over the other methods in filtering and a one-order-of-magnitude improvement in the operator learning task for the smooth case. The one-shot learning example for the smooth case further confirms that effective filtering is essential for accurately capturing the operator. While PINN-SR and SINDy provided a reasonable reconstruction of the solution, the superior filtering of the 1-step KEqL method resulted in a significantly better operator learning error for a new initial condition. We recall that the variational solution proposed in (2.13) was used to solve the

		Method		
	Ex.	1-step	PINN-SR	SINDy
$\mathcal{R}_{\text{filter}}$	1	$3.9e^{-1}$	$3.4e^{-1}$	$6.2e^{-1}$
	1a	$4.7e^{-1}$	$3.6e^{-1}$	$6.2e^{-1}$
	1b	$3.7e^{-1}$	$3.2e^{-1}$	$6.2e^{-1}$
	2	$3.5e^{-3}$	$1.5e^{-1}$	$2.9e^{-1}$
$\mathcal{R}_{\text{opl}}$	2	$1.9e^{-2}$	$4.7e^{-1}$	$5.4e^{-1}$

Figure 2.11: Quantitative values of relative filtering and operator learning errors for Burgers' PDE Equation (2.19) with different ICs corresponding to the one-shot learning examples Ex. 1 (Section 2.9.2), its ablation studies Ex. 1a–b, in which features are removed and added (Section 2.9.2), and Ex. 2 (Section 2.9.2) in Figure 2.3(C–E).

recovered PDE in each of the methods for this experiment, employing the same kernel U , to report the operator learning results. Convergence history of the optimization algorithm for learning the equation is shown in Figure 2.12(D).

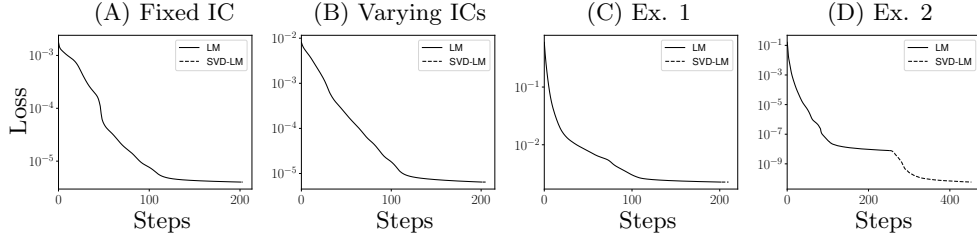


Figure 2.12: Convergence history of the LM algorithm for 1-step KEqL for Burgers' PDE (2.19): (A–B) Sample cases using fixed and varying IC, using $M = 1$ training pairs with number of observations $N_{\mathcal{Y}} = 20$ corresponding to the two cases depicted in the first column in Figure 2.7; (C) Shock development (Section 2.9.2); (D) One-shot smooth case where only observed on the boundary (Section 2.9.2).

2.9.3 Darcy's flow PDE

In this example we performed a systematic study of the performance of 1-step (its reduced version) and 2-step KEqL for learning an elliptic PDE with a variable diffusion coefficient. In particular, we investigate the ID and OOD performance in terms of filtering, equation learning, and operator learning.

Let us recall our problem setting. Here we took $\mathcal{Y} = (0, 1) \times (0, 1)$ and considered the problem

$$\begin{cases} \mathcal{P}(u) = \operatorname{div}(a \nabla u) = f, & \text{in } \mathcal{Y}, \\ u(y) = g(y), & \text{for } y \in \partial \mathcal{Y}, \end{cases} \quad (2.48)$$

where $a(y) = \exp(\sin(\cos(y_1) + \cos(y_2)))$. To generate the training data we drew $M = 2, 4, 8, 16, 32$ functions $u \sim \mathcal{GP}(0, \mathbf{U}_{\text{RBF}})$ with $\Sigma = 0.5^2 I$ and took $f := \mathcal{P}(u)$, also considering the value of u at $\partial\mathcal{Y}$ as the boundary condition. Each u was then subsampled observation points with $N_{\mathcal{Y}} = 2, 4, 8$ in the interior and fixed $N_{\partial\mathcal{Y}} = 56$ at the boundary that were picked randomly from a 15×15 uniform collocation grid where the PDE was enforced. Some training tuples with their observation and collocation points are depicted in Figure 2.13. We recall that the ID data was drawn from the same distribution that the training data was drawn from. The test data sets were generated similarly, with the OOD data drawn from a less regular GP using an RBF kernel with $\Sigma = 0.4^2 I$. Some ID and OOD samples are shown at Figure 2.14(A). The fine grid $\mathcal{Y}_{\text{test}}$ where we tested the equation and operator learning error was chosen to be a uniform grid of size 100^2 .

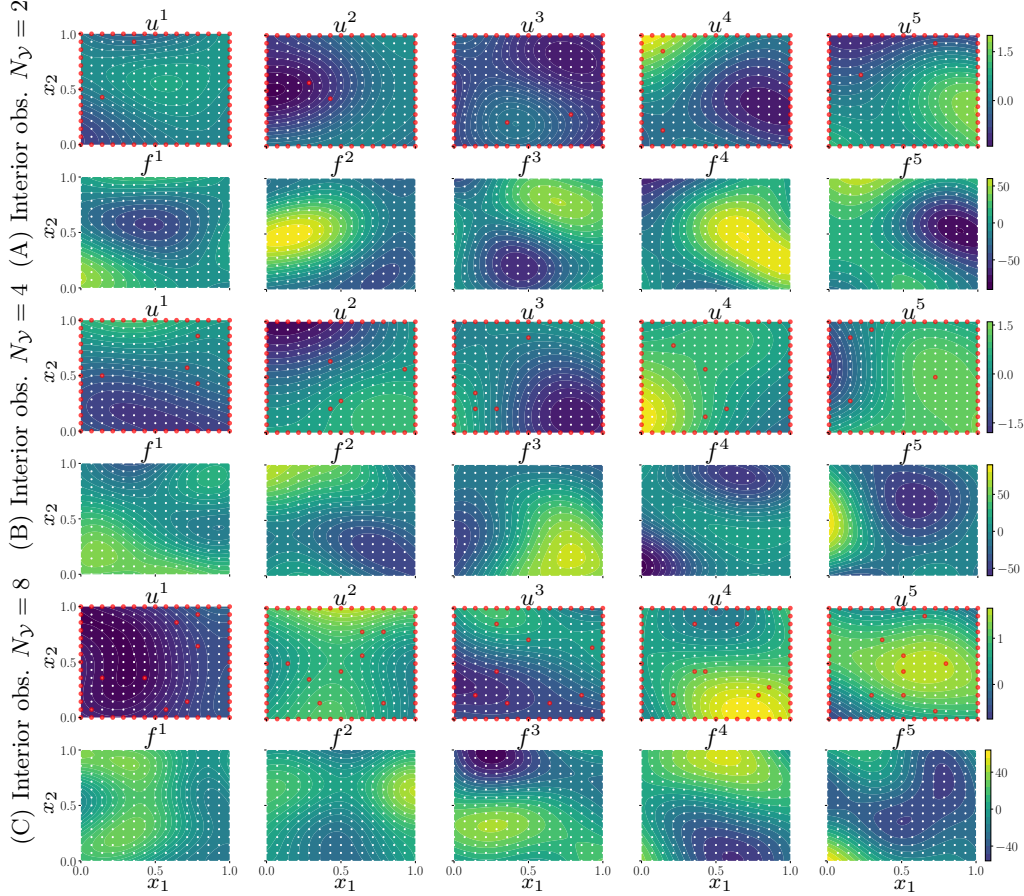


Figure 2.13: Samples of training tuples (u^m, f^m) with different number of observation points for Darcy's flow PDE (2.20): (A-C) Training tuples observation points (red) and collocation points (white).

We used $\mathbf{U} = \mathbf{U}_{\text{RBF}}$ with $\Sigma = 0.5^2 I$ for learning u and $\mathbf{P} = \mathbf{P}_{\text{hybrid}}$ where an RBF with

$\Sigma = 0.4^2 I$ was used for the spatial variables x_1, x_2 and a polynomial kernel of first degree for the variables $\{L_1 u, \dots, L_6 u\}$ where the shift was chosen to be $c = \text{mean}(S_{2,:})$ and the scaling matrix $B = (\text{diag}(\text{cov}(S_{2,:})))^{-1}$ where

$$S = \begin{bmatrix} \Phi(\hat{u}_2^1, Y^1) \\ \vdots \\ \Phi(\hat{u}_2^M, Y^M) \end{bmatrix} \in \mathbb{R}^{\sum_{m=1}^M |Y^m| \times 8}.$$

where $\hat{u}_2^m$ is the learned solution obtained from the first step of the 2-step KEqL method, using the MLE-fitted hyperparameters mentioned above. In summary, we also used the hyperparameter information from the 2-step KEqL methodology to guide the selection of hyperparameters for the kernels in 1-step KEqL.

To scale up equation learning when using many function tuples, we employed a fixed basis (see Section 2.8.2) with $I = 200$ for learning the equation. Additionally, we leveraged the block arrowhead matrix structure within the optimization algorithm (see Section 2.8.4). For the loss function (2.39), we followed the definitions in (2.43) except we took $\sigma_P^2 = \sqrt{I/\theta_P}$ and set $\theta_u = 5e - 12$ and $\theta_P = 1e - 12$ as hyperparameters, selected manually.

To further illustrate the equation and operator learning errors in relation to the error plot in Figure 2.4, we show samples of the contour error plots in Figure 2.14 and the convergence history in Figure 2.15 that uses the block arrowhead matrix structure LM-type algorithm for the case where the equation is learned using $M = 8$ training tuples and $N_y = 8$ observation points of the solution. We recall that the variational solution proposed in (2.13) was used to solve the recovered PDE for both the 1-step and 2-step KEqL methods, employing the same kernel U , to report the operator learning results.

In Figure 2.14(A), we show some ID and OOD solution samples used to test the learned equation for both the 1-step and 2-step KEqL methods. In Figure 2.14(B), we present the equation learning errors, where the values typically differ by one order of magnitude between 1-step and 2-step KEqL in favor of the joint methodology, with larger errors appearing near the boundaries. Similarly, in Figure 2.14(C), we observe that the 1-step method consistently outperforms the 2-step method by an order of magnitude. Notably, for this case the equation learning errors were an order of magnitude smaller than the operator learning errors.

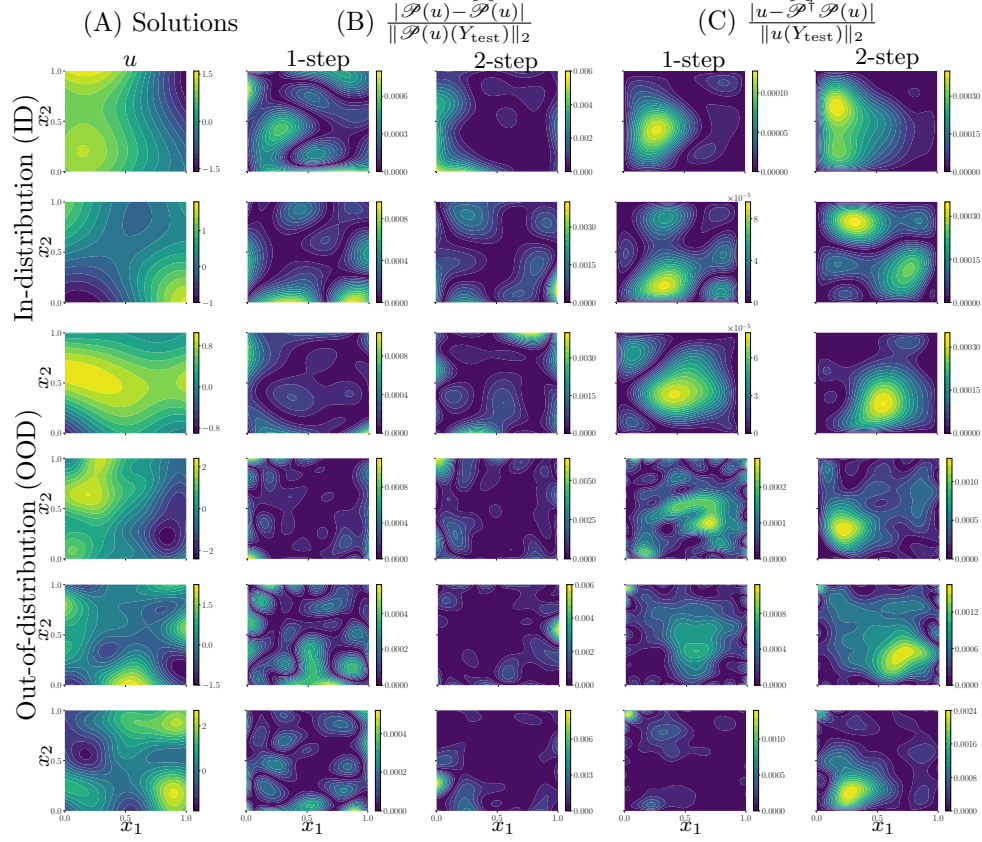


Figure 2.14: Samples of test solutions u (ID and OOD) and their contours of the equation and operator learning errors for 1-step and 2-step KEqL methods for Darcy's flow PDE (2.20): (A) Sample solutions u ID and OOD for the PDE ;(B) Contour plots for equation learning errors; (C) Contour plots for operator learning errors.

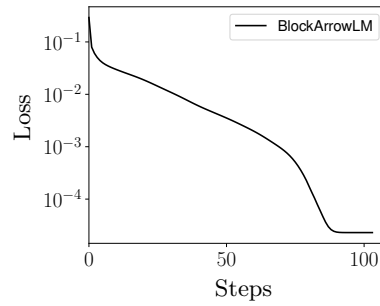


Figure 2.15: Convergence history of the block arrowhead matrix structure LM-type algorithm for the reduced 1-step KEqL when using $M = 8$ solution tuples at $N_{\mathcal{Y}} = 8$ observation points in the interior for Darcy's flow PDE (2.20).

Chapter 3

A JOINT OPTIMIZATION APPROACH TO IDENTIFYING SPARSE DYNAMICS USING LEAST SQUARES KERNEL COLLOCATION

Preface

In this chapter, from [3] which was joint work with Ike Griss Salas, Jacob M. Stevens-Haas, J. Nathan Kutz, Aleksandr Aravkin, and Bamdad Hosseini, we build upon the least squares kernel collocation approach of Chapter 2 to develop a framework for jointly identifying sparse ordinary differential equations and reconstructing the underlying state from scarce, partial, and noisy observations. We consider a function $\mathbf{x}(t)$ and measurements $\mathbf{y}_n$ satisfying

$$\frac{d^p}{dt^p}\mathbf{x}(t) = f(D[\mathbf{x}](t)), \quad \mathbf{y}_n = \mathbf{V}_n^\top \mathbf{x}(t_n) + \boldsymbol{\epsilon}_n, \quad (3.1)$$

$$D[\mathbf{x}](t) := \left(\mathbf{x}(t), \frac{d}{dt}\mathbf{x}(t), \dots, \frac{d^{p-1}}{dt^{p-1}}\mathbf{x}(t) \right), \quad (3.2)$$

where p is the order of the ODE, $\mathbf{V}_n$ are measurement matrices, t_n are observation times, and $\boldsymbol{\epsilon}_n$ are i.i.d. zero-mean noise. We develop algorithms for recovering the function f .

We formulate a similar joint least squares collocation objective as in Chapter 2, but instead represent the unknown differential equation as a sparse linear combination of functions from a fixed library. This representation has two major benefits. It improves the sample complexity when the true function f possesses a sparse representation in the library, and results in a differential equation that can be read off and directly interpreted by a practitioner. In order to enforce this sparsity of the solution, we apply an optimization algorithm which alternates between a Levenberg–Marquardt algorithm on a fixed-support nonlinear least squares problem, and applying an arbitrary sparse regression procedure to identify an improved support.

From extensive experiments, we see significant improvements in terms of data requirements and noise robustness compared to existing algorithms when applicable, while also extending the flexibility of existing methods in settings of partially observed and higher-order dynamics where existing methods often do not apply.

3.1 Introduction

The identification of ordinary differential equations (ODEs) and dynamical systems is a fundamental problem in control [146, 261, 262], data assimilation [263, 264], and more recently in scientific machine learning (ML) [265–267]. While algorithms such as Sparse Identification of Nonlinear Dynamics (SINDy) and its variants [268] are widely used by practitioners, they often fail in scenarios where observations of the state of the system are scarce, indirect, and noisy. In such scenarios modifications to SINDy-type methods

are required to enforce additional constraints on the recovered equations to make them consistent with the observational data. Put simply, traditional SINDy-type methods work in two steps: (1) the data is used to filter the state of the system and estimate the derivatives, and (2) the filtered state is used to learn the underlying dynamics. In the regime of scarce, noisy and incomplete data, step 1 is inaccurate, which can propagate to poor results in the subsequent step 2.

In this paper, we propose an all-at-once approach to filtering and equation learning based on collocation in a reproducing kernel Hilbert space (RKHS) which we term Joint SINDy (JSINDy), and show that the issues above can be mitigated by performing both steps together. This joins a broader class of dynamics-informed methods that integrate the governing equations directly into the learning objective, either as hard constraints or as least-squares relaxations, which couples the problems of state estimation and model discovery. Representative examples include physics-informed and sparse-regression frameworks based on neural networks, splines, kernels, finite differences, and adjoint methods [2, 141, 159, 160, 266, 269, 270].

3.1.1 Problem setup

Consider a process $\mathbf{x}(t) = (x_1(t), \dots, x_d(t)) \in C^p([0, T], \mathbb{R}^d)$, for $d \geq 1$, along with the observables $(\mathbf{y}_n)_{n=1}^N$ that are connected via the model

$$\frac{d^p}{dt^p} \mathbf{x}(t) = f(D[\mathbf{x}](t)), \quad \mathbf{y}_n = \mathbf{V}_n^\top \mathbf{x}(t_n) + \boldsymbol{\epsilon}_n, \quad (3.3)$$

$$\text{where } D[\mathbf{x}](t) := \left(\mathbf{x}(t), \frac{d}{dt} \mathbf{x}(t), \dots, \frac{d^{p-1}}{dt^{p-1}} \mathbf{x}(t) \right), \quad (3.4)$$

with known measurement matrices $\mathbf{V}_n$ (taken to be the identity when the entire state is observed), observation times $\mathbf{t} = (t_n)_{n=1}^N$, and measurement noise $\boldsymbol{\epsilon}_n$ which are assumed to be independent with zero mean and bounded variance. Given these observations, our goal is to estimate the nonlinear map $f : \mathbb{R}^{dp} \mapsto \mathbb{R}^d$ from an instance of the $\mathbf{y}_n$, i.e., to find $\hat{f} \approx f$, and simultaneously infer the process $\mathbf{x}(t)$, i.e., find $\hat{\mathbf{x}}(t) \approx \mathbf{x}(t)$. Additionally, we ask that our representation of $\hat{f}$ is *sparse* in some basis or library of functions $\Phi = \{\phi_1, \dots, \phi_Q\}$. That is, we assume

$$\hat{f}(D[\mathbf{x}](t))_m = \sum_{q=1}^Q \hat{\theta}_{m,q} \phi_q(D[\mathbf{x}](t)), \quad \hat{f}(D[\mathbf{x}]) = (\hat{f}(D[\mathbf{x}](t))_1, \dots, \hat{f}(D[\mathbf{x}](t))_d), \quad (3.5)$$

where only a small number of $\hat{\theta}_{m,q}$ are nonzero in order to have an interpretable symbolic form, following the seminal work [138, 148, 265, 271]. As we do not know *a priori* which features ϕ_q are active, we parametrize $\hat{f}$ using the full feature library and promote sparsity in the coefficients $\hat{\boldsymbol{\theta}} := (\hat{\theta}_{m,q}), m = 1, \dots, d, q = 1, \dots, Q$.

Our setup poses a number of fundamental challenges:

1. Broadly speaking we would like to find $\hat{f}$ such that

$$\hat{f}(D[\mathbf{x}](t)) \approx \frac{d^p}{dt^p} \mathbf{x}(t).$$

However, the input to $\hat{f}$ as well as the right hand side of the above display depend on $\mathbf{x}(t)$ and its derivatives, both of which are unknown and not observed directly.

2. Exacerbating this issue is the fact that the measurements $\mathbf{y}_n$ are sparse, indirect, and noisy, which leads to further challenges in estimating $\mathbf{x}$ and its derivatives.

Estimating and dealing with the errors in the inputs to $\hat{f}$ and the target $\frac{d^p}{dt^p}\mathbf{x}(t)$ is the fundamental property of system identification problems, which differentiates it from standard approximation or ML tasks. If *all* of these variables were fully observed, i.e., we have direct access to the exact values of $((D[\mathbf{x}](t_n))_{n=1}^N$, as well as unbiased estimates of $\frac{d^p}{dt^p}\mathbf{x}(t_n)$, then this would reduce to a standard sparse regression problem.

We overcome the above challenges by proposing a joint estimation approach where an approximation $\hat{\mathbf{x}}(t)$ to the true state $\mathbf{x}(t)$ is obtained at the same time as the approximation $\hat{f}$ to the nonlinear dynamics. The key idea in our approach is to enforce self consistency of the $\hat{\mathbf{x}}$ and $\hat{f}$, i.e., $\hat{\mathbf{x}}$ should approximately satisfy the ODE defined by $\hat{f}$ while matching the observed data. Since the data is limited and the problem is severely ill-posed, we propose an optimization approach to solve the problem with appropriate regularization terms for both the $\boldsymbol{\theta}$ parameters and the approximate state $\hat{\mathbf{x}}$, in particular, we use a sparse regularizer for the former and a reproducing kernel Hilbert space (RKHS) penalty for the latter.

3.1.2 Summary of contributions

In view of SINDy and other existing works on equation discovery [155, 159, 160, 265, 269, 270], we summarize our contributions as follows:

1. We develop an RKHS regularized least-squares approach to learning general ODEs directly from sparse and partial observations. (see Section 3.2).
2. We establish a representer theorem that converts the infinite-dimensional problem into a finite-dimensional nonlinear least-squares system, and design an alternating optimization algorithm based on the Levenberg–Marquardt (LM) method and sparsifying iterations for support selection (see Sections 3.2.2 and 3.3).
3. We demonstrate the flexibility and robustness of JSINDy across a broad range of problems, including scarce and partial observations, higher-order systems, and misspecified models, demonstrating accurate recovery in various challenging data regimes (see Section 3.4).

Our formulation of the problem differs from previous works in two important ways. We allow for arbitrary order ODEs, while previous works are tailored to $p = 1$, and we specifically consider partial, scarce, and scattered observations while previous works often assume access to high resolution full state observations, i.e., $\mathbf{y}_n = \mathbf{x}(t_n) + \boldsymbol{\epsilon}_n$ on a fine grid.

3.1.3 Literature Review

Learning continuous time dynamical systems from time series data has been a prominent problem in system identification [145–147]. Seminal works using symbolic regression [143, 144] through genetic programming sought to discover physical laws from observational data. The works [138, 148, 265, 271] developed and popularized the use of a sparsity prior over a feature library to promote parsimony in the discovered model, drawing on ideas from compressive sensing [272–274]. Bayesian approaches have also been considered, which allow for uncertainty quantification and motivate new strategies for model selection [152–157].

Many of these works consider a two-step framework which first estimates time-derivatives, then performs regression to estimate the dynamics. These approaches are straightforward to implement and often succeed when measurements of the process are sufficiently dense but fail when the data is scarcely sampled and noisy. These approaches first solve a smoothing problem to estimate the state and its derivatives, using methods such as total variation regularization [275], kernel regression [180], Savitzky–Golay filters, smoothing splines [276], or Kalman smoothing [277]. These approaches inevitably trade bias for robustness: as shown in [278], noise-resistant estimators lie on a Pareto frontier balancing error magnitude and error correlation due to implicitly assuming an inaccurate underlying dynamical model. An alternative to estimating derivatives, but that still lies within a two-step framework is to adopt a weak-form formulation [233], which avoids explicit differentiation by matching integrals against test functions. However, this still requires estimating integrals from noisy and scarcely sampled data, a problem that remains challenging despite being more well posed than differentiation. Consequently, any two-step approach to learning dynamics is inherently limited by the absence of a proper dynamical model in its initial estimation stage.

To address limitations of two-step approaches, dynamics-informed approaches jointly estimate the dynamics and the state, allowing each to reinforce the other. These include Physics-Informed Neural Networks-Sparse Regression (PINN-SR) [160] and Physics-Informed Spline Learning (PiSL) [159] which use least squares collocation applied to neural networks and splines respectively, while obtaining sparsity in the coefficients of the ODE by alternating between applying variants of gradient descent and sequentially thresholded ridge regression. [2] takes a least squares collocation approach using RKHS methods, while also representing the unknown differential equation in an RKHS. The ODR-BINDy method in [269] applies a least squares relaxation to a finite difference discretization, and handles feature selection by greedily maximizing a Bayesian evidence over models defined by feature sets. [279] also applies a least squares relaxation to a finite difference discretization, but optimizes using a variational annealing approach, and focuses on identifying dynamics with hidden variables. Notably, our objective function (3.10) is closely related to that of [159, 269, 279, 280] and [160], though we apply different discretization, variable selection, and optimization strategies. [270] applies a finite difference discretization but imposes hard ODE constraints by solving the KKT system associated to the discretized problem. They promote sparsity via ℓ_p regularization, and optimize using iteratively reweighted least squares.

Aside from [270], which considers decreasing sampling frequencies, and [279], which considers hidden variables, none of these works deal with very low sampling frequencies, partial observations through general linear measurements, and higher order dynamics in the context of ODEs. These issues are, however, considered heavily in [160] and [2] in the

context of PDEs, where partial information on the state (being infinite dimensional) and higher order derivatives are fundamental to the problem. [270] crucially point out that decoupling the level of discretization in their finite difference stencil from the sampling rate of the data is required for mitigating bias in estimating the ODE when the sampling rate of the data is low, and demonstrate this in a restricted setting on the Van der Pol system. We generally consider sampling rates over an order of magnitude lower than other works and flexibly handle partial observations and higher order ODEs.

3.1.4 Outline

We derive JSINDy in Section 3.2, describe the algorithmic implementation in Section 3.3, give extensive numerical results in Section 3.4, and discuss the ODE relaxation in Section 3.5.

3.1.5 Notation

We will use boldface for vectors, matrices, and vector-valued trajectories and use subscripts to denote indices, e.g., $\mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_d(t))$. We adopt semicolon notation for parametrized functions, such as $\mathbf{x}(t; \mathbf{z})$, and use a dot $(\cdot)$ in place of an argument in order to treat an object as a function. For a function k of two arguments, and vectors $\mathbf{t}, \boldsymbol{\tau}$, we will write $k(\mathbf{t}, \boldsymbol{\tau})$ to mean the matrix with entries $k(t_i, \tau_m)$. We will write ∂_j^r to denote r -th derivative with respect to the j -th argument.

3.2 Joint state estimation and system identification

In this section, we formulate our approach as an infinite dimensional optimization problem in the product of an RKHS with $\mathbb{R}^{d \times Q}$. We discretize this objective via a quadrature rule and develop a representer theorem which reduces the optimization to a finite dimensional subspace, and prove existence of solutions to both the discretized and continuous problems.

3.2.1 The abstract formulation

Our approach centers around the joint computation of the estimated state $\hat{\mathbf{x}}(t)$, henceforth called *state estimation*, as well as the right hand side $\hat{f}$, henceforth called *system identification*. We will require that the pair $\hat{\mathbf{x}}$ and $\hat{f}$ are *self consistent*, in the sense that $\hat{\mathbf{x}}$ should (approximately) satisfy the ODE defined by $\hat{f}$:

$$\frac{d^p}{dt^p} \hat{\mathbf{x}}(t) = \hat{f} \left(\hat{\mathbf{x}}(t), \frac{d}{dt} \hat{\mathbf{x}}(t), \dots, \frac{d^{p-1}}{dt^{p-1}} \hat{\mathbf{x}}(t) \right).$$

To this end, we impose self-consistency using a least squares penalty method [141, 281–285], which is discretized via collocation in an RKHS. Imposing that $\hat{f}$ has a sparse representation in the library $\Phi = \{\phi_1, \dots, \phi_Q\}$, we adopt a parametrized notation for how it operates on $\mathbf{x}$ and its derivatives:

$$f(D[\mathbf{x}](t); \boldsymbol{\theta}) := \left(\sum_{q=1}^Q \theta_{1,q} \phi_q(D[\mathbf{x}](t)), \dots, \sum_{q=1}^Q \theta_{d,q} \phi_q(D[\mathbf{x}](t)) \right), \quad \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}. \quad (3.6)$$

For instance, if we consider the first order constant coefficient linear ODE $\frac{d}{dt}\mathbf{x}(t) = \mathbf{A}\mathbf{x}$, then $D[\mathbf{x}](t) = \mathbf{x}$, and we would be looking for $\boldsymbol{\theta} = \mathbf{A}$.

We define our estimators $(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}})$ as solutions to the following optimization problem.

$$\begin{aligned}
& \underset{\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} && \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}) \\
& \text{subject to} && \text{supp}(\boldsymbol{\theta}) \subset \mathcal{S}(\mathbf{x}), \\
& && \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}) := \alpha \mathcal{L}_D(\mathbf{x}) + \beta \mathcal{L}_C(\mathbf{x}, \boldsymbol{\theta}) + \frac{\lambda}{2} \|\mathbf{x}\|_{\mathcal{X}}^2 + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2, \\
& && \mathcal{L}_D(\mathbf{x}) := \frac{1}{2} \sum_{n=1}^N \left| \mathbf{V}_n^\top \mathbf{x}(t_n) - \mathbf{y}_n \right|^2, \\
& && \mathcal{L}_C(\mathbf{x}, \boldsymbol{\theta}) := \frac{1}{2} \int_0^T \left| \frac{d^p}{dt^p} \mathbf{x}(t) - f(D[\mathbf{x}](t); \boldsymbol{\theta}) \right|^2 dt.
\end{aligned} \tag{3.7}$$

where

- $\alpha, \beta, \lambda, \mu > 0$ are relative weights trading off data fidelity, self consistency through the ODE, and regularization of the trajectory estimate and coefficients.
- $\mathcal{L}_D$ is a data misfit term enforcing that the estimated trajectory $\hat{\mathbf{x}}$ is consistent with the observed data.
- $\mathcal{L}_C$ is a least squares penalty enforcing that $\hat{\mathbf{x}}$ approximately satisfies the ODE induced by $\hat{f}$.
- $D[\mathbf{x}](t)$, defined in eq. (3.4) expands the trajectory $\mathbf{x}(t)$ into the states and derivatives.
- $\mathcal{X}$ is a vector-valued RKHS used to represent the trajectories, whose elements are smooth enough such that $D[\mathbf{x}](t)$ is well defined pointwise for all $t \in [0, T]$, and $\|\mathbf{x}\|_{\mathcal{X}}$ promotes regularity in the estimates of the trajectory. See the discussion below and Section 3.6.2 for more technical details.
- $\mathcal{S}$ is a feature selection procedure for choosing active terms from the feature library Φ given a fixed trajectory estimate, and maps to a subset of indices. This can be seen as applying a traditional SINDy procedure, with an arbitrary choice of sparsifier, and using exact derivatives of the approximation of the trajectory in order to choose the subset of features which are active.

Because we handle sparsity through the feature selection operator $\mathcal{S}$ we do not need to apply regularization to $\boldsymbol{\theta}$ to promote sparsity. In principle, however, the Frobenius norm $\frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2$ could be replaced by a more general sparsity promoting regularizer $R(\boldsymbol{\theta})$ which would supplant $\mathcal{S}$.

The inclusion of a feature selection procedure is motivated by a natural algorithmic idea for sparse nonlinear optimization. Consider the fixed point iteration

$$\begin{aligned} (\mathbf{x}^{(k+1)}, \boldsymbol{\theta}^{(k+1)}) &\in \begin{cases} \arg \min_{\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}} & \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}) \\ \text{subject to} & \text{supp}(\boldsymbol{\theta}) \subset S^{(k)} \end{cases} \\ S^{(k+1)} &= \mathcal{S}(\mathbf{x}^{(k+1)}), \end{aligned} \quad (3.8)$$

which alternates between optimizing on an active set of features, and choosing features to use in the active set. Given a solution $(\mathbf{x}^*, \boldsymbol{\theta}^*)$ to (3.7), define $S^* = \mathcal{S}(\mathbf{x}^*)$. If

$$(\mathbf{x}^*, \boldsymbol{\theta}^*) \in \arg \min_{\substack{\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q} \\ \text{supp}(\boldsymbol{\theta}) \subset S^*}} \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}), \quad (3.9)$$

then the pair $(\mathbf{x}^*, \boldsymbol{\theta}^*)$ is a fixed point of (3.8). This is a strong condition, but is likely to hold whenever $\mathcal{S}$ is a reasonably good sparsifier that is consistent with the collocation loss $\mathcal{L}_C$. This is also a desirable property in itself; we want for our estimate fixed to the active set to be the best estimate possible on that active set.

The ADO algorithms of [159, 160] are exactly of the form (3.8) and motivate it as a surrogate to ℓ_0 minimization. Similarly, the greedy evidence maximization in [269] is related in that it solves a sequence of fixed-support nonlinear least squares problems, though in each iteration, separate subproblems are solved for each candidate feature to remove. This has the benefit of minimizing a fixed objective function, and ensuring convergence due to the monotone feature removal process. This is discussed in detail algorithmically in Section 3.3.1 and Section 3.3.3.

In our theoretical analysis, we will remove the constraint and work with

$$\underset{\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} \quad \mathcal{L}(\mathbf{x}, \boldsymbol{\theta}), \quad (3.10)$$

which describes the first piece of the fixed point iteration and still entails solving a challenging nonlinear programming problem.

This approach combines regularization on the trajectory estimate with a least squares relaxation of the differential equation, which prevents trajectory estimates from diverging away from the observed data. Further, this relaxation allows the observed data itself to guide the optimization in the early iterations, pulling our state estimates towards a neighborhood of the true trajectory even before our estimates effectively approximate the solution to the ODE.

To represent the d -dimensional trajectory $\mathbf{x} : [0, T] \rightarrow \mathbb{R}^d$, we work in a vector-valued RKHS $\mathcal{X}$ of functions $[0, T] \mapsto \mathbb{R}^d$ generated by the d -fold Cartesian product of an RKHS $\mathcal{H}_k$, associated to a symmetric positive definite kernel k . This is equivalent to modeling each coordinate $x_i(t)$ of the trajectory $\mathbf{x}(t)$ separately in the RKHS $\mathcal{H}_k$, or by taking $\mathcal{X}$ to be the vector-valued RKHS with matrix valued kernel $K(t, t') = k(t, t')I_d$. We will require that for derivative orders $j = 0, \dots, p$, and times $t \in [0, T]$, the derivative evaluation map $\mathbf{x} \mapsto \frac{d^j}{dt^j} \mathbf{x}(t)$ is a bounded linear operator on $\mathcal{X}$ (see Section 3.6.2). It suffices to take k such

that $\mathcal{H}_k$ is compactly embedded in $C^p([0, T])$, which is satisfied if $k \in C^{p,p}([0, T] \times [0, T])$ [286].

We choose the kernel k to be the sum of a high-order Matérn kernel [287] and a constant kernel to account for systems where the time-average is nonzero. More precisely, with h_ν denoting the ν -order Matérn covariance function and ℓ as a length scale parameter, we use kernels of the form

$$k(t, t') = c_0 + c_1 h_\nu \left(\frac{|t - t'|}{\ell} \right), \quad h_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r \right)^\nu K_\nu \left(\sqrt{2\nu} r \right) \quad (3.11)$$

where K_ν is the modified Bessel function of the second kind, and $c_0, c_1 > 0$. These kernels induce RKHSs which satisfy the boundedness condition of the derivative evaluation maps whenever $\nu > p + \frac{1}{2}$. See Section 3.6.2 for more details on kernels and smoothness.

As a preliminary step in our model, we choose the kernel hyperparameters c_0, c_1, ℓ , and the variance of the observation noise $\sigma^2 = \mathbb{V}[\epsilon]$ using maximum likelihood and treat the distribution of ϵ as Gaussian, as is common in Gaussian process regression; see for example [287]. From there, $\|\mathbf{x}\|_{\mathcal{X}}^2$ corresponds to the regularization induced by the RKHS norm associated to this fitted Gaussian process.

3.2.2 Discretization

In order to obtain a computable objective function, we discretize the ODE residual $\mathcal{L}_C(\mathbf{x}, \boldsymbol{\theta})$ using a quadrature rule, setting

$$\mathcal{L}_C(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{2} \int_0^T \left| \frac{d^p}{dt^p} \mathbf{x}(\tau) - f(D[\mathbf{x}](\tau); \boldsymbol{\theta}) \right|^2 d\tau \quad (3.12)$$

$$\approx \frac{1}{2} \sum_{m=1}^M w_m \left| \frac{d^p}{dt^p} \mathbf{x}(\tau_m) - f(D[\mathbf{x}](\tau_m); \boldsymbol{\theta}) \right|^2 := \mathcal{L}_C^\tau(\mathbf{x}, \boldsymbol{\theta}) \quad (3.13)$$

where we refer to $\boldsymbol{\tau} = (\tau_m)_{m=1}^M$ as collocation points, and $\mathbf{w} = (w_m)_{m=1}^M$ as collocation weights. In all of our examples, we take uniformly spaced collocation nodes and uniform weights; this is discussed further in Section 3.5. Further, for notational convenience and without loss of generality, we assume that the observation times $\mathbf{t} = (t_n)_{n=1}^N$ are a subset of the collocation points $\boldsymbol{\tau}$, which can be enforced by augmenting $\boldsymbol{\tau}$, and taking some w_m to be zero.

Substituting $\mathcal{L}_C^\tau$ for $\mathcal{L}_C$ in eq. (3.10), we obtain the optimization problem

$$\begin{aligned} \underset{\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} \quad \mathcal{L}^\tau(\mathbf{x}, \boldsymbol{\theta}) &:= \frac{\alpha}{2} \sum_{n=1}^N \left| \mathbf{V}_n^\top \mathbf{x}(t_n) - \mathbf{y}_n \right|^2 \\ &+ \frac{\beta}{2} \sum_{m=1}^M w_m \left| \frac{d^p}{dt^p} \mathbf{x}(\tau_m) - f(D[\mathbf{x}](\tau_m); \boldsymbol{\theta}) \right|^2 \\ &+ \frac{\lambda}{2} \|\mathbf{x}\|_{\mathcal{X}}^2 + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2 \end{aligned} \quad (3.14)$$

which we use to compute our estimators. At this point, the reproducing property and boundedness of derivative evaluations in the RKHS $\mathcal{X}$ give a natural parametrization of $\hat{\mathbf{x}}$ which is justified by the following *representer theorem*, which converts optimization over the infinite-dimensional space $\mathcal{X}$ into a problem on a finite-dimensional subspace. This result is more general than the typical representer theorems often found in ML, such as in [287], which only make use of pointwise evaluations, but is very similar to the representer theorem given in [142], and follows from the more general results of [166]. The purpose of this result is to reduce the optimization problem over the infinite-dimensional space $\mathcal{X}$ into an optimization problem over the finite-dimensional subspace spanned by an explicit set of basis functions.

Theorem 3.1. *Consider the optimization problem (3.14) for $\alpha, \beta, \lambda, \mu > 0$. Let $\mathcal{X}$ be the vector-valued RKHS associated to the d -fold Cartesian product of an RKHS $\mathcal{H}_k$ associated to the positive definite kernel k . Assume that linear operators of the form $\mathbf{x} \mapsto \frac{d^r}{dt^r} \mathbf{x}(\tau)$ for $r = 0, \dots, p$ and $\tau \in [0, T]$ are bounded in the RKHS $\mathcal{X}$. Also, assume that all functions ϕ_q in the library Φ are continuous on $\mathbb{R}^{dp}$.*

1. *There exists a minimizer $(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}})$ to the optimization problem*

$$\underset{\mathbf{x} \in \mathcal{X}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} \mathcal{L}^\tau(\mathbf{x}, \boldsymbol{\theta}).$$

2. *Define the basis functions*

$$D_\tau[k](t) := (k(t, \tau_1), \partial_2 k(t, \tau_1), \dots, \partial_2^p k(t, \tau_1), \dots, k(t, \tau_M), \dots, \partial_2^p k(t, \tau_M)) \in \mathbb{R}^{(p+1)M}, \quad (3.15)$$

and for $\mathbf{z} \in \mathbb{R}^{(p+1)M \times d}$, set

$$\mathbf{x}(t; \mathbf{z}) := D_\tau[k](t)^\top \mathbf{z}. \quad (3.16)$$

Then problem (3.14) is equivalent to

$$\underset{\mathbf{z} \in \mathbb{R}^{(p+1)M \times d}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} \mathcal{L}^\tau(\mathbf{x}(\cdot; \mathbf{z}), \boldsymbol{\theta}), \quad (3.17)$$

in the sense that they have the same optimal value and $(\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}})$ is a minimizer if and only if $\hat{\mathbf{x}}$ admits a representation $\hat{\mathbf{x}}(t) = D_\tau[k](t)^\top \hat{\mathbf{z}}$ for some $\hat{\mathbf{z}}$, where $(\hat{\mathbf{z}}, \hat{\boldsymbol{\theta}})$ is a solution to (3.17).

3. *For all $\mathbf{z} \in \mathbb{R}^{(p+1)M \times d}$,*

$$\|\mathbf{x}(\cdot; \mathbf{z})\|_{\mathcal{X}}^2 = \text{tr}(\mathbf{z}^\top \mathbf{K} \mathbf{z}), \quad \mathbf{K} = \begin{bmatrix} k(\boldsymbol{\tau}, \boldsymbol{\tau}) & \partial_2^1 k(\boldsymbol{\tau}, \boldsymbol{\tau}) & \cdots & \partial_2^p k(\boldsymbol{\tau}, \boldsymbol{\tau}) \\ \partial_1^1 k(\boldsymbol{\tau}, \boldsymbol{\tau}) & \partial_1^1 \partial_2^1 k(\boldsymbol{\tau}, \boldsymbol{\tau}) & \cdots & \partial_1^1 \partial_2^p k(\boldsymbol{\tau}, \boldsymbol{\tau}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1^p k(\boldsymbol{\tau}, \boldsymbol{\tau}) & \partial_1^p \partial_2^1 k(\boldsymbol{\tau}, \boldsymbol{\tau}) & \cdots & \partial_1^p \partial_2^p k(\boldsymbol{\tau}, \boldsymbol{\tau}) \end{bmatrix} \in \mathbb{R}^{(p+1)M \times (p+1)M}. \quad (3.18)$$

Proof. We have that

$$\mathcal{L}^\tau(\mathbf{x}, \boldsymbol{\theta}) = \alpha \mathcal{L}_D(\mathbf{x}) + \beta \mathcal{L}_C^\tau(\mathbf{x}, \boldsymbol{\theta}) + \frac{\lambda}{2} \|\mathbf{x}\|_{\mathcal{X}}^2 + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2. \quad (3.19)$$

By nesting, problem (3.14) is equivalent to

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} \inf_{\mathbf{x} \in \mathcal{X}} \mathcal{L}^\tau(\mathbf{x}, \boldsymbol{\theta}) \quad (3.20)$$

and we can thus focus solely on the dependence with respect to $\mathbf{x}$, studying the inner variational problem, as the outer problem is already finite dimensional. Going term by term, we have

$$\mathcal{L}_D(\mathbf{x}) = \frac{1}{2} \sum_{n=1}^N \left| \mathbf{V}_n^\top \mathbf{x}(t_n) - \mathbf{y}_n \right|^2, \quad (3.21)$$

which depends only on point evaluations of the form $\mathbf{x}(\tau_m)$ (recall that we assume the observation times $\mathbf{t}$ are a subset of τ). Next, we have that

$$\mathcal{L}_C^\tau(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{2} \sum_{m=1}^M w_m \left| \frac{d^p}{dt^p} \mathbf{x}(\tau_m) - f(D[\mathbf{x}](\tau_m); \boldsymbol{\theta}) \right|^2, \quad (3.22)$$

depends on $\mathbf{x}$ only through point evaluations of $\mathbf{x}$ and its derivatives at $(\tau_m)_{m=1}^M$, i.e., the values of $D[\mathbf{x}](\tau_m)$ and $\frac{d^p}{dt^p} \mathbf{x}(\tau_m)$.

Recall that linear operators of the form $\mathbf{x} \mapsto \frac{d^r}{dt^r} \mathbf{x}(\tau)$ for $r = 0, \dots, p$ and $\tau \in [0, T]$ are assumed to be bounded on the RKHS $\mathcal{X}$. Thus, by the reproducing property [286],

$$\frac{d^r}{dt^r} x_i(\tau_m) = \langle \partial_2^r k(\cdot, \tau_m), x_i(\cdot) \rangle_{\mathcal{H}_k}. \quad (3.23)$$

Since the objective $\mathcal{L}^\tau$ depends only on the values of $\frac{d^r}{dt^r} \mathbf{x}(\tau_m)$ for $r = 0, 1, \dots, p$, $m = 1, \dots, M$, with the addition of the squared RKHS norm, this implies that

$$\hat{x}_i \in \text{span} \left\{ \partial_2^r k(\cdot, \tau_m) : m = 1, \dots, M, r = 0, \dots, p \right\}, \quad (3.24)$$

by Theorem 3.3. This is precisely the subspace spanned by the basis $D_\tau[k](\cdot)$ in (3.15). By the reproducing property, we also have that

$$\langle \partial_2^r k(\cdot, \tau_m), \partial_2^s k(\cdot, \tau_{m'}) \rangle_{\mathcal{H}_k} = \partial_1^r \partial_2^s k(\tau_m, \tau_{m'}), \quad (3.25)$$

which gives the structure of the matrix $\mathbf{K}$, and the RKHS norm of an individual component of the trajectory $x_i(\cdot) \in \text{span} \{ \partial_2^r k(\cdot, \tau_m), r = 0, 1, \dots, p, m = 1, \dots, M \}$. Stacking the individual components together and noting that

$$\mathbf{x}(t) = (x_1(t), \dots, x_d(t)) \text{ and } \|\mathbf{x}\|_{\mathcal{X}}^2 = \sum_{i=1}^d \|x_i\|_{\mathcal{H}_k}^2 \quad (3.26)$$

produces the compact representation in (3.16) and the formula for the RKHS norm in (3.18).

We obtain existence of minimizers from continuity and coercivity. The objective is continuous with respect to $\mathbf{z}$ and $\boldsymbol{\theta}$ due to the continuity of the feature maps ϕ_q . If $\mathbf{K}$ is invertible, then the objective is coercive and thus attains a minimum. If it is not invertible, then we may, without affecting $\mathcal{L}_D$ or $\mathcal{L}_C^\tau$, restrict $\mathbf{z}$ to the range of $\mathbf{K}$; this is essentially applying a finite-dimensional representer theorem to the subspace spanned by $D_\tau[k]$. Under this additional constraint, the objective becomes coercive, giving existence of minima. Finally, the equivalence of the optimization problems allows us to transfer existence from problem 3.17 to problem 3.14. $\square$

A direct application of Theorem 3.1 gives the finite-dimensional objective function

$$\begin{aligned} \mathcal{J}(\mathbf{z}, \boldsymbol{\theta}) := \mathcal{L}^\tau(\mathbf{x}(\cdot; \mathbf{z}), \boldsymbol{\theta}) &= \frac{\alpha}{2} \sum_{n=1}^N \left| \mathbf{V}_n^\top \mathbf{x}(t_n; \mathbf{z}) - \mathbf{y}_n \right|^2 \\ &+ \frac{\beta}{2} \sum_{m=1}^M w_m \left| \frac{d^p}{dt^p} \mathbf{x}(\tau_m; \mathbf{z}) - f(D[\mathbf{x}](\tau_m; \mathbf{z})) \right|^2 \\ &+ \frac{\lambda}{2} \|\mathbf{z}\|_{\mathbf{K}}^2 + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2, \quad \|\mathbf{z}\|_{\mathbf{K}}^2 := \text{tr}(\mathbf{z}^T \mathbf{K} \mathbf{z}) \end{aligned} \quad (3.27)$$

We also give a theorem on the existence of solutions to the continuous problem eq. (3.10).

Theorem 3.2. *Consider the optimization problem (3.10). Let $\alpha, \beta, \lambda, \mu > 0$. Let $\mathcal{X}$ be the vector-valued RKHS associated to the d -fold Cartesian product of an RKHS $\mathcal{H}_k$ associated to the positive definite kernel k . Assume that $\mathcal{H}_k$ is compactly embedded in $C^p([0, T], \mathbb{R}^d)$. Assume that all functions ϕ_q in the library Φ are continuous on $\mathbb{R}^{dp}$. Then there exists a minimizer $(\mathbf{x}^*, \boldsymbol{\theta}^*) \in \mathcal{X} \times \mathbb{R}^{d \times Q}$.*

Proof. Because the features ϕ_q are assumed to be continuous functions, and point evaluations of derivatives up to order p are bounded on $\mathcal{H}_k$ due to the compact embedding property, $\mathcal{L}_C$ and $\mathcal{L}_D$ are continuous on both $\mathcal{X} \times \mathbb{R}^{d \times Q}$ and $C^p([0, T], \mathbb{R}^d) \times \mathbb{R}^{d \times Q}$. Additionally, both are lower bounded by zero. Since $\frac{\lambda}{2} \|\mathbf{x}\|_{\mathcal{X}}^2 + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2$ is coercive on $\mathcal{X} \times \mathbb{R}^{d \times Q}$, we obtain that $\mathcal{L}$ is coercive as well. For a minimizing sequence $(\mathbf{x}^{(k)}, \boldsymbol{\theta}^{(k)})$, we may extract a subsequence that strongly converges in $C^p([0, T], \mathbb{R}^d) \times \mathbb{R}^{d \times Q}$ to a candidate minimizer $(\mathbf{x}^*, \boldsymbol{\theta}^*)$. Lower semicontinuity of $\mathcal{L}$ yields that $\lim_{k \rightarrow \infty} \mathcal{L}(\mathbf{x}^{(k)}, \boldsymbol{\theta}^{(k)}) = \mathcal{L}(\mathbf{x}^*, \boldsymbol{\theta}^*)$ giving existence of a minimizer. $\square$

3.3 Algorithms and implementation details

Reintroducing the sparse selection procedure, we obtain the optimization problem

$$\begin{aligned} &\underset{\mathbf{z} \in \mathbb{R}^{(p+1)M \times d}, \boldsymbol{\theta} \in \mathbb{R}^{d \times Q}}{\text{minimize}} && \mathcal{J}(\mathbf{z}, \boldsymbol{\theta}) \\ &\text{subject to} && \text{supp}(\boldsymbol{\theta}) \subset \mathcal{S}(\mathbf{x}). \end{aligned} \quad (3.28)$$

In order to numerically compute solutions to eq. (3.28), we alternate between applying a Levenberg-Marquardt (LM) algorithm to solve smooth nonlinear least squares problems

with the support of $\boldsymbol{\theta}$ restricted, and a sparse regression algorithm to choose the support of $\boldsymbol{\theta}$. We first write $\mathcal{J}$ as a regularized nonlinear least squares objective. Define the residual functions

$$\mathbf{F}_D(\mathbf{z}) := \begin{bmatrix} \mathbf{V}_1^\top \mathbf{x}(t_1; \mathbf{z}) - \mathbf{y}_1 \\ \vdots \\ \mathbf{V}_N^\top \mathbf{x}(t_N; \mathbf{z}) - \mathbf{y}_N \end{bmatrix}, \quad \mathbf{F}_C(\mathbf{z}, \boldsymbol{\theta}) = \begin{bmatrix} \sqrt{w_1} \left(\frac{dp}{dt^p} \mathbf{x}(\tau_1; \mathbf{z}) - f(D[\mathbf{x}](\tau_1; \mathbf{z}); \boldsymbol{\theta}) \right) \\ \vdots \\ \sqrt{w_M} \left(\frac{dp}{dt^p} \mathbf{x}(\tau_M; \mathbf{z}) - f(D[\mathbf{x}](\tau_M; \mathbf{z}); \boldsymbol{\theta}) \right) \end{bmatrix} \quad (3.29)$$

With

$$\mathbf{F}(\mathbf{z}, \boldsymbol{\theta}) = \begin{bmatrix} \sqrt{\alpha} \mathbf{F}_D(\mathbf{z}) \\ \sqrt{\beta} \mathbf{F}_C(\mathbf{z}, \boldsymbol{\theta}) \end{bmatrix}, \quad (3.30)$$

observe that

$$\mathcal{J}(\mathbf{z}, \boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{F}(\mathbf{z}, \boldsymbol{\theta})\|^2 + \frac{\lambda}{2} \text{tr}(\mathbf{z}^\top \mathbf{K} \mathbf{z}) + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2. \quad (3.31)$$

The addition of the quadratic regularization term $\frac{1}{2} \mathbf{z}^\top \mathbf{K} \mathbf{z}$ is completely benign in the nonlinear least squares setting, and may be treated explicitly in Gauss-Newton Hessian approximations. We further perform a change of variables using the Cholesky factors of $\mathbf{K}$ in order to transform the quadratic regularization into a multiple of the identity, see Section 3.6.3. Additionally, note that $\mathbf{F}(\mathbf{z}, \boldsymbol{\theta})$ is affine with respect to $\boldsymbol{\theta}$, which makes it natural to apply methods for sparse linear least squares for the feature selection procedure $\mathcal{S}$; see Section 3.3.3.

3.3.1 Optimization

We compute minimizers of (3.17) by alternating between *active-set iterations*, and *sparsifying iterations*.

1. **Active-set iterations** refine the estimates of $\mathbf{z}, \boldsymbol{\theta}$ under the condition that $\text{supp}(\boldsymbol{\theta}) \subset S^{(k)}$ for a fixed active set $S^{(k)}$:

$$(\mathbf{z}^{(k+1)}, \boldsymbol{\theta}^{(k+1)}) = \arg \min_{\mathbf{z}, \boldsymbol{\theta}} \mathcal{J}(\mathbf{z}, \boldsymbol{\theta}), \text{ subject to } \text{supp}(\boldsymbol{\theta}) \subset S^{(k)} \quad (3.32)$$

These iterations drop the features not selected in step (2) and solve smooth nonlinear least-squares problems restricted to the identified support using LM. Taking previous parameter values as a warm start greatly improves efficiency. More algorithmic details are discussed in Section 3.3.2.

2. **Sparsifying iterations** freeze $\mathbf{z} = \mathbf{z}^{(k)}$ and compute

$$S^{(k+1)} = \mathcal{S}(\mathbf{x}(\cdot; \mathbf{z}^{(k)})) \quad (3.33)$$

in order to estimate a new active set. Once the estimate $\mathbf{x}(\cdot, \mathbf{z})$ of the trajectory is fixed, a natural strategy is to target the sparse regression problem of finding sparse $\boldsymbol{\theta}$ which approximately minimizes the collocation error $\mathcal{L}_C^\tau(\mathbf{x}(\cdot; \mathbf{z}^{(k)}), \boldsymbol{\theta})$. This objective then has a *linear* least squares structure, and $S^{(k+1)}$ is taken to be the support

of the identified solution. This approach can be seen as applying a traditional 2-step SINDy approach to the current trajectory estimate $\mathbf{x}(\tau_m; \mathbf{z}^{(k)})$, using the *exact* derivatives of the *estimated* trajectory.

We initialize with a dense support by setting $S^{(0)}$ to be all available indices of $\boldsymbol{\theta}$, and alternate these two steps until the support $S^{(k)}$ stabilizes, i.e. $S^{(k)} = S^{(k-1)}$, which corresponds to a fixed point. In most of our examples, we take $\mathcal{S}$ to be the result of sequentially thresholded ridge regression [265], which is also the choice taken in [159, 160] and makes our algorithm template match their ADO approach. Options for sparsifiers are discussed further in Section 3.3.3.

3.3.2 Smooth nonlinear least squares

To solve the smooth subproblems arising in the active-set refinement eq. (3.32), we use an LM method, which is a standard and well studied algorithm for nonlinear least squares [258, 288–290]. For notational convenience, stack the variables $\mathbf{z}, \boldsymbol{\theta}$ and restrict to the current support:

$$\boldsymbol{\eta} = \begin{bmatrix} \text{vec}(\mathbf{z}) \\ \text{vec}(\boldsymbol{\theta}_{S^{(k)}}) \end{bmatrix}, \quad \mathbf{G}(\boldsymbol{\eta}) = \mathbf{F}(\mathbf{z}, \boldsymbol{\theta}). \quad (3.34)$$

The subproblem in eq. (3.32) takes the form

$$\min_{\boldsymbol{\eta}} \mathcal{J}(\boldsymbol{\eta}) := \frac{1}{2} \|\mathbf{G}(\boldsymbol{\eta})\|^2 + \frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q} \boldsymbol{\eta}, \quad (3.35)$$

where we define the positive semi-definite matrix

$$\mathbf{Q} := \begin{bmatrix} \lambda (I_d \otimes \mathbf{K}) & 0 \\ 0 & \mu I_{|S^{(k)}|} \end{bmatrix}, \quad \text{so that } \frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q} \boldsymbol{\eta} = \frac{\lambda}{2} \text{tr}(\mathbf{z}^\top \mathbf{K} \mathbf{z}) + \frac{\mu}{2} \|\boldsymbol{\theta}\|_F^2.$$

Note that we have dropped rows/columns for inactive coefficients, i.e., we only keep $\boldsymbol{\theta}_{S^{(k)}}$.

Let $\mathbf{M}$ be a positive definite matrix defining a step metric. In our case, we set $\mathbf{M} = \mathbf{Q}$. LM linearizes inside of the least-squares term and adds a quadratic damping $\gamma^{(k)}$ to control step size and force convergence. The iteration is given by

$$\boldsymbol{\eta}^{(k+1)} = \arg \min_{\boldsymbol{\eta}} m^{(k)}(\boldsymbol{\eta}) + \frac{\gamma^{(k)}}{2} (\boldsymbol{\eta} - \boldsymbol{\eta}^{(k)})^\top \mathbf{M} (\boldsymbol{\eta} - \boldsymbol{\eta}^{(k)}), \quad (3.36)$$

$$m^{(k)}(\boldsymbol{\eta}) := \frac{1}{2} \left\| \nabla \mathbf{G}(\boldsymbol{\eta}^{(k)}) (\boldsymbol{\eta} - \boldsymbol{\eta}^{(k)}) + \mathbf{G}(\boldsymbol{\eta}^{(k)}) \right\|^2 + \frac{1}{2} \boldsymbol{\eta}^\top \mathbf{Q} \boldsymbol{\eta}. \quad (3.37)$$

We update $\gamma^{(k)}$ adaptively. Define the *gain ratio* (observed vs. predicted decrease)

$$\rho^{(k)} = \frac{\mathcal{J}(\boldsymbol{\eta}^{(k)}) - \mathcal{J}(\boldsymbol{\eta}^{(k+1)})}{\mathcal{J}(\boldsymbol{\eta}^{(k)}) - m^{(k)}(\boldsymbol{\eta}^{(k+1)})}. \quad (3.38)$$

If $\rho^{(k)} < 0.01$, we reject the step and increase $\gamma^{(k)}$; otherwise we accept and update $\gamma^{(k)}$ for the next iteration. One practical choice, analyzed in [258], is

$$\gamma^{(k+1)} = \begin{cases} \gamma^{(k)} c & \rho^{(k)} < 0.4, \\ \gamma^{(k)} & \rho^{(k)} \in [0.4, 0.9], \\ \gamma^{(k)} / c & \rho^{(k)} > 0.9, \end{cases} \quad (3.39)$$

for some $c > 1$; we use $c = 1.2$.

3.3.3 Sparsifying iterations

As mentioned in Section 3.3.1, we alternate between solving the fixed support, smooth non-linear least squares problem (3.32) and updating the support of the coefficient set to obtain S^{k+1} , applying a sparse feature selection method in eq. (3.33). Hence, a crucial component of our algorithm is the choice of regularizer or feature selection procedure in eq. (3.33). The work that introduced SINDy [265] suggested ridge regularized sequentially thresholded least squares (STLSQ). This alternates between solving a ridge regularized least squares problem, and evicting features whose coefficients fall below a given threshold. Another approach is to add regularization which promotes sparsity. These include LASSO [291] which regularizes with an ℓ_1 penalty, SR3 [292, 293] which minimizes smoothed versions of nonsmooth penalties, and MIO SR [294], which directly solves ℓ_0 constrained problems using mixed integer programming. Ensembling the approaches above via bagging and feature subsampling can sometimes improve the performance of such algorithms [151], and provide some uncertainty quantification [295].

An alternative approach to feature selection is to define an appropriate sparsity promoting prior distribution and likelihood, and approximately sample from the posterior. These include spike-and-slab priors [296] and the regularized horseshoe prior [155].

In most of our experiments, we apply ridge regularized STLSQ as it empirically performed the best, and was easy to tune. In section 3.4.5 we applied SR3 [292] with an L1 penalty—we found this to be more consistent on varying trajectory lengths and noise levels without any retuning, but it can be difficult to intuit good parameters a priori. In certain cases, including ensembling within our sparsification procedure greatly improved the robustness of our method, see section 3.4.5. At the cost of losing a straightforward variational problem, considering an algorithmic sparsification framework allows the specific choice of sparsifier to become an implementation detail that is easy for users to modify and provides substantial flexibility.

3.4 Numerical results

We validate JSINDy on a range of examples that illustrate its capability in handling noisy, scarce and partial observations, higher order systems, and model misspecification.¹

¹An implementation of JSINDy and the code used to generate the experiments in this section are available at <https://github.com/AHsu98/jsindy>.

For each example, we generate a true trajectory $\mathbf{x}(t)$ solving the ODE $\frac{d^p}{dt^p}\mathbf{x}(t) = f(D[\mathbf{x}], \boldsymbol{\theta})$, and take noisy measurements $\mathbf{y}_n = \mathbf{V}_n^\top \mathbf{x}(t_n) + \boldsymbol{\epsilon}_n$ by adding Gaussian noise $\boldsymbol{\epsilon}_n$. Fitting our model results in state estimates $\hat{\mathbf{x}}(t)$, and system coefficients $\hat{\boldsymbol{\theta}}$. We evaluate the state estimates, $\hat{\mathbf{x}}(t)$ against the true trajectory $\mathbf{x}(t)$ using root mean square error (RMSE) averaged across coordinates, and normalized by the variance across time of each coordinate.

$$\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x})^2 := \frac{1}{T \cdot d} \sum_{m=1}^d \frac{\int_0^T (\hat{x}_m(t) - x_m(t))^2 dt}{\int_0^T (\bar{x}_m - x_m(t))^2 dt}, \quad \bar{x}_m := \frac{1}{T} \int_0^T x_m(s) ds \quad (3.40)$$

Where appropriate, we evaluate the fitted coefficient matrix $\hat{\boldsymbol{\theta}}$ using the normalized Frobenius norm compared to the true coefficients $\boldsymbol{\theta}$.

$$\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) := \frac{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_F}{\|\boldsymbol{\theta}\|_F} \quad (3.41)$$

This is not relevant in Section 3.4.2 due to unidentifiability, and in Section 3.4.4, as there is no “true” coefficient matrix. All models, for each experiment, use a sum of a scaled constant kernel and Matérn kernel with $\nu = 11/2$, which corresponds to five times differentiable functions.

We test our algorithm on a variety of systems and observation patterns. We start with the Lorenz 63 system, which is commonly used for testing methods for identifying nonlinear dynamics, using a relatively standard sampling frequency and noise level. Next, we consider the regime of scarce sampling, with sampling frequencies far below what have been previously considered in the literature, working with the Lorenz and Lotka–Volterra dynamics. We then consider problems with partial observations, where we never observe the entire state at once (even with noise), with cases of alternating between observing the different coordinates, and observing only a subset. Finally, we consider the related problem of identifying higher order ODEs from only observations of the state variable, but not its derivatives, working with the Van der Pol system, a higher dimensional coupled version of the Duffing system, and the nonlinear pendulum under model misspecification. The problem of identifying higher order ODEs from point observations can be seen as a version of a partially observed problem with additional structure between dimensions if reduced to first order.

In all of these examples, we find that the assumption that $\mathbf{x}(t)$ satisfies *some* autonomous ODE is a strong prior that greatly improves inference, even when the ODE must itself be identified. Identification of the ODE itself is further facilitated by the assumption of sparsity in the dynamics which significantly restricts the class of models under consideration. The benefits of a joint approach to learning as opposed to two-step estimation have been observed in many works [2, 159, 160, 269, 270, 297]. The problem of data scarcity was considered in [2, 160, 270]. The SIDDS algorithm of [270] found that such an approach can reach the Cramer-Rao lower bound and demonstrated the necessity of decoupling the discretization of the ODE from the grid of observations in order to reduce bias when the sampling period is large. Our work expands the flexibility to even scarcer data and partial observations.

We largely use the same hyperparameters throughout. We set the trajectory regularization $\lambda = 10^{-3}$, the data weight $\alpha = 1$, the collocation penalty $\beta = 10^5$, and use 500 uniformly spaced collocation points with uniform weight. Except for Section 3.4.5 where we apply SR3, we use STLSQ as a sparsifier, with thresholds and ridge regularization parameters chosen for each problem. See Table 3.1 for full details of hyperparameters and problem parameters.

Additionally, we include two systematic experiments: one varying the time horizon and noise level on the Lorenz system, and another varying the noise level on a nonlinear damped oscillator system. These show promising empirical results compared to other competitive models.

Experiment	Noise abs	Noise rel (%)	Order	Δt	t_1	Obs.	Obs. Pattern	Thr	Ridge	RMSE _{Var} ($\hat{x}, x$)	RMSE _{Fro} ($\hat{\theta}, \theta$)
Lorenz	2.0	24.60	1	0.05	10	200 ($\times 3$)	Full	0.5	0.01	2.35e-2	3.97e-2
Scarce Lorenz	0.0	0.00	1	0.5	10	21 ($\times 3$)	Full	0.5	0.01	1.61e-3	2.36e-3
Lotka–Volterra	0.2	19.68	1	3	50	17 ($\times 2$)	Full	0.1	0.1	1.13e-2	3.32e-2
Lorenz Streams	0.32	6.92	1	0.025	10	404	One at a time	0.25	0.01	5.69e-3	1.96e-2
Lorenz Unobserved	0.32	3.98	1	0.025	10	404	3rd coord. unobs.	0.1	0.01	6.84e-3	N/A
Van der Pol	0.1	7.39	2	2	100	50	Pos.-only	0.1	0.01	4.54e-2	1.86e-1
Coupled Duffing	0.2	16.85	2	0.2	30	150 ($\times 5$)	Pos.-only	0.1	0.01	3.55e-2	3.63e-2
Nonlin. pend. (linear)	0.1	4.17	2	1	100	100	Pos.-only	0.1	0.01	1.25e-01	N/A
Nonlin. pend. (cubic)	0.1	4.17	2	1	100	100	Pos.-only	0.1	0.01	2.22e-2	N/A
Damped nonlin. osc.	0.32	38.22	1	0.05	10	200 ($\times 2$)	Full	0.08	100.00	6.60e-2	2.47e-1

Table 3.1: Summary of experiments. The $n(\times d)$ indicates that the observations were made at n time points of the d dimensional state. **Thr.** and **Ridge** are the threshold and ridge regularization strength used in the STLSQ sparsifier.

Lorenz System. As a first example which is commonly used to test methods for identifying nonlinear dynamics, we considered the Lorenz system:

$$\begin{aligned}
 \frac{dx_1}{dt} &= -10x_1 + 10x_2, \\
 \frac{dx_2}{dt} &= 28x_1 - x_2 - x_1x_3, \\
 \frac{dx_3}{dt} &= -\frac{8}{3}x_3 + x_1x_2.
 \end{aligned} \tag{3.42}$$

We use the initial condition $\mathbf{x}(0) = (-8, 8, 27)$ and took full observations with a sampling period $\Delta t = 0.05$, and Gaussian noise with variance $\sigma^2 = 4$.

Our approach obtained the equations

$$\begin{aligned}
 \frac{dx_1}{dt} &= -9.607x_1 + 9.689x_2 \\
 \frac{dx_2}{dt} &= 29.122x_1 - 1.239x_2 - 1.033x_1x_3 \\
 \frac{dx_3}{dt} &= -2.624x_3 + 0.999x_1x_2.
 \end{aligned} \tag{3.43}$$

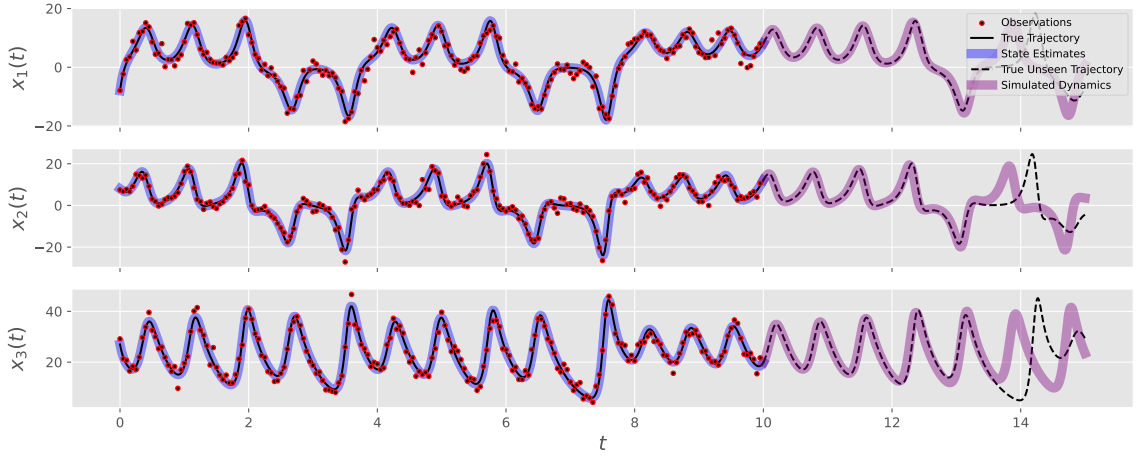


Figure 3.1: Results for Lorenz 63 system. Observations sampled at $\Delta t = 0.05$ until $t = 10$ with added Gaussian noise of variance $\sigma^2 = 4$. Simulated dynamics from eq. (3.43) are compared against true dynamics from $t = 10$ until $t = 15$.

with $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 2.353 \cdot 10^{-2}$ and $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 3.975 \cdot 10^{-2}$. Results can be seen in Figure 3.1. This is essentially a baseline example. For this problem, the sampling rate was at the lower end of the standard range that is considered in the literature. Our approach both identified the correct active features, and approximated the coefficient values to reasonable accuracy. In the rest of our numerical examples, we largely consider examples where traditional two-step methods are not applicable, or will obviously break down.

3.4.1 Scarce observations

We now consider cases of very scarcely sampled data with the Lorenz 63 and the Lotka–Volterra systems.

Lorenz 63 Scarce

We modified the previous example, instead sampling 21 observations of the entire state over the interval $[0, 10]$ with a sampling period of $\Delta t = 0.5$, and no noise. Results can be seen in Figure 3.2 with the following learned system below:

$$\begin{aligned} \frac{dx_1}{dt} &= -10.008 x_1 + 10.006 x_2 \\ \frac{dx_2}{dt} &= 28.073 x_1 - 1.012 x_2 - 1.003 x_1 x_3 \\ \frac{dx_3}{dt} &= -2.663 x_3 + 1.001 x_1 x_2 \end{aligned} \tag{3.44}$$

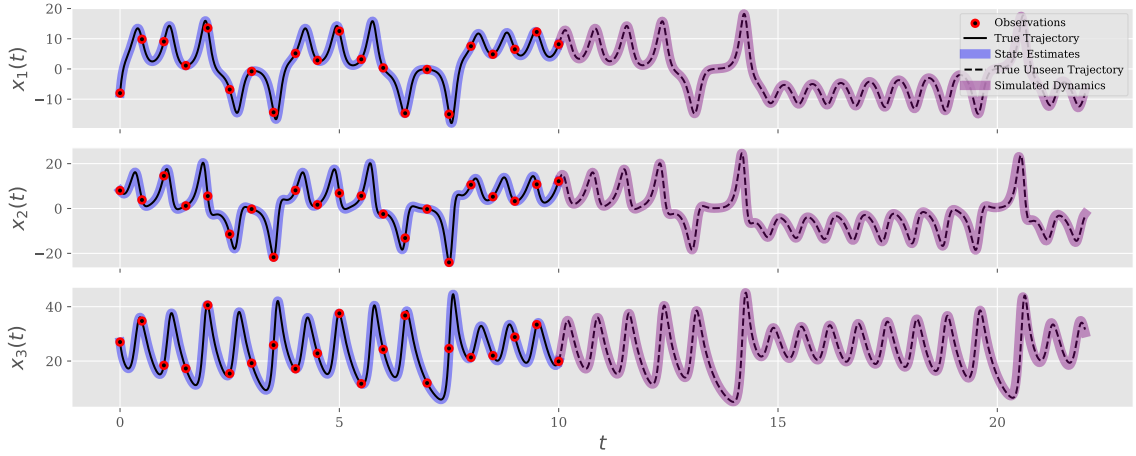


Figure 3.2: Lorenz 63 system with observations sampled at $\Delta t = 0.5$ until $t = 10$ with no added noise. Simulated dynamics of (3.44) are plotted against true, unseen, trajectory from $t = 10$ to $t = 25$.

with $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 1.608 \cdot 10^{-3}$ and $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 2.356 \cdot 10^{-3}$. These results are very accurate, owing to the lack of noise. Indeed, a parameter counting argument suggests that once the support is identified, this problem is essentially a noiseless and well-posed though ill-conditioned inversion problem, where we should expect good algorithms to attain very high accuracy.

Lotka–Volterra

The Lotka–Volterra system is a nonlinear oscillator with periodic solutions, describing population dynamics of a predator-prey ecology, and is a common benchmark for system identification [151, 277, 294, 297]. This system governs the populations of prey x_1 and predators x_2 :

$$\begin{aligned} \frac{dx_1}{dt} &= x_1 - x_1 x_2, \\ \frac{dx_2}{dt} &= -x_2 + x_1 x_2, \end{aligned} \tag{3.45}$$

We again considered scarce sampling, with approximately 2 observations per cycle, taken at a sampling period of $\Delta t = 3$ in the interval $t \in [0, 50]$. Gaussian noise with variance $\sigma^2 = 0.04$ was added to the observations. We set the feature library to include monomials up to second order:

$$\Phi = \{1, x_1, x_2, x_1^2, x_1 x_2, x_2^2\}.$$

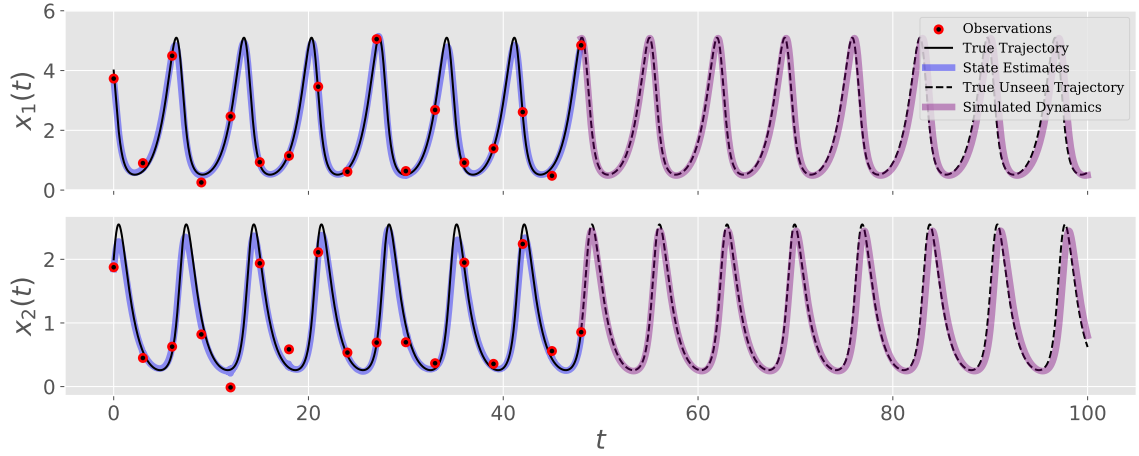


Figure 3.3: Results for the Lotka–Volterra system of section 3.4.1. Samples are taken at the rate $\Delta t = 3$ up to $t = 48$, including additive noise with $\sigma = 0.2$. Simulated dynamics are shown up to $t = 100$, comparing the results from (3.46) with the true dynamics.

Our approach obtained the system

$$\begin{aligned} \frac{dx_1}{dt} &= 1.009x_1 - 1.056x_1x_2 \\ \frac{dx_2}{dt} &= -0.980x_2 + 0.500x_1x_2. \end{aligned} \tag{3.46}$$

with $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 1.131 \cdot 10^{-1}$ and $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 3.323 \cdot 10^{-2}$. We also compare the phase portrait of the learned model to the true dynamics, simulating multiple trajectories from various initial conditions in Figure 3.4.

3.4.2 Partial observations

We now consider problems with only partial observations. Rather than observing the entire state $\mathbf{x}(t_n)$, we observe a sequence of linear measurements of the state, $\mathbf{y}_n = \mathbf{V}_n^\top \mathbf{x}(t_n)$. We test our algorithm on the Lorenz dynamics with two challenging observation patterns. In these examples, we took each $\mathbf{V}_n$ to be a standard basis vector, observing one coordinate at a time. In both of these examples, we took the same initial conditions as above, set the sampling period $\Delta t = 0.025$, and set the noise $\sigma^2 = 0.1$.

Alternating coordinates

In the first case, we alternated between periods of observing each of the three coordinates one at a time for 10 time steps each. For each coordinate, this leads to relatively long gaps (of duration $t = 0.5$) where no observations of the state are made. However, our approach allows the dynamics to effectively fill in this missing information, correlating the

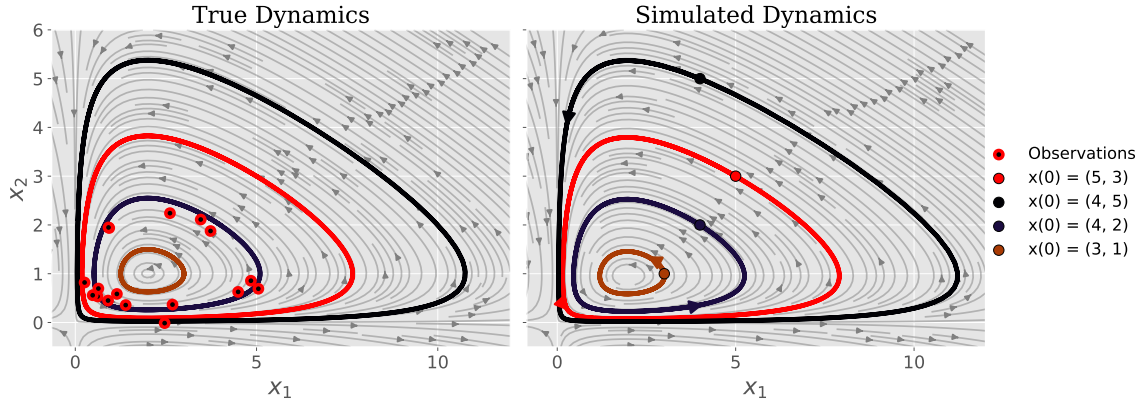


Figure 3.4: Left: True Lotka–Volterra plotted from multiple initial conditions on phase portrait using (3.45). Red trajectory and measurements are those from Figure 3.3. Right: Using learned model (3.46) to simulate same trajectories provided same initial conditions.

data available between the different coordinates. Notably, if the dynamics were *known ahead of time*, then applying a high order nonlinear Kalman smoothing method could be similarly effective, but this is not applicable here, since there is no a priori knowledge of the governing equations [298, 299]. Our model learned the system in (3.47) with $\text{RMSE}_{\text{Fro}}(\hat{\theta}, \theta) = 5.692 \cdot 10^{-3}$ and $\text{RMSE}_{\text{Var}}(\hat{x}, x) = 1.961 \cdot 10^{-2}$. Results can be seen in Figure 3.5.

$$\begin{aligned} \frac{dx_1}{dt} &= -9.987 x_1 + 9.951 x_2, \\ \frac{dx_2}{dt} &= 28.104 x_1 - 1.030 x_2 - 1.003 x_1 x_3, \\ \frac{dx_3}{dt} &= -0.606 \cdot 1 - 2.642 x_3 + 1.005 x_1 x_2. \end{aligned} \tag{3.47}$$

Hidden coordinates

We now consider the problem of identifying dynamics with a hidden variable that is never observed. In this case, the observations alternate between the first two coordinates on each time step, and the third is never observed. [279] considered a very similar setup, introduced a variational annealing approach to identify dynamics in the presence of hidden variables, and also tested on variants of the Lorenz dynamics. [174] approached the hidden variable problem using the lens of computational graph completion [300] and kernel learning, though only considered problems in discrete time without noise. In this case, we cannot expect to recover the values of the hidden third coordinate due to strong identifiability issues. For example, rescaling the third coordinate by an arbitrary scalar, and modifying the first two equations appropriately gives rise to an infinite family of equations of the same sparsity level which reproduce exactly the original dynamics, restricted to the first two coordinates.

Nevertheless, we can learn an extra latent coordinate; when combined with the first two

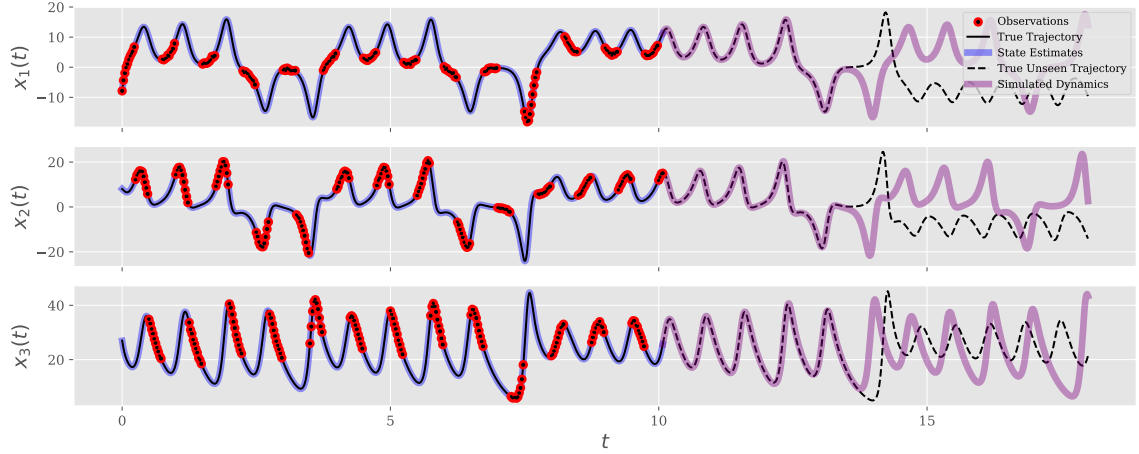


Figure 3.5: Results for Lorenz 63 system with streams of partial observations. Observations alternate between each coordinate every 10 samples at sampling rate $\Delta t = 0.025$ up to $t = 10$ with added noise set to $\sigma^2 = 0.1$. Learned dynamics are simulated from $t = 10$ until $t = 20$.

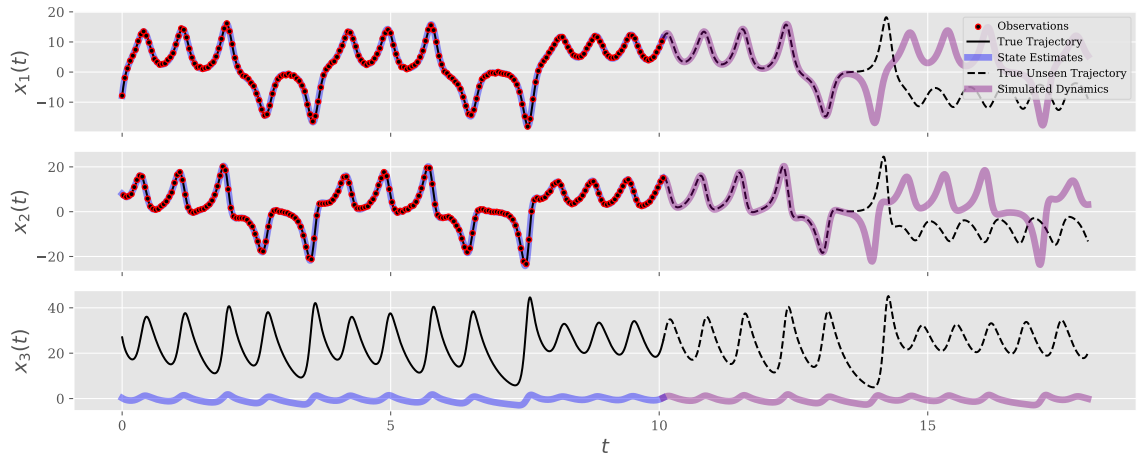


Figure 3.6: Results for Lorenz 63 system with an unobserved coordinate. Observations alternate between first two coordinates at sampling rate $\Delta t = 0.025$ up to $t = 10$, with no knowledge of the third, except that the system is known to be three dimensional. Our algorithm constructs a third, latent coordinate and corresponding dynamics which, when evolved together with the first two coordinates, reproduces the observed dynamics. Further, we can effectively extrapolate the time series forward by using our learned dynamics and imputing the value of our artificial coordinate as a new initial condition, generating good predictions for the first two coordinates.

observed coordinates, these together satisfy a modified ODE that reproduces the dynamics on the observed coordinates. This is demonstrated by taking our state estimate at $t = 10$, the last observed data-point, and simulating the learned ODE forward in time to predict the evolution of the first two states. Doing so produces accurate trajectories for four additional cycles, including a lobe switch, before diverging from the true trajectory due to chaos. Note that modeling only the first two states with an ODE is insufficient—the flow would be self-intersecting and cannot satisfy any autonomous ODE without a third coordinate (no chaos in two dimensions). The results can be seen in Figure 3.6 and eq. (3.48). Because the true hidden state is unidentifiable, we compute the state estimation error restricted to the first two coordinates and obtain $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 6.840 \cdot 10^{-3}$. Similarly, since we cannot expect to recover the third coordinate, we omit the coefficient error.

$$\begin{aligned}\frac{dx_1}{dt} &= 0.722 \cdot 1 - 9.969 x_1 + 9.936 x_2 + 0.430 x_3, \\ \frac{dx_2}{dt} &= -0.456 \cdot 1 - 0.453 x_1 - 0.979 x_2 - 0.316 x_3 + 0.408 x_1 x_2 - 8.832 x_1 x_3 + 0.161 x_3^2, \\ \frac{dx_3}{dt} &= -8.626 \cdot 1 + 0.057 x_2 - 2.683 x_3 + 0.133 x_1 x_2 - 0.421 x_1 x_3.\end{aligned}\tag{3.48}$$

Traditional approaches to handling partially observed systems and modeling time series often enrich the dynamics by inserting time delay coordinates [301–308]. However, that comes with its own challenges, such as selecting the number and spacing of delay coordinates, difficulties in continuous time, especially when observations are not evenly spaced, and introducing bias due to possibly noisy inputs. Our approach has its own drawback of requiring a separate state estimation problem to make predictions, and producing a somewhat uninterpretable third coordinate. We remark that it is possible to include time delay embeddings in our approach as well, essentially turning our collocation method for differential equations into a collocation approach for a *delay* differential equations—we leave this to future work.

3.4.3 Higher order systems

Our approach is also capable of learning higher order differential equations given only state measurements. While devices exist that can directly measure first and second order derivatives (e.g. Doppler radar, gyroscopes, accelerometers), measurements of the system state are the most common, and second order laws are ubiquitous in physics. We consider the Van der Pol oscillator, a coupled version of the Duffing system, and the nonlinear pendulum under model misspecification.

Van der Pol oscillator

The Van der Pol oscillator is governed by the second-order differential equation

$$\frac{d^2 x}{dt^2} = \mu(1 - x^2) \frac{dx}{dt} - x.\tag{3.49}$$

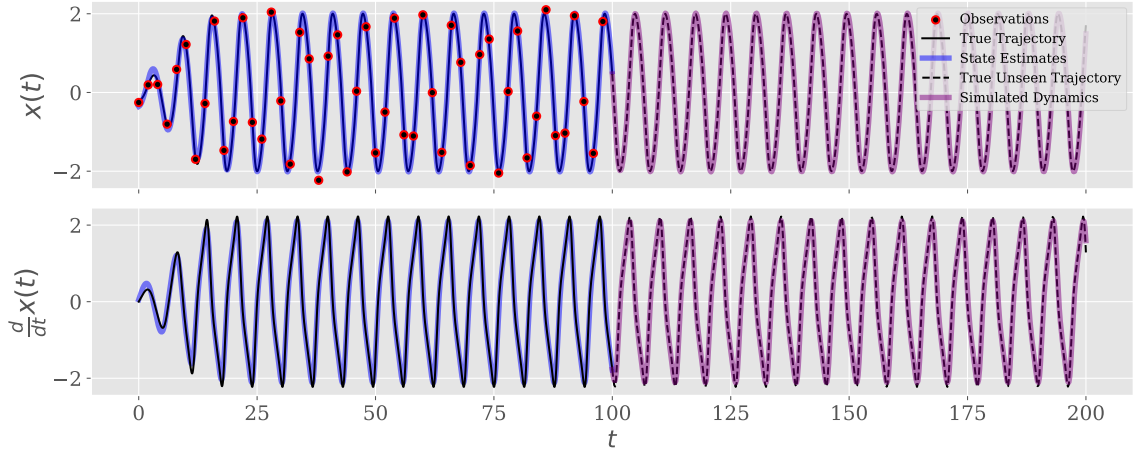


Figure 3.7: Results for the Van der Pol oscillator. Sampling is done only on system state, $\mathbf{x}$, with sampling rate $\Delta t = 2$ up to $t = 100$ with simulated dynamics from $t = 100$ to $t = 200$ using the learned equations from eq. (3.50).

We set $\mu = 0.5$ and note that for $\mu > 0$, Van der Pol systems demonstrate an attractive limit cycle with fast and slow branches. We trained on equally sampled measurements with $\Delta t = 2$ for $t \in [0, 100]$ with additive Gaussian noise with variance $\sigma^2 = 0.1$. We took the feature library to contain cubic polynomials of x and $\frac{d}{dt}x$. Our approach obtained the equation

$$\frac{d^2x}{dt^2} = -0.982x + 0.343\frac{dx}{dt} - 0.337\frac{dx}{dt}x^2 \quad (3.50)$$

with coefficient error $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1.858 \cdot 10^{-1}$, and state estimation error $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 4.540 \cdot 10^{-2}$. Our algorithm identifies nearly the correct equation with exactly the correct support.

Coupled Duffing oscillators

As an example of a larger second order ODE, we consider a ring of d Duffing oscillators with nearest-neighbor coupling. For each dimension $i = 1, \dots, d$, $x_i(t)$ satisfies the differential equation

$$\frac{d^2}{dt^2}x_i(t) = -\alpha_i x_i(t) - \beta_i x_i(t)^3 + \mu (x_{i-1}(t) - 2x_i(t) + x_{i+1}(t)), \quad (3.51)$$

with periodic indexing $x_{d+1} := x_1$. These dynamics are Hamiltonian. Set $\mathbf{v} := \frac{d}{dt}\mathbf{x}$, and

$$\mathcal{H}(\mathbf{x}, \mathbf{v}) = \sum_{i=1}^d \frac{1}{2} v_i^2 + \sum_{i=1}^d \left(\frac{\alpha_i}{2} x_i^2 + \frac{\beta_i}{4} x_i^4 \right) + \frac{\mu}{2} \sum_{i=1}^d (x_{i+1} - x_i)^2, \quad (3.52)$$

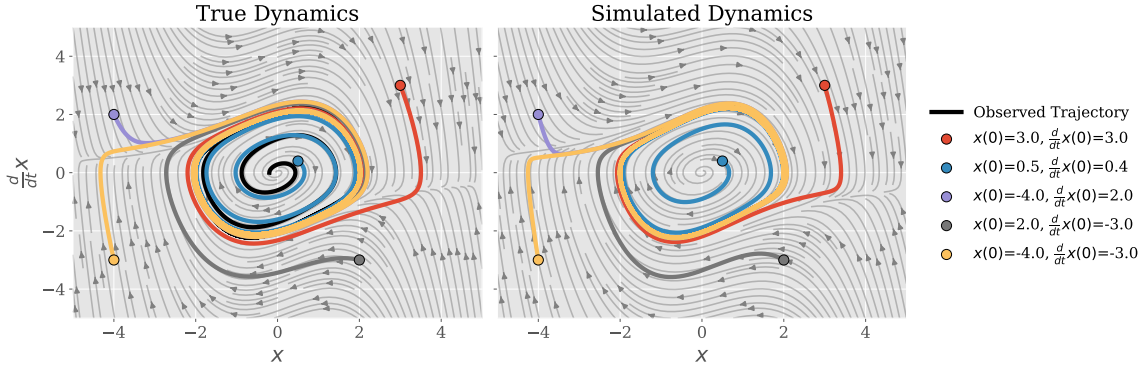


Figure 3.8: Left: True Van der Pol oscillator trajectories plotted from multiple initial conditions on phase portrait using eq. (3.49). Right: Using learned model eq. (3.50) to simulate same trajectories provided same initial conditions.

the dynamics satisfy

$$\frac{d}{dt}\mathbf{x} = \nabla_{\mathbf{v}}\mathcal{H}(\mathbf{x}, \mathbf{v}), \quad \frac{d}{dt}\mathbf{v} = -\nabla_{\mathbf{x}}\mathcal{H}(\mathbf{x}, \mathbf{v}) \quad (3.53)$$

We remark that the nonlinearity has a significant qualitative effect in the dynamics. If we consider a linear version where we drop the cubic restoring force, the dynamics would be governed by the matrix $M := \text{Diag}((\alpha_i)_{i=1}^n) + \mu K$, where K is the graph Laplacian on a cycle. Taking $\mu < 0$ causes the coupling term to encourage exponential blow up, and with μ negative enough relative to α_i , M will have both positive and negative eigenvalues. This implies unbounded level sets in the Hamiltonian, and divergent trajectories in the linear version. However, the positive cubic restoring terms ($\beta_i > 0$) are quartic in the Hamiltonian, which guarantees compact level sets, bounded trajectories [309].

We considered an example in $d = 5$ dimensions with parameters $\mu = -2$, $\alpha_i \stackrel{\text{i.i.d.}}{\sim} \text{Unif}[2, 4]$, $\beta_i = 2$. We initialized the system with all coordinates at the same position and velocity, $x_i(0) = 1$ and $\frac{d}{dt}x_i(0) = 0$ for $i = 1, \dots, d$. We took equally spaced measurements of the position values $x_i(t)$ with a sampling period $\Delta t = 0.2$ between $t = 0$ and $t = 30$ as the input data, and include additive Gaussian noise of variance $\sigma^2 = 0.04$ which corresponds to a signal to noise ratio of 39.2.

Trajectory regularization, data weight, collocation penalties and points remain as in the previous example $\lambda = 10^{-3}$, $\alpha = 1$, $\beta = 10^5$, $m = 1000$. We took the sparsifier to be STLSQ with threshold 0.1, and ridge regularization parameter 0.01. The feature library includes monomials up to cubic order of x_i and $\frac{d}{dt}x_i$, but excludes interaction terms,

$$\Phi = \left\{ 1, x_1, \frac{d}{dt}x_1, \dots, x_5, \frac{d}{dt}x_5, x_1^2, \left(\frac{d}{dt}x_1\right)^2, \dots, x_5^3, \left(\frac{d}{dt}x_5\right)^3 \right\}. \quad (3.54)$$

This library contains 31 functions (down from 286 if we were to include all 10-variable polynomials up to cubic order). Our algorithm correctly identifies the correct support in all

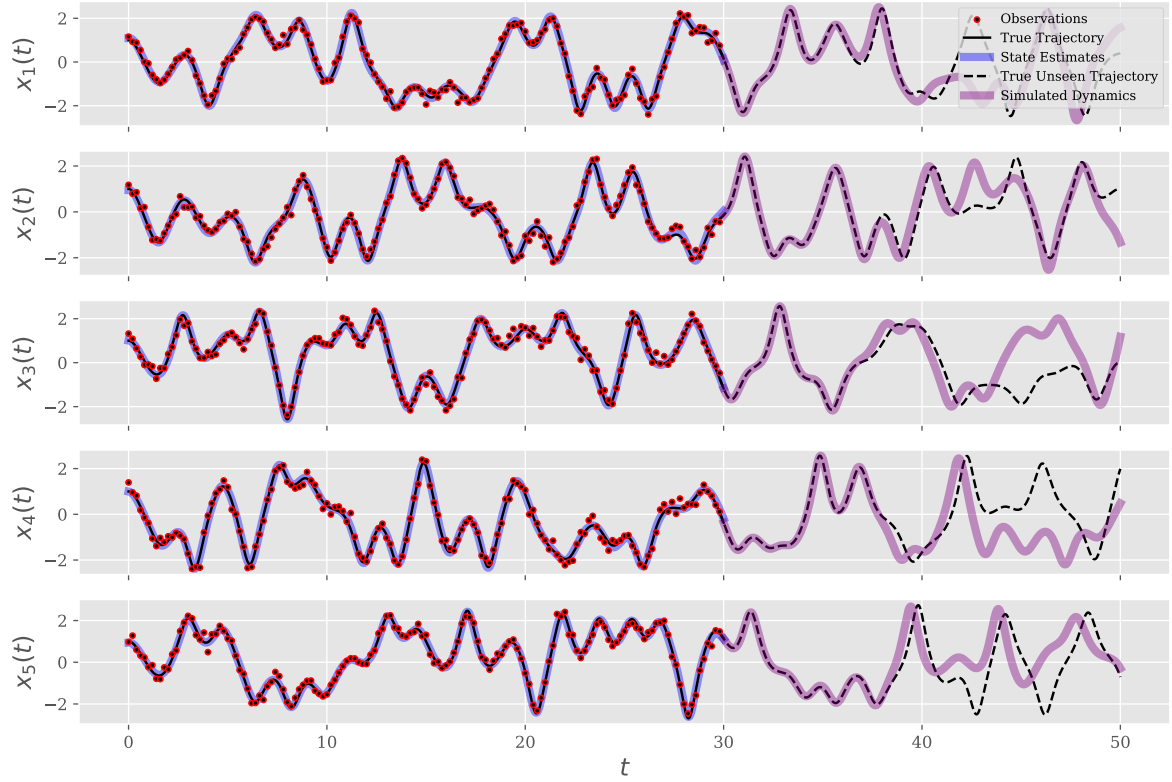


Figure 3.9: Position observations sampled at rate $\Delta t = 0.2$ up to $t = 30$ with added noise set to $\sigma^2 = 0.04$ equating to a relative noise of 2.55%. Learned and true dynamics from eq. (3.55) are simulated from $t = 30$ to $t = 50$.

five variables which amounts to finding the 20 active variables out of 155 with a maximum coefficient error of 0.12.

We show the learned equations on the left, and the true equations on the right.

$$\begin{aligned}
\frac{d^2x_1}{dt^2} &= 1.272x_1 - 2.074x_2 - 2.083x_5 - 1.890x_1^3, & \frac{d^2x_1}{dt^2} &= 1.392x_1 - 2x_2 - 2x_5 - 2x_1^3 \\
\frac{d^2x_2}{dt^2} &= -1.951x_1 + 0.695x_2 - 2.010x_3 - 1.977x_2^3, & \frac{d^2x_2}{dt^2} &= -2x_1 + 0.711x_2 - 2x_3 - 2x_2^3 \\
\frac{d^2x_3}{dt^2} &= -2.016x_2 + 1.862x_3 - 1.956x_4 - 2.045x_3^3, & \frac{d^2x_3}{dt^2} &= -2x_2 + 1.81x_3 - 2x_4 - 2x_3^3 \\
\frac{d^2x_4}{dt^2} &= -2.023x_3 + 1.679x_4 - 2.072x_5 - 2.019x_4^3, & \frac{d^2x_4}{dt^2} &= -2x_3 + 1.521x_4 - 2x_5 - 2x_4^3 \\
\frac{d^2x_5}{dt^2} &= -1.921x_1 - 1.961x_4 + 1.925x_5 - 2.054x_5^3, & \frac{d^2x_5}{dt^2} &= -2x_1 - 2x_4 + 1.988x_5 - 2x_5^3.
\end{aligned} \tag{3.55}$$

We show estimates of the derivatives of one of the state variables and the error in the forward simulations of these learned dynamics starting from the last observation at $t = 30$ in Figure 3.9 with $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 3.554 \cdot 10^{-2}$ and $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 3.627 \cdot 10^{-2}$. Our model produces very accurate derivative estimates despite only having access to noisy observations of the original state variables. The regularized least squares collocation applied to a space of five times differentiable functions allows for derivative estimates up to fourth order to remain accurate, despite the dynamics being only second order. This level of performance is comparable to data simulation approaches where we know the model *beforehand* and apply high order nonlinear Kalman smoothing. The key point is that our approach simultaneously discovers the model and estimates the states.

3.4.4 A misspecified model

We considered a case where the dynamics are misspecified by fitting polynomials to nonlinear pendulum dynamics. Our least squares relaxation of the dynamics allows some of the unmodeled dynamics to be accommodated as an error term, as in eq. (3.63). Robustness to misspecification is very useful; in many experimental or designed processes we may have little to no prior knowledge of the true functional form of the underlying dynamics. In these cases, we hope to maintain highly accurate state estimation (despite the misspecified dynamics), while producing a reasonable model for simulating the future.

We consider the nonlinear pendulum,

$$\frac{d^2x}{dt^2} = -\sin(x), \tag{3.56}$$

and fit two models, one that is linear and another that is cubic. Simulated trajectories were created with initial displacement of $x(0) = 3$ and no initial velocity $\frac{d}{dt}x(0) = 0$. Observations were made with a sampling period of $\Delta t = 1$ over the interval $t \in [0, 100]$, with Gaussian noise ($\sigma^2 = 0.01$).

The relatively large initial condition $x(0) = 3$ means that the nonlinearity has a significant effect; replacing $\sin(x)$ with a cubic Taylor series centered around zero results in finite

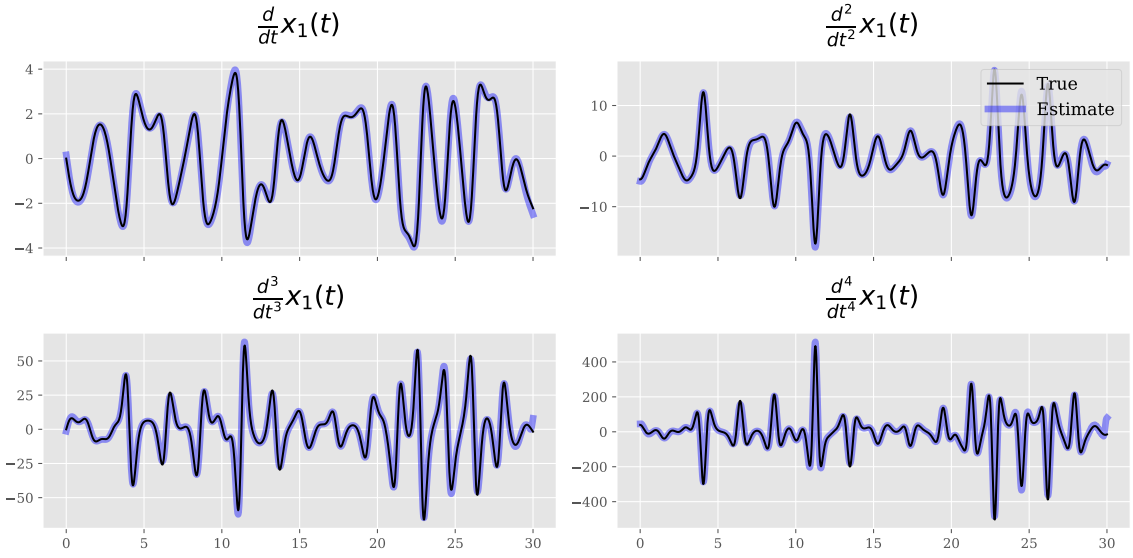


Figure 3.10: Results for estimating time derivatives of the coupled Duffing system. The relative mean square errors of the estimates of the first, second, third, and fourth derivatives are $2.32 \cdot 10^{-3}$, $3.68 \cdot 10^{-3}$, $6.57 \cdot 10^{-3}$ and $1.1 \cdot 10^{-2}$ respectively. In this case, our approach resolves the fine structure in estimating the state.

time blowup with these initial conditions. We obtained the following models:

$$\text{Linear: } \frac{d^2x}{dt^2} = -0.151x \quad (3.57)$$

$$\text{Cubic: } \frac{d^2x}{dt^2} = -0.796x + 0.086x^3. \quad (3.58)$$

Both of these perform very well in state estimation, despite only forming a somewhat crude approximation to the dynamics as evaluated through forward simulation (see Figure 3.11). The filtering error of the linear model was $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 1.250 \cdot 10^{-1}$ and for the cubic model, was $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 2.221 \cdot 10^{-2}$. The linear model matched the period of the true nonlinear dynamics eq. (3.56), but did not have enough degrees of freedom to match the amplitude and misrepresented the shapes of the peaks. The cubic model performed quite well, especially in state estimation and matching the shapes of the oscillations, but has a slightly incorrect period.

3.4.5 Benchmarking

We now consider two systematic experiments testing the robustness of our method to noise, and comparing primarily to ODR-BINDy [269].

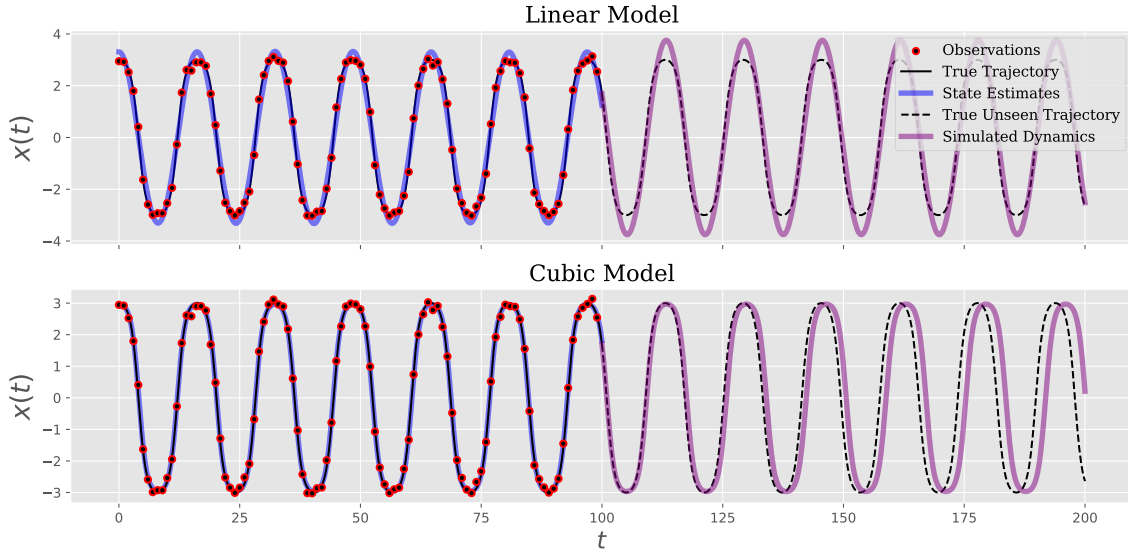


Figure 3.11: Results for misspecified linear and cubic models for nonlinear pendulum. Observations were sampled at a rate $\Delta t = 1$ up to $t = 100$ with added noise with variance $\sigma^2 = 0.01$. Learned dynamics for both linear and cubic model, eq. (3.57) and eq. (3.58), are simulated from $t = 100$ to $t = 200$ where the linear model successfully captures periodicity of the true trajectory, but fails to capture the correct amplitude, whereas this is flipped for the cubic model.

Damped nonlinear oscillator

We considered increasing noise levels with fixed trajectory length and observation frequency for a nonlinear damped oscillator that was studied in [269]. The dynamics are given by

$$\begin{aligned} \frac{d}{dt}x_1 &= -0.1x_1^3 + 2x_2^3 \\ \frac{d}{dt}x_2 &= 2x_1^3 - 0.1x_2^3. \end{aligned} \tag{3.59}$$

We took the initial conditions $\mathbf{x}(0) = (2, 0)$, and took samples of both variables with sampling period $\Delta t = 0.05$ for $t \in [0, 10]$. We first considered an example with noise level $\sigma = 0.32$. We set trajectory regularization $\lambda = 10^{-5}$, data weight $\alpha = 10$, collocation penalty $\beta = 10^5$, and used STLSQ with threshold 0.08 and ridge parameter 100. This is a surprisingly challenging example, especially with this level of noise—our model does not identify the correct support, but still performs quite well in estimating the state. We obtained $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x}) = 6.60 \times 10^{-2}$, $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 2.47 \times 10^{-1}$, and the equations

$$\begin{aligned} \frac{d}{dt}x_1 &= 0.078x_2 - 0.560x_1x_2^2 + 1.859x_2^3 \\ \frac{d}{dt}x_2 &= -0.008x_1 - 0.212x_1^2 - 1.950x_1^3 - 0.284x_1x_2^2. \end{aligned} \tag{3.60}$$

The oscillatory terms were recovered with reasonable accuracy, but the small damping was hard to capture correctly. From Figure 3.12, we can see that the state estimates and reproduced dynamics are still close, which suggests some ill-posedness in the sparse recovery problem.

Next, we tested on a sequence of noise levels $\sigma = 0.02, 0.04, 0.08, 0.16, 0.32, 0.64$ and averaged the values of $\text{RMSE}_{\text{Var}}(\hat{\mathbf{x}}, \mathbf{x})$ and $\text{RMSE}_{\text{Fro}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ over 32 random seeds. In this case, we found that including ensembling [151] in the inner sparsifier greatly improved the consistency of the results for higher noise levels. We compared five models, weak SINDy [233], weak SINDy with ensembling, JSINDy, JSINDy with ensembling, and ODR-BINDy [151, 269]. We used PySINDy for ensembling and weak SINDy [268], and the GitHub repository² associated to [269] for ODR-BINDy. We also tested with other two-step variants of SINDy, but they all performed worse than weak SINDy, so we omit them. As ODR-BINDy also produces state estimates, we compare the state estimation error between JSINDy, JSINDy with ensembling and ODR-BINDy.

We note that our ensembling only occurs inside of the sparsifier $\mathcal{S}$, and therefore does not incur much additional computational cost, which is dominated by the nonlinear least squares problems rather than the sparsifying iterations. In cases where the model is already successful, ensembling can introduce additional extraneous features, but for challenging problems, it seems to improve the robustness. Perhaps surprisingly, the error of weak ensemble SINDy was quite high for low levels of noise, but did not degrade as noise was increased for the scale that we considered.

In terms of the coefficient error, and without ensembling the sparsifier, our model performs noticeably worse than ODR-BINDy; maximizing the Bayesian evidence seems to be a good model selection criterion. Upon introducing ensembling, our model selection stabilizes and the coefficient errors become comparable. In terms of state estimation, JSINDy both with and without ensembling performs well across varying levels of noise, regardless of the coefficient accuracy, owing to the stabilizing effect of the RKHS regularization and least squares relaxation.

Lorenz 63 example

We also tested the performance of our method on varying levels of noise and durations of observations with fixed observation frequency on the Lorenz 63 dynamics (3.42) with results in Figure 3.14. We applied JSINDy with data weight $\alpha = 1$, collocation weight $\beta = 10^5$, regularization $\lambda = 10^{-5}$, and used SR3 as a sparsifier with an L1 penalty with weight 5, and relaxation coefficient $\nu = 0.05$ [292]. We compare to ensemble weak SINDy and ODR-BINDy. The setup is identical to that of [269, Sec. 3.2], though we did not exclude the constant function from the library for weak SINDy. ODR-BINDy performed best in variable selection, likely owing to their approach of Bayesian evidence maximization. While also better than JSINDy in coefficient error, the difference is somewhat less pronounced. A much larger difference can be seen between JSINDy and ensemble weak SINDy, as well as the other models shown in [269, Sec. 3.2].

²<https://github.com/llfung/ODR-BINDy>

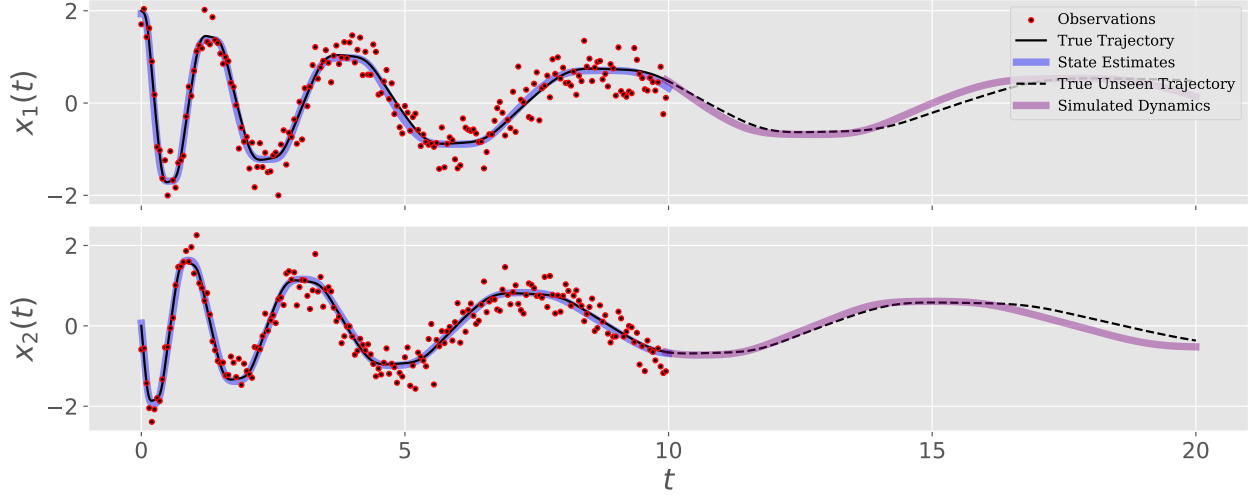


Figure 3.12: Results for the nonlinear damped oscillator with full state observations, $\Delta t = 0.05$ in the interval $[0, 10]$, and $\sigma = 0.32$ additive noise. While the learned equations in eq. (3.60) do not match the true dynamics in terms of the coefficient values, JSINDy effectively infers the states, and produces a reasonable extrapolation in forward simulation.

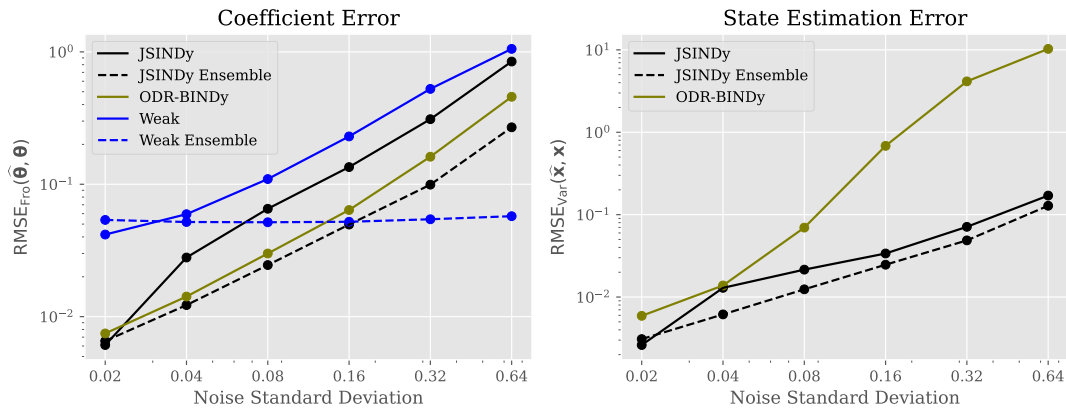


Figure 3.13: Results for the damped nonlinear oscillator across varying levels of noise. Left: coefficient error $\text{RMSE}_{\text{Fro}}(\hat{\theta}, \theta)$, Right: state estimation error $\text{RMSE}_{\text{Var}}(\hat{x}, x)$. In this case, ensembling significantly improves the coefficient error for both weak SINDy and JSINDy. For JSINDy, including ensembling inside of the sparsifier brings the coefficient error from noticeably worse than ODR-BINDy to slightly better. Overall, JSINDy performs very well in state estimation across varying levels of noise.

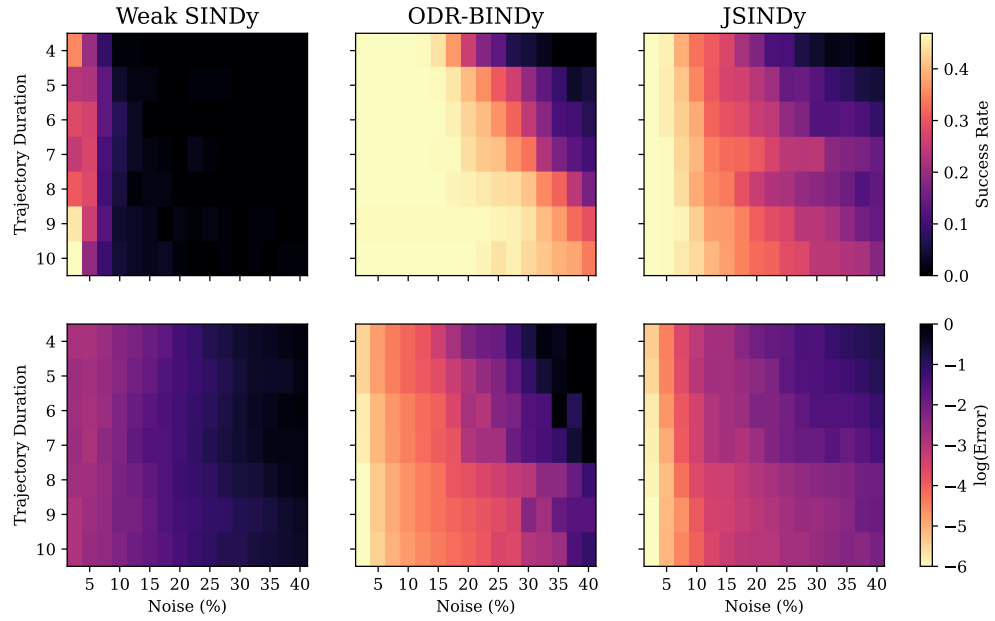


Figure 3.14: Experiment across different trajectory lengths and noise levels, averaged over 128 independent trials and $\Delta t = 0.01$. **Top:** Proportion of trials where the correct support was exactly recovered **Bottom:** Relative coefficient error $\text{RMSE}_{\text{Fro}}(\hat{\theta}, \theta)$ of identified system.

3.5 Discussion and conclusion

A key technique used in this work is the least squares relaxation of the ODE constraint:

$$\frac{d^p}{dt^p} \mathbf{x} = f(D[\mathbf{x}](t); \boldsymbol{\theta}) \quad (3.61)$$

into the penalty term

$$\mathcal{L}_C(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{2} \int_0^T \left| \frac{d^p}{dt^p} \mathbf{x} - f(D[\mathbf{x}](t); \boldsymbol{\theta}) \right|^2 dt. \quad (3.62)$$

In this section, we provide some insight into the mathematical underpinnings of this approach, discuss the theoretical advantages, and highlight the role of regularization, while providing additional context and reference to trace many of these ideas in the literature.

The least squares relaxation has been applied in the PDE inverse problem setting [282], where it was found to mitigate the effects of local minima [284]. It was further analyzed in [281], which presented a representer theorem for relaxations of parametric linear PDEs and cast the inversion problem as minimizing a special RKHS norm. The least squares approach is also common in the ML literature, most notably through physics-informed neural networks [141], as well as approaches to learning differential equations and parameter estimation [2, 159, 160, 269, 310, 311]. In our case, the additional error associated to the ODE from applying a least-squares approach can be beneficial to the overall learning process, rather than a limitation. We prefer the relaxation approach to alternatives, such as reduced or shooting methods that view the ODE as a constraint and enforce feasibility from the start, which can be seen in neural ODEs, inverse problems, and optimal control [312–315], or formulations that enforce the hard constraint within an all-at-once formulation such as [270], which optimizes by solving the discretized KKT conditions.

The RKHS regularized least squares approach that we take has two major advantages over reduced approaches. First, state estimates for any length of time are well defined for a much broader class of ODEs, which can be seen from the existence of bounded solutions in Theorem 3.2. Further, the state estimates are allowed to be self-correcting in that the additional data misfit terms may inform the state estimates. In contrast, in reduced approaches, certain parameter values may lead to finite time divergence of the ODE. Second, this approach improves the regularity of the objective with respect to the system parameters $\boldsymbol{\theta}$. For chaotic dynamics, we can generically expect that small perturbations to parameters of the system will lead to exponentially large changes over long periods of time relative to the Lyapunov exponent. Several works have observed that this leads to extremely ill-conditioned and non-convex objectives in reduced formulations, especially in chaotic regimes where the mismatch between simulated and observed trajectories can grow exponentially with time [316–318]. As a result, gradient-based optimization often stalls, converges to spurious local minima, or fails entirely. These issues can be mitigated significantly through multiple-shooting and multipoint-penalty methods [317–319], which break a long trajectory into shorter segments, while simulating within each segment. The full space approaches considered in this work can, in a sense, be seen as the limit of many segments in a multiple-shooting method, but we leave a more systematic comparison to future work.

The least squares relaxation avoids simulating forward trajectories altogether and instead distributes the residual over time, substantially smoothing the optimization landscape and improving robustness. Further, it makes the objective a *linear* least squares regression problem in the coefficients $\boldsymbol{\theta}$ for any fixed state estimate $\mathbf{x}$. This natural perspective also appears in [281] in the context of recovering a coefficient field in a PDE-constrained inverse problem. [269] points out the connection to orthogonal distance regression in all-at-once approaches, casting optimization with respect to the state variables as adjusting the inputs to correct an error-in-variables bias. Indeed, in the linear case, such an approach appears as early as 1950 in Householder’s method for exponential fitting [320], where the error-in-variables issue in Prony’s method is corrected by jointly fitting state estimates and discrete time dynamics.

Least squares relaxation of the ODE constraint can also be seen as minimizing a backwards error of the ODE. Introducing an auxiliary forcing function $\mathbf{r}(t)$, we may equivalently write the least squares penalty as

$$\mathcal{L}_C(\mathbf{x}, \boldsymbol{\theta}) = \begin{cases} \min_{\mathbf{r} \in L^2([0, T], \mathbb{R}^d)} \int_0^T |\mathbf{r}(t)|^2 dt \\ \text{subject to } \frac{d^p}{dt^p} \mathbf{x}(t) = f(D[\mathbf{x}](t); \boldsymbol{\theta}) + \mathbf{r}(t) \end{cases}, \quad (3.63)$$

where the ODE is understood in the weak sense.

When this formulation of $\mathcal{L}_C$ is embedded into the optimization problem eq. (3.7) with an appropriate lifting to optimize with respect to $\mathbf{r}$, this can be interpreted as finding a control or forcing term $\mathbf{r}(t)$ with small L^2 norm which is optimized to steer $\mathbf{x}$ to match the data and have small RKHS norm. This view is also well motivated when the model is misspecified, where $\mathbf{r}$ can act as an error term in the dynamics.

While one can control the error in terms of the size of the residual $\mathbf{r}$ and a Lipschitz assumption on f via classical error analysis for ODEs using Grönwall’s inequality, this will generally result in overly pessimistic error estimates which grow exponentially in time as it assumes worst case behavior of $\mathbf{r}$. Rather, $\mathbf{r}$ largely pushes the trajectory to match the data, so its effects should be small when the model is well specified, and large in misspecified cases, where it is necessary. By assimilating the observation data over a long trajectory, the error can be compared to that of Kalman filtering/smoothing (see e.g. [321]) under very low system noise and a reasonably accurate system model.

In our case, the regularity of $\mathbf{r}$ follows from the smoothness of $\mathbf{x}$ lent by the RKHS regularization and continuity of f , noting that $\mathbf{r}(t) = \frac{d^p}{dt^p} \mathbf{x}(t) - f(D[\mathbf{x}](t); \boldsymbol{\theta})$, the difference of two continuous functions. However, without regularization of $\mathbf{x}$, there is no reason to believe that $\mathbf{r}$ is continuous, so $\mathbf{x}$ will only be piecewise p -times differentiable; for first order ODEs, one should expect kinks in the state estimates without higher order regularization. The closely related approach of ODR-BINDy [269] interprets the least squares approach as a stochastic relaxation, modeling error in the ODE as independent white noise. However, it can be difficult to rigorously interpret the noise stochastically taking the function space limit of inferring $\mathbf{x}(\cdot)$; in the random setting and without regularization, the covariance of the residuals becomes increasingly important as the mesh is refined and $\mathbf{r}(t)$ is considered as a random process. In the case that the residuals are uncorrelated (e.g. $\mathbf{r}(t), \mathbf{r}(t')$ are

independent for $t \neq t'$), then either the pointwise variances have to blow up, or the noise completely washes out under the integral. The probabilistic model can be seen as applying maximum a posteriori estimation to the stochastic differential equation

$$d\mathbf{x} = f(\mathbf{x})dt + \epsilon d\mathbf{B}_t. \quad (3.64)$$

In this model, $\frac{d}{dt}\mathbf{x}$ will not be defined pointwise, possessing only the regularity of white noise. Of course, these considerations do not manifest computationally; the error terms end up correlated, there is no “real” Brownian motion happening, and most reasonable discretizations will implicitly regularize the trajectory estimate. However, this does raise the question of determining a computationally effective and complete stochastic model in the vein of probabilistic numerics [322], as well as the idea of penalizing the residual in a stronger norm than L^2 .

3.5.1 Conclusion

We proposed and demonstrated the effectiveness of JSINDy in both system identification and filtering by performing both jointly. We addressed a number of challenging problems in continuous time system identification within a single framework based on collocation in RKHSs, including scarce sampling, far below what has been considered in previous works, partially observed states, and higher order dynamics. Building on other recent advancements in all-at-once system identification methods [159, 160, 269, 270, 279], this work provides more strong evidence for the superiority of all-at-once methods for system identification, and demonstrates significant flexibility. In short, the assumption that the data satisfy some dynamics is a very good prior to use when it is true. Overall, the RKHS formulation provides a flexible and effective tool for modeling while providing a principled approach to regularization, resulting in stable optimization, and allowing for seamlessly incorporating prior assumptions, data, and dynamics.

Currently, the primary drawback of JSINDy is that dense kernel matrices arise from the global approximation, leading to cubic complexity in the trajectory length. This can, however, be mitigated through a state-space representation of Matérn Gaussian processes as in [323], which essentially amounts to a finite difference approximation to the associated Sobolev norm, and would result in banded Jacobian matrices.

Further improvements are available on the side of optimization. The objective function in eq. (3.10) is amenable to partial minimization [258, 324, 325] with respect to the coefficients $\boldsymbol{\theta}$ in solving the fixed-support nonlinear least squares problems. Our approach of alternating with an algorithmic sparsifier is another choice which could use more investigation. In particular, incorporating Bayesian evidence as a model selection criteria could be highly effective, while raising interesting questions for efficiently solving the discrete optimization problems that arise [269, 326, 327].

Acknowledgments

AH and BH acknowledge support from the National Science Foundation under awards 2208535 (Machine Learning for Bayesian Inverse Problems) and 2337678 (CAREER: Gaussian Processes for Scientific Machine Learning: Theoretical Analysis and Computational

Algorithms). AH acknowledges support from a Carl E. Pearson Fellowship. JMSH was partially supported by the Veterans Administration via the Post-9/11 GI Bill. JNK was supported in part by the NSF AI Institute for Dynamical Systems (dynamicsai.org), grant 2112085. The numerical experiments involving ODR-BINDy were completed on Hyak, UW's high performance computing cluster, which is funded by the UW student technology fee.

3.6 Theory

We provide some more details of the RKHS theory that underlies our approach and the least squares relaxation approach we take for the ODE.

3.6.1 RKHS and representer theorems

We first give a version of the representer theorem in the Hilbert space setting which is appropriate for Theorem 3.1. These theorems are well known and appear in various forms throughout the literature; we state this for completeness. The variant that we present can be found in [166, 328, 329], and similar results appear in [142, 260, 286].

Theorem 3.3 (Representer). *Let $\mathcal{H}$ be a separable Hilbert space. Let $F : \mathcal{H} \mapsto \mathbb{R}$ be given by*

$$F(u) = f(\langle u, w_1 \rangle, \dots, \langle u, w_m \rangle) + J(\|u\|) \quad (3.65)$$

where $J : \mathbb{R}^+ \mapsto \mathbb{R} \cup \{\infty\}$ is strictly increasing, $\langle \cdot, w_i \rangle$ are bounded linear functionals, and $f : \mathbb{R}^m \mapsto \mathbb{R} \cup \{\infty\}$. Then if F admits a minimizer u^ with $F(u^*) < \infty$, then*

$$u^* = \sum_{i=1}^m c_i w_i. \quad (3.66)$$

3.6.2 Matérn kernels and Sobolev spaces

Here, we review some prerequisite RKHS and Sobolev space theory, giving more details on the kernel functions we use, and discussing the smoothness of functions in RKHS $\mathcal{X}$ associated to the kernels we define in eq. (3.11). Much of this will pull directly from [260, 330] where we direct the interested reader for additional details. First, recall from eq. (3.11) that we take the sum of a constant kernel with a Matérn kernel

$$k(t, t') = c_0 + c_1 h_\nu \left(\frac{|t - t'|}{\ell} \right), \quad h_\nu(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r \right)^\nu K_\nu \left(\sqrt{2\nu} r \right). \quad (3.67)$$

From [260, Ex. 2.6], the RKHS $\mathcal{H}_k$ associated to this kernel is norm equivalent to the Sobolev space $H^{\nu+\frac{1}{2}}([0, T])$, functions which are $\nu + \frac{1}{2}$ times differentiable in the mean-square sense. Compactness of the domain means that adding a constant to the kernel function does not affect the equivalence. By the Sobolev embedding theorem [331, Thm. 5.6.6, p. 270], this implies that $\mathcal{H}_k$ is also compactly embedded in the Hölder space $C^p([0, T])$ for $p < \nu$.

When ν is a half integer (i.e. $\nu = p + \frac{1}{2}$), h_ν reduces to the product of a polynomial and exponential in r .

$$h_\nu(r) = \exp\left(-\sqrt{2\nu}r\right) \frac{p!}{(2p)!} \sum_{i=0}^p \frac{(\nu - \frac{1}{2} + i)!}{i!(\nu - \frac{1}{2} - i)!} \left(2\sqrt{2\nu}r\right)^{\nu - \frac{1}{2} - i}. \quad (3.68)$$

Stably computing high order derivatives of this h_ν can still be challenging due to the absolute value inside of h_ν , and directly applying autodiff to this representation can cause catastrophic cancellation. However, for $j < 2\nu$, $\frac{d^j}{dt^j} h_\nu(\frac{|t-t'|}{\ell})$, can also be written as the product of a polynomial and exponential function; we compute this representation of the derivatives in SymPy [332]. We then convert the resulting expressions to Jax [333] using Sympy2Jax [334].

3.6.3 Cholesky factorization and choice of basis

In our numerical implementation, we perform a linear change of variables in order to simplify bookkeeping and mitigate some of the effects of ill-conditioning and numerical roundoff on convergence by reparametrizing with the Cholesky factors of $\mathbf{K}$. That is, write

$$\mathbf{K} = \mathbf{C}\mathbf{C}^\top \quad (3.69)$$

for the Cholesky factorization of $\mathbf{K}$, and set

$$\tilde{\mathbf{z}} = (\tilde{z}_1, \dots, \tilde{z}_d), \quad \tilde{z}_m = \mathbf{C}^\top \mathbf{z}_m. \quad (3.70)$$

Under this parametrization, we have that

$$\mathbf{x}(\cdot; (\mathbf{C}^{-T}\tilde{\mathbf{z}}_1, \dots, \mathbf{C}^{-T}\tilde{\mathbf{z}}_d)) = \mathbf{x}(\cdot; \mathbf{z}). \quad (3.71)$$

It is clear that these are interchangeable *mathematically*. Computationally, they relegate some of the bookkeeping with tracking the kernel matrix to an initial Cholesky factorization, which, in our implementation, is absorbed into the parametrization. Even though this still entails factorizations of ill-conditioned matrices, and direct solvers are also applied downstream, this empirically seems to mitigate some of the effects of ill-conditioned Jacobian matrices in solving regularized least-squares problems.

3.7 Computational tools

We perform all of our computations using Jax [333], and solve ODEs using the Tsit5 [335] solver with a PID controller for adaptive timestepping [336, 337] as recommended by, and implemented in Diffrax [312, 338]. Comparisons with SINDy use the PySINDy package [268]. Comparisons with ODR-BINDy were made using the implementation³ from the original work [269]. Our experiments and current implementation are available in the GitHub repository [339]; an optimized implementation of JSINDy will soon be added to PySINDy. Our computations with JAX used an RTX-5000 Ada Generation GPU in a Lambda machine.

³<https://github.com/llfung/ODR-BINDy>

Chapter 4

SHARP BOUNDS ON THE FAILURE OF THE HOT SPOTS
CONJECTURE*Preface*

In this chapter, from [4] which is joint work with Jaume de Dios Pont and Mitchell A. Taylor, we determine the sharp value of the hot spots ratio in every dimension and study its asymptotic behaviour as the dimension grows. This collaboration began at the Mathematisches Forschungsinstitut Oberwolfach in the fall of 2024, at the Arbeitsgemeinschaft on Quantum Signal Processing and Nonlinear Fourier Analysis organized by András Gilyén, Lin Lin and Christoph Thiele.

The hot spots conjecture, posed by Rauch at a conference in 1974, posits that the maximum of the first nonconstant Neumann Laplacian eigenfunction occurs on the boundary of the domain. While known to be false in general [340], precisely quantifying the degree of failure remained open: How much larger can the maximum on the interior be than the maximum on the boundary?

Let Ω denote a Lipschitz domain, and ψ_1 be the first nontrivial Neumann eigenfunction on this domain, satisfying

$$\begin{aligned} -\Delta\psi_1(x) &= \mu_1\psi_1(x), & x \in \Omega \\ \partial_n\psi_1(x) &= 0, & x \in \partial\Omega \end{aligned}$$

for $\mu_1 > 0$. Define

$$R(\Omega) := \max_{x \in \Omega} \psi_1(x) / \max_{x \in \partial\Omega} \psi_1(x) \text{ and } S_d := \sup_{\text{Lipschitz } \Omega \subset \mathbb{R}^d} R(\Omega).$$

The original work [341] of Steinerberger that posed this question proved that $S_d < 60$ for all d , which was improved in subsequent years [342]. Our work determined the exact value of S_d for every dimension d , proving both new upper bounds and showing they are sharp through explicit constructions. The proof of the upper bound combines a comparison principle for elliptic equations with Talenti's rearrangement inequality. A quantitative form [343] of the Szegő–Weinberger eigenvalue inequality [344, 345] — that the ball uniquely maximizes the first nontrivial Neumann eigenvalue among domains of equal volume — then converts near-equality in the upper bound into a rate that extremizing sequences must approach a ball, in terms of Fraenkel asymmetry. The construction achieving the lower bound uses perforated domains that, in an appropriate limit, allow for jump discontinuities in the eigenfunction, effectively creating separate interior and exterior regions, and reducing the eigenvalue problem on a perforated disk into a radial problem with an interior jump condition.

4.1 Introduction

Rauch's *hot spots* conjecture, first posed in his 1974 lectures at Tulane University [346], is a statement about the long-term behavior of solutions to the heat equation in an insulated domain. Given a bounded, connected domain $\Omega \subset \mathbb{R}^d$, let ψ_Ω denote the first nontrivial eigenfunction of the Laplace operator in Ω with Neumann boundary conditions. In physical terms, ψ_Ω represents the fundamental mode of heat dissipation in the insulated body Ω , and its behavior dominates the asymptotic temperature fluctuations for generic initial temperature distributions. Rauch's *hot spots* conjecture asserted that, for sufficiently regular domains, the points where ψ_Ω attains its maximal and minimal values must lie on the boundary $\partial\Omega$. In other words, for generic initial data, the extrema of the temperature fluctuations should migrate towards $\partial\Omega$ as $t \rightarrow \infty$.

Initially, this conjecture was believed to hold for *all* sufficiently regular domains in $\mathbb{R}^d$. Burdzy and Werner [340] disproved this in two dimensions by constructing a counterexample with three holes (with later examples with one hole provided in [347]). Most attention then turned towards convex domains, where the conjecture was thought to hold in arbitrary dimensions. This was recently disproved by the first author [348] in all sufficiently high dimensions.

Despite these counterexamples, the conjecture has been proven for various classes of domains. In $\mathbb{R}^2$, these include all triangles [349, 350], domains that are long and thin in different senses [351–353], and convex sets with one axis of symmetry [354, 355]. It is widely believed that the conjecture holds for all simply connected sets in $\mathbb{R}^2$.

In higher dimensions, positive results have been established for cylinder sets [356], certain thin sets that are rotationally symmetric in all but one dimension [357], and domains generalizing the two-dimensional class introduced by Atar and Burdzy [352], such as [358, 359].

The existence of counterexamples in various settings raises the following fundamental question: *How badly does the hot spots conjecture fail?* This question has been around since the first counterexamples were found in the late 90s, and its importance was reiterated in recent years by Steinerberger [341], who introduced the hot spots ratio, a quantitative measure of the conjecture's failure.

Definition 4.1 ([341, 342]). Let S_d be the supremum, among all connected, bounded domains $\Omega \subset \mathbb{R}^d$ with Lipschitz boundary, of the ratio

$$\frac{\max_{x \in \Omega} \psi_\Omega(x)}{\max_{x \in \partial\Omega} \psi_\Omega(x)},$$

where ψ_Ω is any of the first non-constant Laplace eigenfunctions of Ω with Neumann boundary conditions.

In this paper, we determine the exact value of S_d in all dimensions. Approximate numerical values are presented in Table 4.1. We further show that extremizers to this ratio do not exist for $d > 1$, and that extremizing sequences must converge to a ball at a precise quantitative rate. As $d \rightarrow \infty$, we show that S_d converges to $\sqrt{e}$, matching the previously best known upper bound from [342].

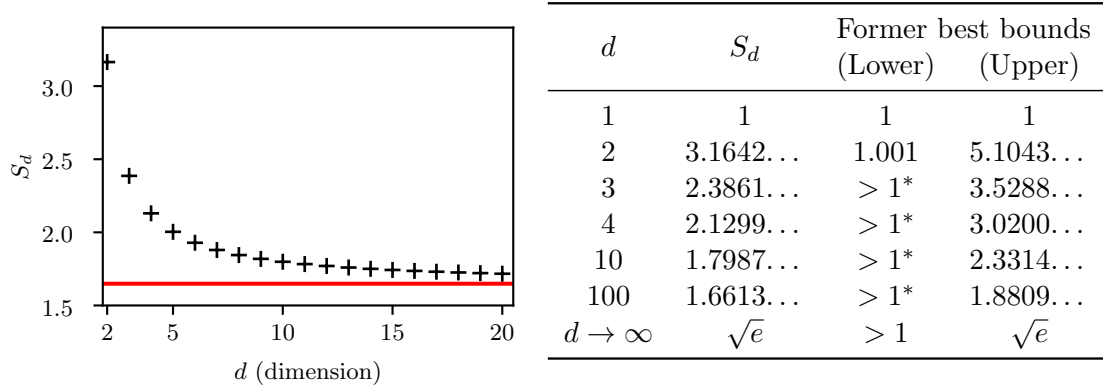


Figure 4.1: (Left) Plot of S_d as a function of d with the asymptotic value $S_d \rightarrow \sqrt{e}$ marked by a red line. (Right) Approximate values of S_d for different dimensions d along with the previously best known bounds. The value > 1 means that no specific value had been computed (to the knowledge of the authors), but the value is known to be > 1 . Results marked with $*$ were *folklore* results. The fact that $\liminf_{d \rightarrow \infty} S_d > 1$, without a specific computed value, follows from the proof of [348]. The lower bound in $d = 2$ comes from numerical experiments [360]. All previously best known upper bounds follow from [342].

To state our main results precisely, we recall the following basic notation and terminology.

Notation for eigenfunctions.

1. We will denote by $\varphi_{\Omega}^{(k)}$ the k -th Dirichlet eigenfunction, of eigenvalue $\lambda_{\Omega}^{(k)}$, with k starting at 1. When k is omitted, it is assumed to be 1.
2. We will denote by $\psi_{\Omega}^{(k)}$ the k -th Neumann eigenfunction, of eigenvalue $\mu_{\Omega}^{(k)}$, with k starting at 0 (so that $\psi_{\Omega}^{(0)}$ is the constant function and $\psi_{\Omega}^{(1)}$ is the first nontrivial eigenfunction). When k is omitted, it is assumed to be 1.
3. When a function (such as $\varphi_{B_1(\mathbb{R}^d)}$) is radial, we will, in an abuse of notation, write

$$\varphi_{B_1(\mathbb{R}^d)}(x) = \varphi_{B_1(\mathbb{R}^d)}(|x|).$$

We will prove that sequences of domains extremizing the hot spots ratio must (in a certain sense) converge to a ball (see Figure 4.2). The reason that the ball is not a maximizer is because its first nontrivial Neumann eigenfunction is not radial. As we will see, the almost extremizing domains for the hot spots constant are slight modifications of the ball that *drop* the first radial eigenvalue right below the first non-radial eigenvalue. The resulting radial eigenfunction then dictates the degree of failure of the *hot spots* conjecture.

Definition 4.2. We will denote by

$$\eta_d(x) := \frac{\varphi_{B_1(\mathbb{R}^d)} \left(x \sqrt{\mu_{B_1(\mathbb{R}^d)} / \lambda_{B_1(\mathbb{R}^d)}} \right)}{\varphi_{B_1(\mathbb{R}^d)} \left(\sqrt{\mu_{B_1(\mathbb{R}^d)} / \lambda_{B_1(\mathbb{R}^d)}} \right)} = \frac{x^{1-\frac{d}{2}} J_{\frac{d}{2}-1} \left(x \sqrt{\mu_{B_1(\mathbb{R}^d)}} \right)}{J_{\frac{d}{2}-1} \left(\sqrt{\mu_{B_1(\mathbb{R}^d)}} \right)}.$$

This function, with domain $B_1(\mathbb{R}^d)$, is the unique solution to the PDE

$$\begin{cases} -\Delta \eta_d(x) = \mu_{B_1(\mathbb{R}^d)} \eta_d(x) & \text{if } |x| < 1, \\ \eta_d(x) = 1 & \text{if } |x| = 1. \end{cases}$$

The function $\eta_d(x)$ is not a Neumann eigenfunction of the ball (in that it does not satisfy the Neumann boundary conditions), but it is a Laplace eigenfunction on the interior. It will arise as the locally uniform limit of a sequence of Neumann eigenfunctions of a sequence of domains converging to a ball.

Our first main result states that the failure modes of the *hot spots* conjecture are tightly controlled by the function $\eta_d(x)$. Given a domain Ω with the same volume as the unit ball, its *Fraenkel asymmetry* $\mathcal{A}(\Omega)$ measures how different Ω is from the ball:

$$\mathcal{A}(\Omega) := \min_{x \in \mathbb{R}^d} |(x + B_1(\mathbb{R}^d)) \Delta \Omega|.$$

Theorem 4.3. *Let $\eta_d(x)$ be as above. Then for any $d \geq 2$,*

$$S_d = \|\eta_d(x)\|_{L^\infty(B_1(\mathbb{R}^d))} = \eta_d(0).$$

Moreover, the value of S_d is not achieved by any set in $\mathbb{R}^d$ with Lipschitz boundary. On the other hand, any extremizing sequence of domains must converge to a ball at a quantitative rate: If Ω has the same volume as the unit ball, then

$$S_d - \frac{\max_{x \in \Omega} \psi_\Omega(x)}{\max_{x \in \partial \Omega} \psi_\Omega(x)} < \epsilon^2 \quad \implies \quad \mathcal{A}(\Omega) \leq C_d \epsilon,$$

where $\mathcal{A}(\Omega)$ is the Fraenkel asymmetry of Ω .

Since $\eta_d(x)$ can be written explicitly in terms of Bessel functions, we can compute the limit $\lim_{d \rightarrow \infty} S_d$.

Corollary 4.4. *One has $\lim_{d \rightarrow \infty} S_d = \sqrt{e}$.*

Theorem 4.3 determines the *gap* in the *hot spots* conjecture in terms of the L^∞ norm. One could, alternatively, try to determine the size of the gap *in measure*: How large is the set where ψ_Ω is too large?

Definition 4.5. For $\alpha \in [1, S_d]$ let $V_d(\alpha)$ be the maximum, among all bounded and connected Lipschitz domains Ω in $\mathbb{R}^d$, of the quantity

$$\frac{|\{x \in \Omega \text{ s.t. } \psi_\Omega^{(1)}(x) \geq \alpha \max_{y \in \partial \Omega} \psi_\Omega^{(1)}(y)\}|}{|\Omega|}.$$

Theorem 4.6. *When $\alpha \in [1, S_d]$, the function $V_d(\alpha)$ is given implicitly by*

$$V_d\left(\eta_d(\alpha^{1/d})\right) = \alpha.$$

Theorem 4.6 implies Theorem 4.3 (without the stability result). The proof of both theorems is the same and it yields an analogous stability result in Theorem 4.6 for any $\alpha \in (1, S_d)$.

Corollary 4.4 shows that the *hot spots* conjecture is, in an L^∞ sense, *uniformly false* as $d \rightarrow \infty$. If instead one uses a measure-theoretic sense (or, as a corollary, an L^p sense, $0 < p < \infty$), the *hot spots* conjecture becomes more and more true as the dimension increases:

Corollary 4.7. *For any $\alpha > 1$, one has $\lim_{d \rightarrow \infty} V_d(\alpha) = 0$. Moreover, the convergence is exponentially fast.*

The construction in [348] of a convex set that violates the *hot spots* conjecture is, in all but two dimensions, radially symmetric. Moreover, if in [348, Definition 2.7] one could choose $\psi_{\Omega, V}$ with enough freedom, one would likely be able to reach a $\sqrt{e}$ bound in [348, Proposition 2.8]. This suggests the following conjecture.

Conjecture 4.8. Let C_d be the *hot spots ratio* of d -dimensional convex sets, that is, the supremum among all convex, bounded domains $\Omega \subset \mathbb{R}^d$ of the ratio

$$\frac{\max_{x \in \Omega} \psi_\Omega(x)}{\max_{x \in \partial\Omega} \psi_\Omega(x)}.$$

Then

$$\lim_{d \rightarrow \infty} C_d = \sqrt{e}.$$

Remark 4.9. Our construction shares some overarching similarities with the counterexample to Payne’s nodal line conjecture in [361] ($d = 2$) or [362] ($d \geq 3$), with more recent constructions in [363], as it involves removing multiple holes from a ball, resulting in a topologically complex domain. In $d \geq 3$, a harmonic capacity argument allows one to connect the holes to the boundary of the ball, giving rise to counterexamples that are homeomorphic to a ball [364]. An analogous phenomenon holds for the hot spots ratio in $d \geq 3$ (but not for $d = 2$, where the hot spots conjecture is still widely expected to hold for simply connected domains). $\diamond$

Organization of the paper: Section 4.2 shows the upper bound in Theorems 4.3 and 4.6. Section 4.3 explains the heuristics for the lower bound and motivates an effective, limiting problem, which we carefully analyze in Section 4.4. Section 4.5 constructs the sequence of sets that witness the lower bound and reduces the analysis to the effective problem from Section 4.4. The sets saturating the hot spots ratio are constructed using a limiting procedure as a spherically symmetric Neumann sieve. In Section 4.6 we compute the asymptotics of S_d and $V_d(\alpha)$ as $d \rightarrow \infty$, completing the proofs of Corollaries 4.4 and 4.7.

LLM usage disclosure: The sharp result arises from two independent proofs, one providing a lower bound through a sequence of examples, one providing an upper bound through a rearrangement proof. The proof of the upper bound was found with significant assistance of a large language model (GPT o1-preview). The example witnessing the lower bound was originally found through a computational search, and then interpreted in the proof. The initial code for the lower bound example was mostly produced by GPT4o and Gemini. The proofs appearing in the manuscript were all human-written.

Acknowledgments: This work was initiated when all three authors were in residence at the Mathematisches Forschungsinstitut Oberwolfach during the fall of 2024, participating in the Arbeitsgemeinschaft “Quantum Signal Processing and Nonlinear Fourier Analysis”. JD was funded by the Simons Collaborations in MPS grant 563916. AH was supported by National Science Foundation grants DMS-2208535 and DMS-2337678 with travel support to Oberwolfach provided by DMS-2230648.

4.2 The upper bound

The upper bound is a refined version of the argument in [341, 342]. We substitute estimates on stopping-time probabilities for the Brownian motion with a combination of Talenti’s rearrangement inequality and the Szégo-Weinberger bound on the first eigenvalue.

Definition 4.10. Given two nonnegative functions f, g with domains of definition Ω_f, Ω_g such that $|\Omega_f| = |\Omega_g|$, we write $f^\#$ to denote the symmetric decreasing rearrangement of f , and $f \leq_\# g$ to denote that $f^\# \leq g^\#$, or, equivalently, that for all $\alpha > 0$, one has $|\{f > \alpha\}| \leq |\{g > \alpha\}|$. In particular, $f \leq g$ implies $f \leq_\# g$.

As an intermediate step in the proof, we compare ψ_Ω with the following function.

Definition 4.11. For a bounded domain $\Omega \subset \mathbb{R}^d$ and any number $0 \leq \mu < \lambda_\Omega$, let $u_{\mu,\Omega}$ be the unique solution to

$$\begin{cases} -\Delta u_{\mu,\Omega}(x) = \mu u_{\mu,\Omega}(x) & \text{for } x \text{ in } \Omega, \\ u_{\mu,\Omega}(x) = 1 & \text{for } x \text{ in } \partial\Omega. \end{cases} \quad (4.1)$$

Existence and uniqueness for (4.1) follow from the fact that for $\mu < \lambda_\Omega$, the operator $-\Delta - \mu$ is positive definite in $H_{\text{Dirichlet}}^1(\Omega)$. With the above definitions at hand, the upper bounds in Theorems 4.3 and 4.6 will follow from showing the chain of inequalities

$$\begin{aligned} \frac{\psi_\Omega}{\max_{x \in \partial\Omega} \psi_\Omega(x)} &\stackrel{(4.2.A)}{\leq} u_{\mu_\Omega, \Omega} \stackrel{(4.2.B)}{\leq_\#} u_{\mu_\Omega, B_1(\mathbb{R}^d)} \\ &\stackrel{(4.2.C)}{\leq} u_{\mu_{B_1(\mathbb{R}^d)}, B_1(\mathbb{R}^d)} = \eta_d. \end{aligned} \quad (4.2)$$

The rest of this section is devoted to the proof of these three inequalities. First, we outline the basic strategy.

(4.2.A) The first inequality in (4.2) follows from a comparison principle (Proposition 4.12).
Let

$$\tilde{\psi}_\Omega := \frac{\psi_\Omega}{\max_{x \in \partial\Omega} \psi_\Omega(x)},$$

so that for all $x \in \partial\Omega$ one has $\tilde{\psi}_\Omega(x) \leq 1 = u_{\mu_\Omega, \Omega}(x)$. Since both $\tilde{\psi}_\Omega$ and $u_{\mu_\Omega, \Omega}$ satisfy the PDE $(\Delta + \mu_\Omega)(\cdot) = 0$, the inequality can be extended to the interior.

(4.2.B) The second inequality is an application of Talenti's inequality (Proposition 4.13).

(4.2.C) The function $(\mu, x) \mapsto u_{\mu, B_1(\mathbb{R}^d)}(x)$ is an analytic nonnegative function in $[0, \lambda_{B_1(\mathbb{R}^d)}) \times B_1(\mathbb{R}^d)$ (Proposition 4.14) satisfying $\partial_\mu u_{\mu, B_1(\mathbb{R}^d)}(x) > 0$ on the interior of its domain. This implies the third inequality.

Moreover, if the third step (4.2.C) is an almost-equality at $x = 0$ (within a difference of ϵ^2) it must be that $\mu_{B_1(\mathbb{R}^d)} - \mu_\Omega \leq C_d \epsilon^2$ as well, so by the stability of the Szégo-Weinberger inequality ([343, Theorem 4.1]) the Fraenkel asymmetry of Ω will be $\lesssim \epsilon$. If equality were to be achieved by a fixed Lipschitz domain Ω , this would imply that $\mathcal{A}(\Omega) = 0$ and $\Omega = B_1(\mathbb{R}^d)$. However, the maximum of $\psi_{B_1(\mathbb{R}^d)}$ appears on the boundary, implying that $\tilde{\psi}_{B_1(\mathbb{R}^d)} = 1 < S_d$ and yielding a contradiction.

We begin the proof by recalling a standard comparison principle (see, e.g., [365]).

Proposition 4.12 (Comparison principle). *Let Ω be a bounded domain in $\mathbb{R}^d$, with first Dirichlet eigenvalue λ_Ω . Let $\mu < \lambda_\Omega$ and let $u \in C^2(\Omega)$ be a function satisfying*

$$\begin{cases} -\Delta u \geq \mu u & \text{in } \Omega^\circ, \\ u \geq 0 & \text{on } \partial\Omega. \end{cases}$$

Then $u \geq 0$ in Ω° .

Proof. Let $w := -u \cdot \mathbb{1}_{u \leq 0}$, a positive function in $H_0^1(\Omega)$ satisfying $-\Delta w \leq \mu w$ on its support. By a formal computation, we may compute that

$$\int_{\{w \geq 0\}} |\nabla w|^2 dx = - \int_{\{w \geq 0\}} w \Delta w dx \leq \mu \int_{\{w \geq 0\}} w^2 dx.$$

If $w \neq 0$, by the Rayleigh quotient characterization of the first eigenvalue, we would have $\lambda_\Omega \leq \mu$, contradicting the hypothesis.

The above *integration by parts* equality requires some regularity of the boundary of the set $\{u \leq 0\}$. This can be achieved, for example, by an approximation argument considering the sets $\{u \leq -\epsilon\}$. By Sard's theorem, the sets $\{u \leq -\epsilon\}$ are generically C^1 , and for these generic ϵ one does have

$$\int_{\{w \geq \epsilon\}} |\nabla w|^2 dx = - \int_{\{w - \epsilon \geq 0\}} (w - \epsilon)_+ \Delta w dx.$$

□

Proposition 4.13. *For any $\mu \leq \mu_{B_1(\mathbb{R}^d)}$ and any Lipschitz domain Ω in $\mathbb{R}^d$ such that $|\Omega| = |B_1(\mathbb{R}^d)|$, we have*

$$u_{\mu,\Omega} \leq_{\#} u_{\mu,B_1(\mathbb{R}^d)}.$$

Proof. Let $T_{\mu,\Omega}f$ be the solution operator given by the PDE

$$\begin{cases} -\Delta T_{\mu,\Omega}(f) = \mu f & \text{in } \Omega, \\ T_{\mu,\Omega}(f) = 1 & \text{on } \partial\Omega. \end{cases}$$

This operator is a contraction in $L^2(\Omega)$ as long as $\mu < \lambda_{\Omega}$ because the operator $\frac{-1}{\mu}\Delta$ with zero Dirichlet boundary conditions has smallest eigenvalue $\frac{\lambda_{\Omega}}{\mu} > 1$. Therefore, it has as a unique fixed point $u_{\mu,\Omega}$. By elliptic regularity, the functions $T_{\mu,\Omega}^n 1$ will converge uniformly to this fixed point $u_{\mu,\Omega}$. It suffices to show, by induction, that for all $n \geq 0$ one has the relation $T_{\mu,\Omega}^n 1 \leq_{\#} T_{\mu,B_1(\mathbb{R}^d)}^n 1$.

- The case $n = 0$ follows from $1^{\#} = 1$.
- For the induction step, we use Talenti's rearrangement inequality. Talenti's inequality implies that $T_{\mu,\Omega}f \leq_{\#} T_{\mu,B_1(\mathbb{R}^d)}f^{\#}$. On the other hand, the maximum principle for Laplace supersolutions implies that whenever $f^{\#} \leq g^{\#}$ one has $T_{\mu,B_1(\mathbb{R}^d)}f^{\#} \leq T_{\mu,B_1(\mathbb{R}^d)}g^{\#}$. By combining both of these observations with the induction hypothesis, we may conclude that

$$T_{\mu,\Omega}(T_{\mu,\Omega}^n 1) \leq_{\#} T_{\mu,B_1(\mathbb{R}^d)}((T_{\mu,\Omega}^n 1)^{\#}) \leq T_{\mu,B_1(\mathbb{R}^d)}(T_{\mu,B_1(\mathbb{R}^d)}^n 1).$$

□

Proposition 4.14. *Let $v(\mu, x) := u_{\mu,B_1(\mathbb{R}^d)}(x)$, with domain $(0, \lambda_{B_1(\mathbb{R}^d)}) \times B_1(\mathbb{R}^d)$. Then v is an analytic function and*

$$v(\mu, x) = \frac{\varphi_{B_1(\mathbb{R}^d)}\left(|x|\sqrt{\mu/\lambda_{B_1(\mathbb{R}^d)}}\right)}{\varphi_{B_1(\mathbb{R}^d)}\left(\sqrt{\mu/\lambda_{B_1(\mathbb{R}^d)}}\right)}.$$

Moreover, $v(\mu, x) \geq 1$ and $\partial_{\mu}v(\mu, x) > 0$ in $(0, \lambda_{B_1(\mathbb{R}^d)}) \times (B_1(\mathbb{R}^d))^{\circ}$.

Proof. The function $\varphi_{B_1(\mathbb{R}^d)}(x)$ is a strictly positive radially decreasing function in $B_1(\mathbb{R}^d)$ satisfying the equation $-\Delta\varphi_{B_1(\mathbb{R}^d)} = \lambda_{B_1(\mathbb{R}^d)}\varphi_{B_1(\mathbb{R}^d)}$. This simultaneously proves that $v(\mu, x)$ satisfies the PDE that defines $u_{\mu,B_1(\mathbb{R}^d)}(x)$ and the fact that $v \geq 1$.

One can take derivatives of the PDE $(-\Delta_x - \mu)v(\mu, x) = 0$ to see that $\partial_{\mu}v(\mu, x)$ satisfies the PDE

$$\begin{cases} (-\Delta_x - \mu)\partial_{\mu}v = v \geq 1, \\ \partial_{\mu}v|_{\partial B_1} = 0. \end{cases}$$

In particular, for some small constant $\epsilon = \epsilon_d$ the function $w = \partial_\mu v - \epsilon(1 - |x|^2)$ satisfies the differential inequalities

$$\begin{cases} (-\Delta_x - \mu)w \geq 0, \\ w|_{\partial B_1} = 0. \end{cases}$$

By Proposition 4.12, this shows that $w \geq 0$, and therefore that $\partial_\mu v \geq \epsilon(1 - |x|^2)$. $\square$

4.3 The lower bound construction

If one believes that the upper bounds in Section 4.2 are sharp, the stability results imply that extremizing sequences to this sharp bound must, in a certain sense, converge to a ball. The ball, however, is not a maximizer: the hot spots ratio for the ball is 1. The main obstruction in the case of the ball is that the first Neumann eigenfunction is not radial, while the upper bounds in Section 4.2 were all achieved by radially symmetric functions. In order to produce a *radially symmetric* first eigenfunction, one *weakly disconnects* the ball into an *inner ball* and an *outer annulus* by removing parts of the domain between both of them (see Figure 4.2). It is then energetically favorable for the first eigenfunction to *greatly change* across this weakly connected region. This favors radially symmetric eigenfunctions that are smooth on either side of the weakly connected region and that *jump* across this region.

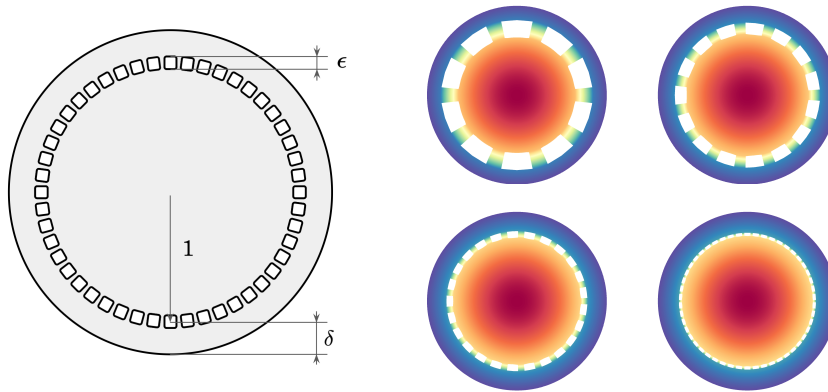


Figure 4.2: (Left) Sketch of the Neumann sieve domains. (Right) First Neumann Laplace eigenfunction for a sequence of sieve domains. In the limit when the number of holes goes to infinity one recovers a radially symmetric *effective problem*. The effective problem is studied in Section 4.4, and the fact that the convergence holds is studied in Section 4.5.

The process of obtaining an effective problem by adding small occlusions to a domain and sending the size of the occlusions to zero in a suitable way is known as a *Neumann sieve* construction, and it has been extensively studied (see, e.g., [366–368] and [369] for a modern treatment including a brief history of the problem). When the size and separation of the

occlusions is sent to zero at the right rate, the limiting *effective problem* has a new Robin boundary condition on a hypersurface, with a Robin parameter that one can tune. The usage of Neumann sieves to build counterexamples to the *hot spots* conjecture was briefly suggested by Jerison and Nadirashvili in [354].

The sharp lower bound for the hot spots constant comes from carefully balancing the *connectivity* between the interior and exterior regions: If the two regions are essentially disconnected, the first eigenvalue will be almost zero. Hence, the first eigenfunction will be almost constant in the inner circle and almost constant in the outer annulus, giving a hot spots ratio of almost 1. If the regions are strongly connected, the most energetically favorable eigenfunction will not be radial (as happens for the ball, where there is no disconnection at all).

Balancing this connectivity parameter gives the sharp hot spots ratio.

For this reason, in our setup we will consider not only a single Neumann sieve, but a family of sieves depending on parameters δ (the thickness of the outer annulus) and β (related to the connectivity strength of the sieve). The analysis proceeds as follows:

1. First, we send the number of holes in Figure 4.2 to infinity and the size of the holes to zero at the right rate, in order to obtain an *effective problem* depending on two parameters: the distance δ of the sieve to the boundary, and a parameter β that characterizes how *connected* the two sides of the Neumann sieve are to each other. This convergence is, in the appropriate sense, uniform, and one can *transfer* lower bounds from this effective hot spots problem to the original hot spots problem.
2. The hot spots ratio of the examples arising in the effective problem can be computed explicitly. By optimizing over these parameters, (which involves sending δ to zero and β to $\mu(B_1(\mathbb{R}^d))$) one achieves the desired lower bound for S_d . Notice that sending $\delta \rightarrow 0$ in these radial sieve domains makes them closer and closer to a ball, which is consistent with the *equality* case in the proofs in Section 4.2 being sharp for a ball.

4.3.1 Convergence to an effective problem

The behavior seen in Figure 4.2 as the number of holes goes to infinity is governed by an *effective bilinear form* $D_{\beta,\delta}$ which can be written as

$$D_{\beta,\delta}(f, f) = \int_{B_1(\mathbb{R}^d)} |\nabla(f)|^2 dx + \int_{B_{1+\delta}(\mathbb{R}^d) \setminus B_1(\mathbb{R}^d)} |\nabla(f)|^2 dx + \beta \delta \oint_{\mathbb{S}^{d-1}} |f(1^+e) - f(1^-e)|^2 de, \quad (4.3)$$

with domain $H^1(B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1})$. Here, the notation $f(1^+e)$ (resp. $f(1^-e)$) denotes the *outer* (resp. *inner*) H^1 trace onto the boundary $\mathbb{S}^{d-1}$. The main purpose of Section 4.5 is to prove the following approximation result.

Proposition 4.15. *Let $\beta, \delta > 0$ be such that $\psi_{\beta,\delta}^{(1)}$, the first non-constant eigenfunction of $D_{\beta,\delta}$, is radially symmetric. Then there exists a sequence of Neumann sieve domains Ω_n in $\mathbb{R}^d$ with first Neumann Laplace eigenfunctions ψ_n such that:*

1. The domains Ω_n converge to a ball in the Hausdorff metric.
2. The boundaries $\partial\Omega_n$ converge to $\mathbb{S}^{d-1} \sqcup (1+\delta)\mathbb{S}^{d-1}$.
3. There exists a choice of signs $\sigma_n \in \{\pm 1\}$ (which we will without loss of generality assume to be 1) such that

- (a) $\sigma_n \psi_n \rightarrow \psi_{\beta,\delta}^{(1)}$ locally uniformly in $\overline{B_{1+\delta}(\mathbb{R}^d)} \setminus \mathbb{S}^{d-1}$.
- (b) $\limsup_{n \rightarrow \infty} \max_{x \in \partial\Omega_n} \sigma_n \psi_n(x) = \max(\psi_{\beta,\delta}^{(1)}(1^-), \psi_{\beta,\delta}^{(1)}(1^+), \psi_{\beta,\delta}^{(1)}(1+\delta))$.

The Neumann sieve construction has been extensively studied in the literature and it is likely possible to deduce Proposition 4.15 from the proofs of known results. However, we were unable to find a result that implied this exact formulation, and therefore we provide a proof of Proposition 4.15 in Section 4.5 for completeness.

4.3.2 Optimizing over the effective problems

Proposition 4.15 reduces the problem of finding lower bounds for the hot spots constant to finding lower bounds for the *effective hot spots constant* associated to the operators $D_{\beta,\delta}^{(d)}$, defined as

$$S_d \geq \tilde{S}_d := \max_{\substack{\beta > 0 \\ \delta > 0 \\ \psi_{\beta,\delta}^{(1)} \text{ is radial}}} \frac{\max_{0 \leq r \leq 1} \psi_{\beta,\delta}^{(1)}(r)}{\max_{r \in \{1^-, 1^+, 1+\delta\}} \psi_{\beta,\delta}^{(1)}(r)},$$

where $\psi_{\beta,\delta}^{(1)}$ is taken with the sign normalization that makes it positive at the origin. In Section 4.4, we will optimize over the possible values of β, δ to obtain the lower bounds in Theorems 4.3 and 4.6. More precisely, we will prove the following proposition.

Proposition 4.16. *For any $d \geq 2$ and all $\beta < \mu_{B_1(\mathbb{R}^d)}$ there is a $\delta_0 := \delta_0(d, \beta)$ such that for all $0 < \delta < \delta_0$ the eigenfunction $\psi_{\beta,\delta}^{(1)}$ is radial. Moreover,*

$$\lim_{\beta \rightarrow \mu_{B_1(\mathbb{R}^d)}^-} \lim_{\delta \rightarrow 0} \frac{\psi_{\beta,\delta}^{(1)}(r)}{\max_{r \in \{1^-, 1^+, 1+\delta\}} \psi_{\beta,\delta}^{(1)}(r)} = \eta_d(r).$$

4.4 The effective problem

Throughout this section, we fix the value of $d \geq 2$. Our aim is to understand the first nontrivial eigenfunction of the effective problem in $L^2(B_{1+\delta}(\mathbb{R}^d))$, which is derived from the bilinear form (4.3). By applying the reductions in Section 4.5 and optimizing over the parameters $\beta, \delta > 0$, we will obtain the sharp lower bound for the hot spots ratio.

The bilinear form $D_{\beta,\delta}(f, g)$ induces a self-adjoint operator and is lower bounded on its domain by the Dirichlet energy. In particular, it has compact resolvent and discrete point spectrum, with eigenfunctions $\psi_{\beta,\delta}^{(0)}, \psi_{\beta,\delta}^{(1)}, \dots$, with $\psi_{\beta,\delta}^{(0)} \equiv 1$. Proposition 4.16 would follow if we were able to establish the following facts.

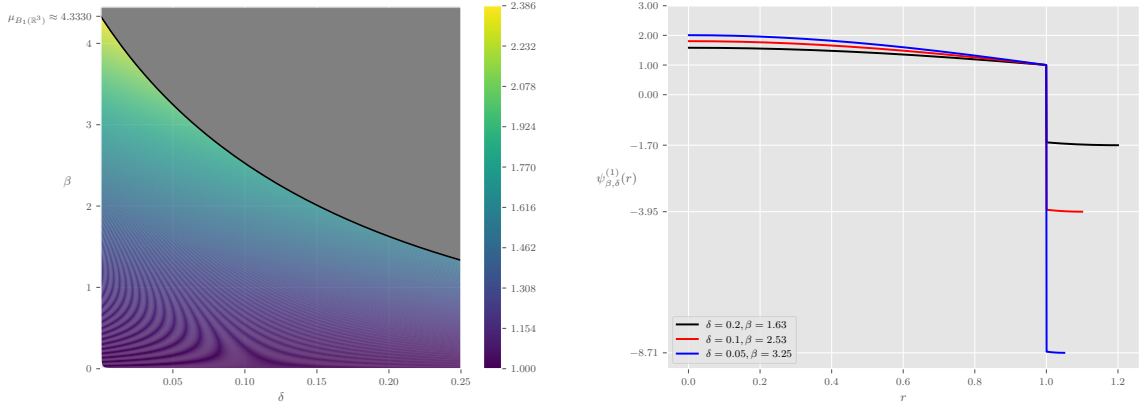


Figure 4.3: (Left) Hot spots ratio in dimension 3 as a function of β and δ . Note that the actual value of S_3 is approximately 2.3861. The black curve represents the boundary of the allowed region, above which the first non-trivial Neumann eigenfunction is not radial and the effective problem yields a hot spots constant equal to 1. The optimal asymptotic domain is achieved by taking $\delta \rightarrow 0$ and $\beta \rightarrow \mu_{B_1(\mathbb{R}^3)}$. (Right) Examples of the radial part of the eigenfunctions $\frac{\psi_{\beta,\delta}^{(1)}(r)}{\psi_{\beta,\delta}^{(1)}(1^-)}$, where β is taken as large as possible while keeping the first eigenfunction radial. This normalization sets the maximum on the boundary to 1, in contrast to the L^2 normalization. As we take $\delta \rightarrow 0$ and $\beta \rightarrow \mu_{B_1(\mathbb{R}^3)}$, $\frac{\psi_{\beta,\delta}^{(1)}(0)}{\psi_{\beta,\delta}^{(1)}(1^-)}$ will approach S_3 while $\frac{\psi_{\beta,\delta}^{(1)}(1^+)}{\psi_{\beta,\delta}^{(1)}(1^-)}$ will tend to $-\infty$ at a rate $\sim -\delta^{-1}$.

Proposition 4.17. *Let $0 < \beta < \mu_{B_1(\mathbb{R}^d)}$. Then as $\delta \rightarrow 0$, we have:*

1. *The eigenvalues $\mu_{\beta,\delta}^{(1)}$ converge to β .*
2. *The functions $\psi_{\beta,\delta}^{(1)}$ are radial for all $\delta > 0$ small enough (depending on β).*
3. *For $|x| < 1$ (and all $\delta > 0$ small enough) the functions $\psi_{\beta,\delta}^{(1)}$ are equal to*

$$\psi_{\beta,\delta}^{(1)}(x) = c_{\beta,\delta} \varphi_{B_1(\mathbb{R}^d)} \left(x \cdot \sqrt{\mu_{\beta,\delta}^{(1)} / \lambda_{B_1(\mathbb{R}^d)}} \right) \quad (4.4)$$

for some non-zero constants $c_{\beta,\delta}$.

4. *For $|x| > 1$ (and all $\delta > 0$ small enough) the functions $\psi_{\beta,\delta}^{(1)}$ are negative.*

Notice that statement 3. follows immediately from 2. and 4.. Since $\psi_{\beta,\delta}^{(1)}$ is the radial Laplace eigenfunction of eigenvalue $\mu_{\beta,\delta}^{(1)}$ for $|x| < 1$, the identity (4.4) follows (with $c_{\beta,\delta}$ potentially zero). The function $\varphi_{B_1(\mathbb{R}^d)} \left(x \cdot \sqrt{\mu_{\beta,\delta}^{(1)} / \lambda_{B_1(\mathbb{R}^d)}} \right)$ is nonnegative. Since $\psi_{\beta,\delta}^{(1)}$ is negative for $|x| > 1$ and it must have mean zero, $c_{\beta,\delta}$ must be positive.

4.4.1 Splitting eigenfunctions by spherical harmonics

The bilinear form $D_{\beta,\delta}$ is invariant under rotations. In particular, there is a basis of eigenfunctions that splits (in polar coordinates) into radial functions times a spherical harmonic. The possible eigenvalues of the spherical harmonics are $\ell(\ell + d - 2)$, for integers $\ell \geq 0$. For $\beta, \delta \geq 0$, this splits the problem into a family of radial problems in one dimension, in $L^2([0, 1 + \delta], x^{d-1}dx)$, with associated forms

$$H_{\ell,\beta,\delta}(f, f) = H_{\beta,\delta}^0(f, f) + \ell(\ell + d - 2) \int_0^{1+\delta} |f(x)|^2 x^{d-3} dx, \quad \ell > 0,$$

where

$$H_{\beta,\delta}^0(f, f) = \int_0^1 |f'(x)|^2 x^{d-1} dx + \beta\delta |f(1^-) - f(1^+)|^2 + \int_1^{1+\delta} |f'(x)|^2 x^{d-1} dx.$$

Note that $H_{\beta,\delta}^0 = H_{0,\beta,\delta}$. When $\delta = 0$, with an abuse of notation, we will define $H_{\beta,0}^0(f, f) := \int_0^1 |f'(x)|^2 x^{d-1} dx$ and $H_{\ell,\beta,0}$ accordingly. Note that $H_{\ell,\beta,0}$ does not depend on β .

The eigenfunctions of $D_{\beta,\delta}$ are the products of a spherical harmonic of eigenvalue $\ell(\ell + d - 2)$ (in the polar coordinate) and an eigenfunction of $H_{\ell,\beta,\delta}$ (in the radial coordinate), with eigenvalue the corresponding eigenvalue of $H_{\ell,\beta,\delta}$. We will denote by $h_{\ell,\beta,\delta}^{(k)}$ the k -th eigenvalue of $H_{\ell,\beta,\delta}$. The eigenvalues of $D_{\beta,\delta}$ are then the (ordered) union of the eigenvalues of $H_{\ell,\beta,\delta}$ over all ℓ , and the remaining question is which of the $H_{\ell,\beta,\delta}$ gives rise to the first nontrivial eigenvalue of $D_{\beta,\delta}$.

The constant function is a radial eigenfunction, showing that $\mu_{\beta,\delta}^{(0)} = h_{0,\beta,\delta}^{(0)} = 0$. Since $H_{\ell,\beta,\delta} \preceq H_{\ell+1,\beta,\delta}$ in the positive semidefinite ordering, we have $h_{1,\beta,\delta}^{(0)} \leq h_{\ell,\beta,\delta}^{(0)}$ for all $\ell \geq 1$. In particular, there are only two possibilities for $\mu_{\beta,\delta}^{(1)}$.

- One possibility is that $\mu_{\beta,\delta}^{(1)} = h_{0,\beta,\delta}^{(1)}$, i.e., the first non-zero eigenvalue of $H_{0,\beta,\delta}(f, f)$ is the first non-zero eigenvalue of $D_{\beta,\delta}$. In this case, the first eigenfunction of $D_{\beta,\delta}$ is radial.
- The alternative possibility is that $\mu_{\beta,\delta}^{(1)} = h_{1,\beta,\delta}^{(0)}$. In this case, the first nontrivial eigenfunction of $D_{\beta,\delta}$ is not radially symmetric. Experiments show that the hot spots ratio is equal to 1 in this case, and therefore it is beyond our interest.

Proving Proposition 4.17 entails showing that we are in the first (radial) case. This will happen as long as $h_{0,\beta,\delta}^{(1)} \leq h_{1,\beta,\delta}^{(0)}$. We will first fix $\beta < \mu_{B_1}(\mathbb{R}^d)$ and send $\delta \rightarrow 0$.

4.4.2 Taking the limit $\delta \rightarrow 0$

The splitting into spherical harmonics reduces Proposition 4.17 to the following:

Proposition 4.18. *Fix $0 < \beta < \mu_{B_1}(\mathbb{R}^d)$. The following statements hold:*

1. We have

$$\lim_{\delta \rightarrow 0} h_{0,\beta,\delta}^{(1)} = \beta < \mu_{B_1(\mathbb{R}^d)} = \lim_{\delta \rightarrow 0} h_{1,\beta,\delta}^{(0)}.$$

In particular, for $\delta > 0$ small enough (depending on β) one has $h_{0,\beta,\delta}^{(1)} < h_{1,\beta,\delta}^{(0)}$.

2. For $\delta > 0$ small enough, the eigenfunction corresponding to the eigenvalue $h_{0,\beta,\delta}^{(1)}$ has constant (and opposite) signs in $[0, 1]$ and in $(1, 1 + \delta]$.

To understand the scheme of the proof of Proposition 4.18, we emphasize the following intuition coming from the simulations in Figure 4.3:

1. The regions $r < 1$ and $r \in (1, 1 + \delta)$ essentially decouple from each other.
2. For $r \in [1, 1 + \delta)$, the first eigenfunction takes the value $\approx -\delta^{-1/2}$. In particular, most of the L^2 mass concentrates in this region.
3. For $r \in [0, 1)$, the first eigenfunction is still proportional to the radial part of a Laplace eigenfunction, with proportionality constant going to zero at a rate of $\sim \delta^{1/2}$.

This motivates a change of variables (in the form of an L^2 isometry) that *focuses* on the $(1, 1 + \delta)$ region.

Proof of Proposition 4.18. We define the isometry

$$\Phi : L^2([0, 1 + \delta], r^{d-1}dr) \rightarrow L^2([0, 1], r^{d-1}dr) \oplus L^2([0, 1], dx)$$

by

$$\Phi : f \mapsto \left(f|_{[0,1]}, C_\delta \delta^{1/2} f(T_\delta(x)) \right), \quad \Phi^{-1}(f, g)(r) = \begin{cases} f(r) & r \in [0, 1), \\ C_\delta^{-1} \delta^{-1/2} g(T_\delta^{-1}(r)) & r \in [1, 1 + \delta], \end{cases}$$

where

$$T_\delta(x) = (1 + d\delta C_\delta^2 x)^{1/d}, \quad C_\delta = \sqrt{\frac{(1 + \delta)^d - 1}{d\delta}}.$$

We note that $T_\delta(x) = 1 + \delta x + O(\delta^2)$ and $C_\delta = 1 + o(1)$. The above isometry breaks $[0, 1 + \delta]$ into an inner function and an outer function, and induces a new bilinear form $\tilde{H}_{\ell,\beta,\delta}(\Phi(\cdot), \Phi(\cdot)) = H_{\ell,\beta,\delta}(\cdot, \cdot)$ in $L^2([0, 1], r^{d-1}dr) \oplus L^2([0, 1], dx)$, given by:

$$\begin{aligned} \tilde{H}_{\ell,\beta,\delta}((f, g), (f, g)) = & H_{\ell,1,0}(f, f) + \beta |C_\delta \delta^{1/2} f(1) - g(0)|^2 \\ & + \ell(\ell + d - 2) \int_0^1 v_\delta(x) |g(x)|^2 dx \\ & + \delta^{-2} \int_0^1 w_\delta(x) |g'(x)|^2 dx, \end{aligned}$$

where

$$v_\delta(x) = \frac{T_\delta(x)^{d-3} T'_\delta(x)}{C_\delta^2 \delta}, \quad w_\delta(x) = \frac{\delta^2}{T'_\delta(x)^2}.$$

Notably, v_δ and w_δ converge in C^∞ to the constant function 1. For any sequence $\delta_n \rightarrow 0$ and any sequence (f_n, g_n) such that $\tilde{H}_{\ell, \beta, \delta_n}((f_n, g_n), (f_n, g_n))$ is uniformly bounded, we must have $\int_0^1 |g'_n(x)|^2 dx \rightarrow 0$, and thus $\|g_n - \bar{g}_n\|_\infty \rightarrow 0$, where $\bar{g}_n$ denotes the mean of g_n .

Let $(\phi_{\ell, \beta, \delta}^{(k)}, \gamma_{\ell, \beta, \delta}^{(k)})$ be the k -th L^2 -normalized eigenfunction associated to $\tilde{H}_{\ell, \beta, \delta}$ with eigenvalue $h_{\ell, \beta, \delta}^{(k)}$ (in other words, $(\phi_{\ell, \beta, \delta}^{(k)}, \gamma_{\ell, \beta, \delta}^{(k)}) = \Phi(\psi_{\ell, \beta, \delta}^{(k)})$, where $\psi_{\ell, \beta, \delta}^{(k)}$ is the corresponding eigenfunction of the original problem). Note that the index $k = 0, 1, \dots$ is for each fixed harmonic ℓ . The function $\psi_{\ell, \beta, \delta}^{(0)}$ will always be constant sign, and will be constant when $\ell = 0$.

Applying Courant-Fischer and restricting the outer portion of the trial function g to be constant (or even zero), we see that $h_{\ell, \beta, \delta}^{(k)} \leq h_{\ell, \beta, 0}^{(k)} + O(1)$. The functions $\gamma_{\ell, \beta, \delta}^{(k)}$ are also uniformly bounded in H^1 , and in particular have a convergent subsequence in $H^{1-\epsilon}$ and L^∞ by Sobolev embedding, so converge (at least subsequentially) to a constant. The only interaction between $\phi_{\ell, \beta, \delta}^{(k)}$ and $\gamma_{\ell, \beta, \delta}^{(k)}$ is given by $2\beta C_\delta \delta^{1/2} \cdot \gamma_{\ell, \beta, \delta}^{(k)}(0) \cdot \phi_{\ell, \beta, \delta}^{(k)}(1)$, which goes to zero. In particular, as $\delta \rightarrow 0$ the spectrum of each $H_{\ell, \beta, \delta}$ decouples into the union of the spectrum of $H_{\ell, \beta, 0}$ (which we recall is independent of β) and the single eigenvalue $\beta + \ell(\ell + d - 2)$ coming from γ being a constant function.

In the limit as $\delta \rightarrow 0$, the first non-zero eigenvalue of $D_{\beta, \delta}$ will come from the smallest of four possibilities: β , $h_{0, \beta, 0}^{(1)}$ (the two possibilities if the minimum is attained at $\ell = 0$), and $h_{1, \beta, 0}^{(0)}$, $\beta + (d - 1)$ (if the minimum is attained for $\ell = 1$, in the non-radial case). Hence, we see that as long as $\beta < h_{1, \beta, 0}^{(0)} = \mu_{B_1(\mathbb{R}^d)}$, the first eigenfunction will be in the subspace of radial functions for $\delta > 0$ small enough.

When $\ell = 0$, $(\phi_{0, \beta, \delta}^{(0)}, \gamma_{0, \beta, \delta}^{(0)}) = c \cdot \Phi(1) = c \cdot (1, C_\delta \sqrt{\delta})$, where c is a constant for L^2 normalization. We require orthogonality for other eigenfunctions, meaning that for $k \geq 1$,

$$\int_0^1 \phi_{0, \beta, \delta}^{(k)}(r) r^{d-1} dr + C_\delta \delta^{1/2} \int_0^1 \gamma_{0, \beta, \delta}^{(k)}(x) dx = 0,$$

which is equivalent to the mean zero condition for Neumann eigenfunctions.

The functions $\phi_{0, \beta, \delta}^{(k)}$ are the radial part of radially symmetric Laplace eigenfunctions. Taking $k = 1$, the uniqueness of radial solutions to the Helmholtz equation implies that $\phi_{0, \beta, \delta}^{(1)}$ must be equal to

$$c_{\beta, \delta} \varphi_{B_1(\mathbb{R}^d)} \left(x \cdot \sqrt{h_{0, \beta, \delta}^{(1)} / \lambda_{B_1(\mathbb{R}^d)}} \right).$$

Since $h_{0, \beta, \delta}^{(1)} < \mu_{B_1(\mathbb{R}^d)} < \lambda_{B_1(\mathbb{R}^d)}$ holds for small enough $\delta > 0$, the function $\phi_{0, \beta, \delta}^{(1)}$ will also be constant-sign (or exactly zero if $c_{\beta, \delta}$ was zero).

We now take the limit $\delta \rightarrow 0$. By the previous argument and up to subsequential limits, $\gamma_{0, \beta, \delta}^{(1)}$ must uniformly approach a constant $K_\beta := K_{0, \beta}^{(1)}$ (which we will assume to be ≤ 0 by the sign symmetry of eigenfunctions) and by the explicit representation of the interior part,

there exist constants M_β, Q_β such that

$$\int_0^1 \varphi_{B_1(\mathbb{R}^d)} \left(r \cdot \sqrt{h_{0,\beta,\delta}^{(1)} / \lambda_{B_1(\mathbb{R}^d)}} \right) r^{d-1} dr = M_\beta + o(1), \quad M_\beta > 0,$$

and

$$\int_0^1 \varphi_{B_1(\mathbb{R}^d)} \left(r \cdot \sqrt{h_{0,\beta,\delta}^{(1)} / \lambda_{B_1(\mathbb{R}^d)}} \right)^2 r^{d-1} dr = Q_\beta + o(1), \quad Q_\beta > 0.$$

From orthogonality and the L^2 normalization, we have

$$c_{\beta,\delta}(M_\beta + o(1)) + \delta^{1/2}(K_\beta + o(1)) = 0, \quad c_{\beta,\delta}^2(Q_\beta + o(1)) + (K_\beta + o(1))^2 = 1.$$

Solving these equations, we obtain

$$c_{\beta,\delta} = -\delta^{1/2} \frac{K_\beta}{M_\beta} (1 + o(1)) \quad \delta \frac{K_\beta^2}{M_\beta^2} Q_\beta + K_\beta^2 = 1 + o(1),$$

and deduce that $K_\beta = -1$ and $c_{\beta,\delta} > 0$ for $\delta > 0$ small enough, with limits independent of the subsequence that we take. This establishes the second point in Proposition 4.18. $\square$

Remark 4.19. The effective radial problems can be analyzed more explicitly as Sturm-Liouville eigenvalue problems with an interior jump condition. The Euler-Lagrange equations associated to

$$H_{\ell,\beta,\delta}(f, f) - \mu \int_0^{1+\delta} f(r)^2 r^{d-1} dr$$

are

$$\begin{aligned} f'(1^-) &= \beta\delta (f(1^+) - f(1^-)), \\ f'(1^-) &= f'(1^+), \\ f'(1+\delta) &= 0, \\ -\frac{d}{dr} \left(r^{d-1} f'(r) \right) + \ell(\ell + d - 2) f(r) r^{d-3} &= h_{\ell,\beta,\delta}^{(k)} r^{d-1} f(r), \quad r \in (0, 1) \cup (1, 1+\delta). \end{aligned}$$

This system of equations can be transformed into a Bessel equation, and by requiring that $f(r) = O(1)$ as $r \rightarrow 0$, we see that, up to scaling,

$$f(r) = \begin{cases} r^{1-\frac{d}{2}} J_{\frac{d}{2}+\ell-1}(r \sqrt{h_{\ell,\beta,\delta}^{(k)}}) & r < 1, \\ ar^{1-\frac{d}{2}} J_{\frac{d}{2}+\ell-1}(r \sqrt{h_{\ell,\beta,\delta}^{(k)}}) + br^{1-\frac{d}{2}} Y_{\frac{d}{2}+\ell-1}(r \sqrt{h_{\ell,\beta,\delta}^{(k)}}) & r > 1, \end{cases}$$

where the coefficients a, b are chosen to match the interior jump and right boundary conditions. This alternative strategy was used in the numerical experiments to determine optimal β, δ configurations and to produce Figure 4.3. An analysis of the asymptotics of these problems as $\delta \rightarrow 0$ leads to an alternative proof of Proposition 4.18. $\diamond$

4.5 The Neumann sieve problem

In this section, we construct a sequence of domains that saturates the upper bound for the hot spots constant from Section 4.2. To build such domains, we use a variant of the “Neumann sieve” construction. Although Neumann sieve methods have been employed several times in the literature to achieve different purposes, they are usually presented in planar geometries (which are not appropriate for us) or in very general scenarios.

For our purposes, a simpler, direct version of the Neumann sieve will suffice. Compared to the sieves considered in the literature, our construction is *thicker*, in the sense that the length of the channels is much larger than their size (see Figure 4.4). This greatly simplifies the homogenization computations.

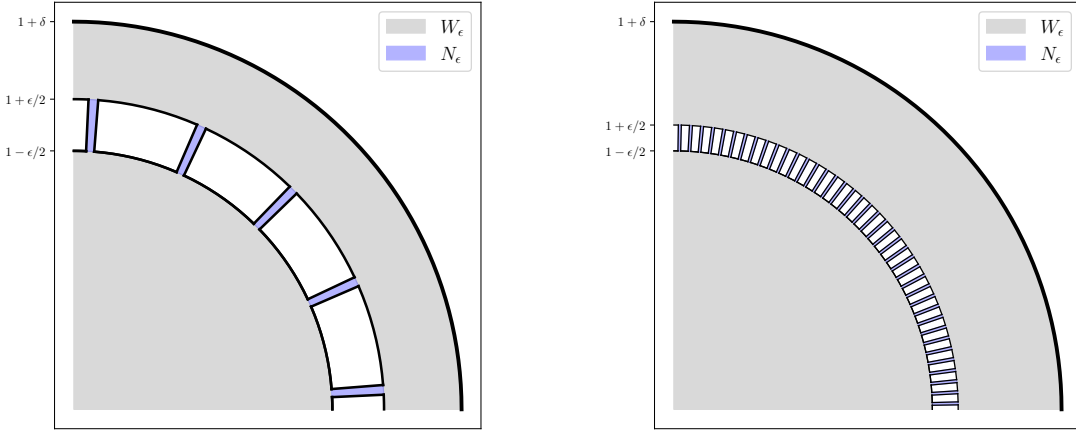


Figure 4.4: The domain Ω_ϵ consists of $B_{1+\delta} := B_{1+\delta}(\mathbb{R}^d)$ with the region $\{x \in B_{1+\delta} \text{ s.t. } |x| \in (1 - \epsilon/2, 1 + \epsilon/2)\}$ removed along with thin channels through that region. We refer to the bulk as W_ϵ and the channels as N_ϵ . The width of the channels must be much smaller than their length, ϵ , being that they contain only a proportion $\epsilon^2\alpha$ of the volume in the shell while being equidistributed on $\mathbb{S}^{d-1}$ at scales $\sim \exp(-\frac{1}{\epsilon})$. In the case $d = 2$, this will consist of exponentially many channels that are getting exponentially narrower as $\epsilon \rightarrow 0$.

Throughout this section, we will fix $d \geq 2$ and $\beta, \delta > 0$ such that the first nontrivial eigenfunction of $D_{\beta, \delta}$ is unique and radial (and in particular the second eigenvalue of $D_{\beta, \delta}$ is strictly larger than the first). We will drop the dependencies on these parameters and let $\alpha := \beta\delta$ be the *connectivity parameter* of the Neumann sieve, i.e., the factor in front of the last summand in (4.3).

Definition 4.20 (Spherical sieve). Let $0 < \epsilon < \alpha^{-1/2}$. A set $S_\epsilon \subset \mathbb{S}^{d-1}$ with smooth boundary is an ϵ -spherical sieve if for any ball $B_{e^{-1/\epsilon}}(x_0)$ of radius $e^{-1/\epsilon}$ in $\mathbb{S}^{d-1}$, we have

that

$$\left| \epsilon^{-2} \frac{|S_\epsilon \cap B_{e^{-1/\epsilon}}(x_0)|}{|B_{e^{-1/\epsilon}}(x_0)|} - \alpha \right| \leq e^{-1/\epsilon}. \quad (4.5)$$

In other words, ϵ -spherical sieves contain $\epsilon^2\alpha$ of the volume of $\mathbb{S}^{d-1}$ and are uniformly distributed in $\mathbb{S}^{d-1}$ up to exponentially small scales $\exp(-1/\epsilon)$ with similarly exponentially small error. Note that a probabilistic argument (splitting the ball into roughly $\exp(-2/\epsilon)$ pieces, keeping each of them with probability $\epsilon^2\alpha$, and smoothing the result) guarantees the existence of ϵ -spherical sieves.

Definition 4.21 (Thick sieve). A set $\Omega_\epsilon \subset B_{1+\delta}$, with $0 < \epsilon < \delta$, is an ϵ -thick sieve if

1. Ω_ϵ is a closed domain with smooth boundary.
2. $B_{1+\delta} \supset \Omega_\epsilon \supset \{x \in B_{1+\delta}(\mathbb{R}^d) \text{ s.t. } |x| \notin [1 - \epsilon/2, 1 + \epsilon/2]\}$.
3. The domain $B_{1+\delta} \setminus \Omega_\epsilon$ is *graphical* over $\mathbb{S}^{d-1}$: For every $e \in \mathbb{S}^{d-1}$, the set $\{r \geq 0 \text{ s.t. } r \cdot e \in B_{1+\delta} \setminus \Omega_\epsilon\}$ is an interval.
4. There exists a spherical sieve S_ϵ approximating Ω_ϵ , in the sense that

$$\{x \in \Omega_\epsilon \text{ s.t. } |x| \in [1 - \epsilon/2 + \epsilon^2, 1 + \epsilon/2 - \epsilon^2]\} = \{e \cdot r \text{ s.t. } e \in S_\epsilon, r \in [1 - \epsilon/2 + \epsilon^2, 1 + \epsilon/2 - \epsilon^2]\}.$$

The set $\{x \in \Omega_\epsilon \text{ s.t. } |x| \in [1 - \epsilon/2, 1 + \epsilon/2]\}$ will be denoted by N_ϵ (the “necks” of the sieve) and its complement (which contains all but $O(\epsilon)$ of the mass in Ω_ϵ) will be denoted by W_ϵ .

Thick sieves can be constructed from spherical sieves using an approximation argument by smoothing the set

$$\{x \in B_{1+\delta} \text{ s.t. } |x| \notin [1 - \epsilon/2, 1 + \epsilon/2] \text{ or } x/|x| \in S_\epsilon\}.$$

4.5.1 Estimating the eigenvalues of Ω_ϵ

We now show that the eigenvalues of Ω_ϵ approach the eigenvalues of $D_{\beta,\delta}$ as ϵ goes to zero, at a rate that depends only on β , δ and d .

Proposition 4.22. *We have $\lim_{\epsilon \rightarrow 0} \mu_{\Omega_\epsilon}^{(k)} = \mu_{\beta,\delta}^{(k)}$, uniformly over the possible choices of thick Neumann sieves Ω_ϵ , with a rate that depends only on β, δ, d, k .*

In order to show that this is the case, we will construct two operators, RI and RE, which approximately conjugate $-\Delta$ in Ω_ϵ and $D_{\beta,\delta}$ to each other.

First, we define the linear interpolation operator $\text{RI} : L^2(B_{1+\delta}(\mathbb{R}^d)) \cap H^1(B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1}) \rightarrow H^1(\Omega_\epsilon)$ as

$$\text{RI} g(x) := \begin{cases} g(x) & \text{if } x \in W_\epsilon, \\ \frac{(1+\epsilon/2-|x|)}{\epsilon} g\left(\left(1 - \frac{\epsilon}{2}\right) \frac{x}{|x|}\right) + \frac{(|x|-1+\epsilon/2)}{\epsilon} g\left(\left(1 + \frac{\epsilon}{2}\right) \frac{x}{|x|}\right) & \text{if } x \in N_\epsilon. \end{cases}$$

In other words, the RI operator *substitutes* the value at the neck N_ϵ by the radial linear interpolation of g .

Analogously, we can *remove* the neck and create a function with domain the whole ball of radius $1 + \delta$ by *stretching* the value. We define the removal operator $\text{RE} : H^1(\Omega_\epsilon) \rightarrow L^2(B_{1+\delta}(\mathbb{R}^d)) \cap H^1(B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1})$ as

$$\text{RE } g(x) := \begin{cases} g((1 - \epsilon/2)x) & \text{if } |x| < 1, \\ g\left(\frac{x}{|x|} \left(1 + \frac{\epsilon}{2} + (|x| - 1)(1 - \delta^{-1}\epsilon/2)\right)\right) & \text{if } |x| > 1. \end{cases}$$

This operator *removes* the data at the neck, mapping the ball $B_{1-\epsilon/2}(\mathbb{R}^d)$ to $B_1(\mathbb{R}^d)$ and the shell $(B_{1+\delta}(\mathbb{R}^d) - B_{1+\epsilon/2}(\mathbb{R}^d))$ to $(B_{1+\delta}(\mathbb{R}^d) - B_1(\mathbb{R}^d))$ by shifting and stretching the domains radially, combining the results, and throwing away the values of g for $|x| \in (1 - \epsilon/2, 1 + \epsilon/2)$.

The operators RI and RE are approximate isometries that approximately conjugate the Laplace operator to $D_{\beta,\delta}$ in the following sense:

Lemma 4.23. *As $\epsilon \rightarrow 0$, the following inequalities hold for any functions $f \in H^1(B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1})$ and $g \in H^1(\Omega_\epsilon)$:*

1. *Mass preservation:*

$$\begin{aligned} |\|\text{RI } f\|_{L^2(\Omega_\epsilon)} - \|f\|_{L^2(B_{1+\delta}(\mathbb{R}^d))}| &\leq C_{\beta,\delta,d}\epsilon \cdot (\|f\|_{L^2(B_{1+\delta}(\mathbb{R}^d))} + D_{\beta,\delta}(f, f)^{1/2}), \\ |\|g\|_{L^2(\Omega_\epsilon)} - \|\text{RE } g\|_{L^2(B_{1+\delta}(\mathbb{R}^d))}| &\leq C_{\beta,\delta,d}\epsilon \cdot (\|g\|_{L^2(\Omega_\epsilon)} + \|\nabla g\|_{L^2(\Omega_\epsilon)}). \end{aligned}$$

2. *Energy control:*

$$\begin{aligned} \|\nabla \text{RI } f\|_{L^2(\Omega_\epsilon)}^2 &\leq D_{\beta,\delta}(f, f) + C_{\beta,\delta,d}\epsilon \cdot (\|f\|_{L^2(B_{1+\delta}(\mathbb{R}^d))}^2 + D_{\beta,\delta}(f, f)), \\ D_{\beta,\delta}(\text{RE } g, \text{RE } g) &\leq \|\nabla g\|_{L^2(\Omega_\epsilon)}^2 + C_{\beta,\delta,d}\epsilon \cdot (\|g\|_{L^2(\Omega_\epsilon)}^2 + \|\nabla g\|_{L^2(\Omega_\epsilon)}^2). \end{aligned}$$

Lemma 4.23 contains all of the geometric information about Ω_ϵ that is needed for the proof of Proposition 4.22. Indeed, once Lemma 4.23 is established, Proposition 4.22 follows from a standard Courant-Fischer argument.

Proof of Proposition 4.22. Let $E_k := \text{span}\{\text{RI } \psi_{\beta,\delta}^{(i)}\}_{i=0}^k$ and fix $v \in E_k$ with norm 1. We may express $v = \text{RI} \left(\sum_{i=0}^k \alpha_i \psi_{\beta,\delta}^{(i)} \right)$. Norm-preservation (together with the fact that any linear combination of the first k eigenfunctions will have bounded energy) guarantees that $\sum_{i=0}^k \alpha_i^2 = 1 + o(1)$. Energy control then guarantees that

$$\|\nabla v\|_{L^2(\Omega_\epsilon)}^2 \leq \mu_{\beta,\delta}^{(k)} \sum_{i=0}^k \alpha_i^2 + o(1)(1 + \mu_{\beta,\delta}^{(k)}) \|v\|_{L^2(\Omega_\epsilon)}^2 \leq (\mu_{\beta,\delta}^{(k)} + o(1)) \|v\|_{L^2(\Omega_\epsilon)}^2.$$

In particular, $\mu_{\Omega_\epsilon}^{(k)} \leq \mu_{\beta,\delta}^{(k)} + o(1)$, where the $o(1)$ term goes to 0 as $\epsilon \rightarrow 0$ at a rate depending only on β, δ, d, k . The symmetric argument (replacing RI with RE) gives $\mu_{\beta,\delta}^{(k)} \leq \mu_{\Omega_\epsilon}^{(k)} + o(1)$ and therefore $\lim_{\epsilon \rightarrow 0} \mu_{\Omega_\epsilon}^{(k)} = \mu_{\beta,\delta}^{(k)}$, uniformly over all possible choices of thick Neumann sieves. Note that for the second argument one needs uniform upper bounds on the eigenvalues of $\mu_{\Omega_\epsilon}^{(k)}$, which are provided by the first argument. $\square$

4.5.2 Proof of the approximation inequalities

The main issue in proving Lemma 4.23 is that the size of the spherical sieve S_ϵ goes to zero as $\epsilon \rightarrow 0$. This becomes a challenge when trying to estimate differences of the form

$$\oint_{\mathbb{S}^{d-1}} g(x) dx - \oint_{S_\epsilon} g(x) dx.$$

The functions of interest arise as squares of traces of H^1 functions. If h is a function in H^1 , then its trace is in $W^{1/2,2} \subset W^{1/4,2+2/(2d-1)}$, and the square of its trace, by Kato-Ponce, is in $W^{1/4,1+1/(2d-1)} \subset W^{1/4,1}$. In particular, for $g : B_{1+\delta}(\mathbb{R}^d) \rightarrow \mathbb{R}$ we have

$$\left| \oint_{\mathbb{S}^{d-1}} g^2(x) dx - \oint_{S_\epsilon} g^2(x) dx \right| \lesssim_{d,\delta} \|g\|_{W^{1/2,2}(\mathbb{S}^{d-1})}^2 \left\| \frac{1}{|\mathbb{S}^{d-1}|} - \frac{1}{|S_\epsilon|} \mathbb{1}_{S_\epsilon} \right\|_{W^{-1/4,\infty}(\mathbb{S}^{d-1})}.$$

Lemma 4.24 (Homogenization estimate). *We have*

$$\left\| \frac{1}{|\mathbb{S}^{d-1}|} - \frac{1}{|S_\epsilon|} \mathbb{1}_{S_\epsilon} \right\|_{W^{-1/4,\infty}(\mathbb{S}^{d-1})} \leq C_{\beta,\delta,d} e^{-\frac{1}{5\epsilon}}.$$

In fact, showing convergence at a *uniform* rate would be enough, and the exponential rate is not necessary. This would require allowing for balls of size ϵ^{N_d} instead of $e^{-1/\epsilon}$ as the balls for which (4.5) holds. This type of argument is not applicable to the more commonly studied *thin* versions of the Neumann sieve, but it leads to a self-contained argument.

We will first show the energy bound estimates and derive the mass preservation estimates from those.

Proof of the energy bound estimates in Lemma 4.23. Let $f \in H^1(B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1})$. Then

$$\int_{\Omega_\epsilon} |\nabla \text{RI} f|^2 dx = (1 + O(\epsilon)) \left(\int_{B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1}} |\nabla f|^2 dx + |\mathbb{S}^{d-1}| \alpha \oint_{S_\epsilon} |f(1^-x) - f(1^+x)|^2 dx \right).$$

The $W^{1/2,1}$ -norm of $|f(1^-x) - f(1^+x)|^2$ is bounded by the $H^1(B_{1+\delta}(\mathbb{R}^d) \setminus \mathbb{S}^{d-1})$ norm of f . The result then follows from the estimate in Lemma 4.24 applied to the last term, because

$$\begin{aligned} & \left| \oint_{S_\epsilon} |f(1^-x) - f(1^+x)|^2 dx - \oint_{\mathbb{S}^{d-1}} |f(1^-x) - f(1^+x)|^2 dx \right| \\ & \leq \| |f(1^-x) - f(1^+x)|^2 \|_{W^{1/2,1}(\mathbb{S}^{d-1})} \\ & \quad \cdot \left\| \frac{1}{|\mathbb{S}^{d-1}|} - \frac{1}{|S_\epsilon|} \mathbb{1}_{S_\epsilon} \right\|_{W^{-1/4,\infty}(\mathbb{S}^{d-1})}. \end{aligned}$$

The second energy bound follows from the same approach: by the Poincaré inequality on segments, for any function $g \in H^1(\Omega_\epsilon)$ we have

$$\begin{aligned} & \alpha \int_{\mathbb{S}^{d-1}} |g((1 + \epsilon/2)x) - g((1 - \epsilon/2)x)|^2 dx \\ & = (1 + O(\epsilon)) \alpha |\mathbb{S}^{d-1}| \oint_{S_\epsilon} |g((1 + \epsilon/2)x) - g((1 - \epsilon/2)x)|^2 dx \\ & \leq \int_{N_\epsilon} |\partial_r g(x)|^2 dx \leq \int_{N_\epsilon} |\nabla g(x)|^2 dx. \end{aligned}$$

This shows that the contributions from the neck N_ϵ asymptotically bound the contributions in the jump.

The norm concentration estimates are proven in essentially the same way. The key estimate states that for any function bounded in H^1 , there cannot be much mass in the necks, because the necks are contained in a set of size $\lesssim \epsilon$:

$$\int_{N_\epsilon} g(x)^2 dx \lesssim \epsilon \left(\int_{\mathbb{S}^{d-1}} g((1 + \epsilon/2)x)^2 + g((1 - \epsilon/2)x)^2 dx + \int_{N_\epsilon} |\nabla g(x)|^2 dx \right).$$

This shows that adding (or removing) the necks will not change the mass by much, thus proving the norm-preservation inequalities. $\square$

Proof of Lemma 4.24. Let $W(x) := \frac{1}{|\mathbb{S}^{d-1}|} - \frac{1}{|S_\epsilon|} \mathbb{1}_{S_\epsilon}$. We know that $\|W(x)\|_{L^\infty(\mathbb{S}^{d-1})} \approx \epsilon^{-2}$, and therefore it suffices to prove that $\|W(x)\|_{W^{-1,\infty}(\mathbb{S}^{d-1})} \lesssim e^{-\frac{4}{5\epsilon}}$. We show this by duality. Let $f \in W^{1,1}(\mathbb{S}^{d-1})$ of norm one. Let A_s be the averaging operator at scale s (averaging f on caps of radius s), which satisfies $\|A_s f - f\|_{L^1(\mathbb{S}^{d-1})} \lesssim s \|f\|_{W^{1,1}(\mathbb{S}^{d-1})}$. We have

$$\begin{aligned} \langle W, f \rangle &\leq \|W\|_{L^\infty(\mathbb{S}^{d-1})} \|(A_{e^{-1/\epsilon}} - I)f\|_{L^1(\mathbb{S}^{d-1})} + \|A_{e^{-1/\epsilon}} W\|_{L^\infty(\mathbb{S}^{d-1})} \|f - \bar{f}\|_{L^1(\mathbb{S}^{d-1})} \\ &\lesssim \epsilon^{-2} \cdot e^{-\frac{1}{\epsilon}} + e^{-1/\epsilon} \cdot 1 \lesssim e^{-\frac{4}{5\epsilon}}. \end{aligned}$$

The bound $\|A_{e^{-1/\epsilon}} W\|_{L^\infty(\mathbb{S}^{d-1})} \lesssim e^{-\frac{1}{\epsilon}}$ is precisely (4.5). The result then follows by interpolation. $\square$

4.5.3 Estimating the first eigenfunction of Ω_ϵ near the sieve

By hypothesis, the first nontrivial eigenvalue of $D_{\beta,\delta}$ is unique, with a non-zero spectral gap. In particular, we have the following lemma.

Lemma 4.25. *The functions $\text{RE}(\psi_{\Omega_\epsilon}^{(1)})$ must converge in $L^2(B_{1+\delta}(\mathbb{R}^d))$ to $\psi_{\beta,\delta}^{(1)}$. By elliptic regularity, this convergence holds in the $C_{loc}^\infty(\overline{B_{1+\delta}(\mathbb{R}^d)} \setminus \mathbb{S}^{d-1})$ topology as well.*

Near $\mathbb{S}^{d-1}$ we do not have direct access to similarly fine estimates, but we can still upper bound $\psi_{\Omega_\epsilon}^{(1)}$ in N_ϵ :

Lemma 4.26. *We have*

$$\limsup_{\epsilon \rightarrow 0} \max_{x \in N_\epsilon} \psi_{\Omega_\epsilon}^{(1)}(x) \leq \max_{r \in \{1^+, 1^-\}} \psi_{\beta,\delta}^{(1)}(r).$$

Proof. Let $s = s(\epsilon)$ be a small parameter with $0 < \epsilon < s$ and let $S_\epsilon := \{x \in \Omega_\epsilon \text{ s.t. } |x| \in [1-s, 1+s]\}$. We will proceed by a comparison principle (in the sense of Proposition 4.12). By the monotonicity of Dirichlet eigenvalues, λ_S is bounded from below by the first Dirichlet

eigenvalue of $\{x \in \mathbb{R}^2 \text{ s.t. } |x| \in [1-s, 1+s]\}$, which is $\gtrsim s^{-2}$. In particular, for s small enough, the hypothesis of Proposition 4.12 is satisfied. If s is small enough, the function

$$b_{s,\epsilon}(x) = ((\mu_{\beta,\delta}^{(1)} + 1)(s^2 - (|x| - 1)^2) + 1) \cdot \max_{\substack{r \in \{1+s, 1-s\} \\ e \in \mathbb{S}^{d-1}}} \psi_{\Omega_\epsilon}^{(1)}(r \cdot e)$$

is a barrier function for $\psi_{\Omega_\epsilon}^{(1)}(x)$. For fixed s , uniform convergence of $\psi_{\Omega_\epsilon}^{(1)}(x)$ shows that

$$\limsup_{\epsilon \rightarrow 0} \max_{x \in N_\epsilon} \psi_{\Omega_\epsilon}^{(1)}(x) \leq (1 + (\mu_{\beta,\delta}^{(1)} + 1)s^2) \max_{r \in \{1+s, 1-s\}} \psi_{\beta,\delta}^{(1)}(r).$$

Sending $s \rightarrow 0$ proves the result. $\square$

4.6 Asymptotics as $d \rightarrow \infty$

Computing our desired asymptotics as $d \rightarrow \infty$ reduces to the following lemma:

Lemma 4.27. *The functions $\eta_d(r)$ converge uniformly for $r \in [0, 1]$ to the function*

$$\eta_\infty(r) := e^{(1-r^2)/2}.$$

Proof. Let $\tilde{\eta}_d(x) := \eta_d(x)/\|\eta_d\|_\infty$. Let σ_d be the uniform probability measure on the sphere of radius $\sqrt{\mu_{B_1}(\mathbb{R}^d)}$. Then $\tilde{\sigma}_d$ is a radial eigenfunction of the Laplace operator, with the same eigenvalue as $\tilde{\eta}_d(x)$, taking the same value at zero. Therefore, $\tilde{\eta}_d(x) = \widehat{\tilde{\sigma}_d}$. Let $\pi_*^1 \sigma_d$ be the projection (push-forward) of σ_d onto the first coordinate. Then $\tilde{\eta}_d(x) = \widehat{\pi_*^1 \sigma_d}(|x|)$, and it suffices to show that $\pi_*^1 \sigma_d \rightarrow_{L^1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

The eigenvalues of the ball have asymptotics $\mu_{B_1}(\mathbb{R}^d) = d + O(1)$, and thus the measures σ_d are uniform on spheres of radius $\sqrt{d} + O(d^{-1/2})$. The marginals of spheres of radius $\sqrt{d}$ converge to standard normals, showing the convergence to a Gaussian. $\square$

The fact that marginals of balls of radius $\sqrt{d}$ converge to a Gaussian is a well-known fact in high-dimensional probability (see, e.g. [370]) but this can also be computed explicitly. The one-dimensional marginal of the ball is

$$\frac{1}{Z_d} \mathbb{1}_{[-\sqrt{d}, \sqrt{d}]} (1 - x^2/d)^{d/2} \sqrt{1 + \frac{x/d}{1 - x^2/d}}$$

which converges (up to the normalization constant) to $e^{-x^2/2}$ as $d \rightarrow \infty$.

Corollary 4.4 follows from the fact that $\max_{|x| < 1} \eta_\infty(x) = \eta_\infty(0) = \sqrt{e}$. Lemma 4.27 implies that, as $d \rightarrow \infty$,

$$\frac{|\{\psi_\Omega^{(1)}(x) > (1 + \delta) \max_{y \in \partial\Omega} \psi_\Omega^{(1)}(y)\}|}{|\Omega|} \leq (1 - \delta + o(1))^d \lesssim e^{-d\delta(1-o(1))},$$

establishing Corollary 4.7.

BIBLIOGRAPHY

- [1] B. Hosseini, A. W. Hsu, and A. Taghvaei. “Conditional optimal transport on function spaces”. In: *SIAM/ASA Journal on Uncertainty Quantification* 13.1 (2025), pp. 304–338.
- [2] Y. Jalalian, J. F. Osorio Ramirez, A. Hsu, B. Hosseini, and H. Owhadi. *Data-Efficient Kernel Methods for Learning Differential Equations and Their Solution Operators: Algorithms and Error Analysis*. 2025. arXiv: 2503.01036.
- [3] A. W. Hsu, I. W. Griss Salas, J. M. Stevens-Haas, J. N. Kutz, A. Aravkin, and B. Hosseini. *A joint optimization approach to identifying sparse dynamics using least squares kernel collocation*. 2025. arXiv: 2511.18555 [math.OC].
- [4] J. de Dios Pont, A. W. Hsu, and M. A. Taylor. *Sharp bounds on the failure of the hot spots conjecture*. 2025. arXiv: 2508.16321.
- [5] C. P. Robert, G. Casella, and G. Casella. *Monte Carlo statistical methods*. Vol. 2. Springer, 1999.
- [6] A. M. Stuart. “Inverse problems: a Bayesian perspective”. In: *Acta Numerica* 19 (2010), pp. 451–559.
- [7] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. “MCMC methods for functions: modifying old algorithms to make them faster”. In: *Statistical Science* 28.3 (2013), pp. 424–446.
- [8] M. Hairer, A. M. Stuart, S. J. Vollmer, et al. “Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions”. In: *The Annals of Applied Probability* 24.6 (2014), pp. 2455–2490.
- [9] T. Cui, K. J. Law, and Y. M. Marzouk. “Dimension-independent likelihood-informed MCMC”. In: *Journal of Computational Physics* 304 (2016), pp. 109–137.
- [10] A. Beskos, M. Girolami, S. Lan, P. E. Farrell, and A. M. Stuart. “Geometric MCMC for infinite-dimensional inverse problems”. In: *Journal of Computational Physics* 335 (2017), pp. 327–351.
- [11] M. Betancourt, S. Byrne, S. Livingstone, and M. Girolami. “The geometric foundations of Hamiltonian Monte Carlo”. In: *Bernoulli* 23.4A (2017), pp. 2257–2298.
- [12] A. Durmus and É. Moulines. “Nonasymptotic convergence analysis for the unadjusted Langevin algorithm”. In: *Annals of Applied Probability* 27.3 (2017), pp. 1551–1587.
- [13] B. Hosseini. “Two Metropolis–Hastings Algorithms for Posterior Measures with Non-Gaussian Priors in Infinite Dimensions”. In: *SIAM/ASA Journal on Uncertainty Quantification* 7.4 (2019), pp. 1185–1223.

- [14] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. “Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler”. In: *SIAM Journal on Applied Dynamical Systems* 19.1 (2020), pp. 412–441.
- [15] B. Hosseini and J. E. Johndrow. “Spectral gaps and error estimates for infinite-dimensional Metropolis–Hastings with non-Gaussian priors”. In: *The Annals of Applied Probability* 33.3 (2023), pp. 1827–1873.
- [16] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. “Variational inference: A review for statisticians”. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.
- [17] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. “Advances in variational inference”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.8 (2018), pp. 2008–2026.
- [18] C. W. Fox and S. J. Roberts. “A tutorial on variational Bayesian inference”. In: *Artificial intelligence review* 38.2 (2012), pp. 85–95.
- [19] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. “Sampling via measure transport: An introduction”. In: *Handbook of Uncertainty Quantification* (2016), pp. 1–41.
- [20] A. Y. Ng and M. I. Jordan. “On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes”. In: *Advances in neural information processing systems*. 2002, pp. 841–848.
- [21] T. Jebara. *Machine learning: discriminative and generative*. Vol. 755. Springer Science & Business Media, 2012.
- [22] I. Goodfellow. “NIPS 2016 tutorial: Generative adversarial networks”. In: *arXiv preprint:1701.00160* (2016).
- [23] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.
- [24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.
- [25] D. Rezende and S. Mohamed. “Variational inference with normalizing flows”. In: *International Conference on Machine Learning*. 2015.
- [26] I. Kobyzev, S. J. Prince, and M. A. Brubaker. “Normalizing flows: An introduction and review of current methods”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 3964–3979.
- [27] G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. “Normalizing Flows for Probabilistic Modeling and Inference.” In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64.
- [28] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265.

- [29] Y. Song and S. Ermon. “Improved techniques for training score-based generative models”. In: *Advances in neural information processing systems* 33 (2020), pp. 12438–12448.
- [30] D. Kingma, T. Salimans, B. Poole, and J. Ho. “Variational diffusion models”. In: *Advances in neural information processing systems* 34 (2021), pp. 21696–21707.
- [31] H. Cao, C. Tan, Z. Gao, G. Chen, P.-A. Heng, and S. Z. Li. “A survey on generative diffusion model”. In: *arXiv:2209.02646* (2022).
- [32] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. “Stochastic interpolants: A unifying framework for flows and diffusions”. In: *arXiv:2303.08797* (2023).
- [33] L. Pardo. *Statistical inference based on divergence measures*. CRC press, 2018.
- [34] C. Villani. *Optimal transport: Old and new*. Springer, 2009.
- [35] G. Peyré and M. Cuturi. “Computational optimal transport: With applications to data science”. In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.
- [36] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. “Stochastic optimization for large-scale optimal transport”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.
- [37] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein generative adversarial networks”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 214–223.
- [38] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. “Improved training of Wasserstein GANs”. In: *International Conference on Neural Information Processing Systems*. 2017.
- [39] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. “Optimal transport mapping via input convex neural networks”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 6672–6681.
- [40] D. Onken, S. Wu Fung, X. Li, and L. Ruthotto. “OT-flow: Fast and accurate continuous normalizing flows via optimal transport”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2021.
- [41] L. Rout, A. Korotin, and E. Burnaev. “Generative Modeling with Optimal Transport Maps”. In: *International Conference on Learning Representations*. 2021.
- [42] T. Uscidda and M. Cuturi. “The Monge Gap: A Regularizer to Learn All Transport Maps”. In: *arXiv:2302.04953* (2023).
- [43] M. Daniels, T. Maunu, and P. Hand. “Score-based generative neural networks for large-scale optimal transport”. In: *Advances in neural information processing systems* 34 (2021), pp. 12955–12965.
- [44] A. Korotin, L. Li, A. Genevay, J. M. Solomon, A. Filippov, and E. Burnaev. “Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark”. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

- [45] B. Amos. “On amortizing convex conjugates for optimal transport”. In: *The Eleventh International Conference on Learning Representations*. 2022.
- [46] F. Coeurdoux, N. Dobigeon, and P. Chainais. “Learning optimal transport between two empirical distributions with normalizing flows”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2022, pp. 275–290.
- [47] A. Korotin, D. Selikhanovych, and E. Burnaev. “Kernel neural optimal transport”. In: *The Eleventh International Conference on Learning Representations*. 2022.
- [48] A. Korotin, D. Selikhanovych, and E. Burnaev. “Neural optimal transport”. In: *arXiv:2201.12220* (2023).
- [49] M. A. Rahman, M. A. Florez, A. Anandkumar, Z. E. Ross, and K. Azizzadenesheli. “Generative Adversarial Neural Operators”. In: *Transactions on Machine Learning Research* (2022).
- [50] G. Kerrigan, J. Ley, and P. Smyth. “Diffusion Generative Models in Infinite Dimensions”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 9538–9563.
- [51] G. Franzese, G. Corallo, S. Rossi, M. Heinonen, M. Filippone, and P. Michiardi. “Continuous-time functional diffusion processes”. In: *Advances in Neural Information Processing Systems* 36 (2024).
- [52] R. Baptista and N. B. Kovachki. “Score-based diffusion models in function space”. In: *arXiv:2302.07400* (2023).
- [53] R. Baptista, B. Hosseini, N. B. Kovachki, and Y. M. Marzouk. “Conditional sampling with monotone GANs: from generative models to likelihood-free inference”. In: *SIAM/ASA Journal on Uncertainty Quantification* 12.3 (2024), pp. 868–900.
- [54] A. Siahkoohi, G. Rizzuti, R. Orozco, and F. J. Herrmann. “Reliable amortized variational inference with physics-based latent distribution correction”. In: *Geophysics* 88.3 (2023), R297–R322.
- [55] M. A. Iglesias, K. Lin, and A. M. Stuart. “Well-posed Bayesian geometric inverse problems arising in subsurface flow”. In: *Inverse Problems* 30.11 (2014), p. 114001.
- [56] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer Science & Business Media, 2005.
- [57] V. I. Bogachev. *Measure Theory*. Vol. 2. Springer, New York, 2007.
- [58] C. Bunne, A. Krause, and M. Cuturi. “Supervised training of conditional Monge maps”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 6859–6872.
- [59] G. Carlier, V. Chernozhukov, and A. Galichon. “Vector quantile regression: an optimal transport approach”. In: *The Annals of Statistics* 44.3 (2016), pp. 1165–1192.
- [60] A. Delalande and Q. Mérigot. “Quantitative stability of optimal transport maps under variations of the target measure”. In: *Duke Mathematical Journal* 172.17 (2023), pp. 3321–3357.

- [61] A. Figalli and F. Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. EMS, 2021.
- [62] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.
- [63] F. Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.
- [64] G. Carlier, A. Galichon, and F. Santambrogio. “From Knothe’s transport to Brenier’s map and a continuation method for optimal transport”. In: *SIAM Journal on Mathematical Analysis* 41.6 (2010), pp. 2554–2576.
- [65] N. Bonnotte. “From Knothe’s rearrangement to Brenier’s optimal transport map”. In: *SIAM Journal on Mathematical Analysis* 45.1 (2013), pp. 64–87.
- [66] H. Knothe. “Contributions to the theory of convex bodies”. In: *Michigan Mathematical Journal* 4.1 (1957), pp. 39–52.
- [67] M. Rosenblatt. “Remarks on a multivariate transformation”. In: *The annals of mathematical statistics* 23.3 (1952), pp. 470–472.
- [68] V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. “Triangular transformations of measures”. In: *Sbornik: Mathematics* 196.3 (2005), p. 309.
- [69] V. I. Bogachev and A. V. Kolesnikov. “Nonlinear transformations of convex measures”. In: *Theory of Probability & Its Applications* 50.1 (2006), pp. 34–52.
- [70] Z. O. Wang, R. Baptista, Y. Marzouk, L. Ruthotto, and D. Verma. “Efficient Neural Network Approaches for Conditional Optimal Transport with Applications in Bayesian Inference”. In: *arXiv:2310.16975* (2023).
- [71] M. Al-Jarrah, B. Hosseini, and A. Taghvaei. “Optimal Transport Particle Filters”. In: *arXiv:2304.00392* (2023).
- [72] M. Al-Jarrah, N. Jin, B. Hosseini, and A. Taghvaei. “Optimal Transport-based Nonlinear Filtering in High-dimensional Settings”. In: *arXiv:2310.13886* (2023).
- [73] J. Alfonso, R. Baptista, A. Bhakta, N. Gal, A. Hou, I. Lyubimova, D. Pocklington, J. Sajonz, G. Trigila, and R. Tsai. “A generative flow for conditional sampling via optimal transport”. In: *arXiv:2307.04102* (2023).
- [74] A. Taghvaei and B. Hosseini. “An optimal transport formulation of Bayes’ law for nonlinear filtering algorithms”. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6608–6613.
- [75] B. Muzellec and M. Cuturi. “Subspace detours: Building transport plans that are optimal on subspace projections”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [76] G. Kerrigan, G. Migliorini, and P. Smyth. “Dynamic Conditional Optimal Transport through Simulation-Free Flows”. In: *arXiv:2404.04240* (2024).
- [77] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. “Flow Matching for Generative Modeling”. In: *arXiv [cs.LG]* (Oct. 6, 2022).

- [78] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Q. Chen. “Multisample Flow Matching: Straightening Flows with Minibatch Couplings”. In: *arXiv [cs.LG]* (Apr. 28, 2023).
- [79] A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. “Improving and generalizing flow-based generative models with minibatch optimal transport”. In: *arXiv [cs.LG]* (Feb. 1, 2023).
- [80] T. A. El Moselhy and Y. M. Marzouk. “Bayesian inference with optimal maps”. In: *Journal of Computational Physics* 231.23 (2012), pp. 7815–7850.
- [81] L. Baldassari, A. Siahkoohi, J. Garnier, K. Solna, and M. V. de Hoop. “Conditional score-based diffusion models for Bayesian inference in infinite dimensions”. In: *arXiv:2305.19147* (2023).
- [82] Z. Yin, R. Orozco, M. Louboutin, and F. J. Herrmann. “Solving multiphysics-based inverse problems with learned surrogates and constraints”. In: *Advanced Modeling and Simulation in Engineering Sciences* 10.1 (2023), p. 14.
- [83] A. Siahkoohi, G. Rizzuti, M. Louboutin, P. A. Witte, and F. J. Herrmann. “Preconditioned training of normalizing flows for variational inference in inverse problems”. In: *arXiv preprint:2101.03709* (2021).
- [84] A. Spantini, R. Baptista, and Y. Marzouk. “Coupling techniques for nonlinear ensemble filtering”. In: *SIAM Review (In Press)* (2019).
- [85] M. Ramgraber, R. Baptista, D. McLaughlin, and Y. Marzouk. “Ensemble transport smoothing—Part 1: unified framework”. In: *arXiv:2210.17000* (2022).
- [86] M. Ramgraber, R. Baptista, D. McLaughlin, and Y. Marzouk. “Ensemble transport smoothing—Part 2: nonlinear updates”. In: *arXiv:2210.17435* (2022).
- [87] D. Grange, M. Al-Jarrah, R. Baptista, A. Taghvaei, T. T. Georgiou, and A. Tannenbaum. “Computational Optimal Transport and Filtering on Riemannian manifolds”. In: *arXiv:2309.08847* (2023).
- [88] B. L. Trippe and R. E. Turner. “Conditional density estimation with Bayesian normalising flows”. In: *arXiv:1802.04908* (2018).
- [89] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling. “Learning likelihoods with conditional normalizing flows”. In: *arXiv:1912.00042* (2019).
- [90] J.-M. Lueckmann, G. Bassetto, T. Karaletsos, and J. H. Macke. “Likelihood-free inference with emulator networks”. In: *Symposium on Advances in Approximate Bayesian Inference*. PMLR. 2019, pp. 32–53.
- [91] M. Mirza and S. Osindero. “Conditional generative adversarial nets”. In: *arXiv:1411.1784* (2014).
- [92] J. Adler and O. Öktem. “Deep Bayesian inversion”. In: *arXiv preprint:1811.05910* (2018).
- [93] S. Liu, X. Zhou, Y. Jiao, and J. Huang. “Wasserstein generative learning of conditional distribution”. In: *arXiv:2112.10039* (2021).

- [94] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann. “Conditional image generation with score-based diffusion models”. In: *arXiv:2111.13606* (2021).
- [95] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. “Image super-resolution via iterative refinement”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022), pp. 4713–4726.
- [96] D. Ray, H. Ramaswamy, D. V. Patel, and A. A. Oberai. “The efficacy and generalizability of conditional GANs for posterior inference in physics-based inverse problems”. In: *arXiv preprint:2202.07773* (2022).
- [97] X. Zhou, Y. Jiao, J. Liu, and J. Huang. “A deep generative approach to conditional sampling”. In: *Journal of the American Statistical Association* (2022), pp. 1–12.
- [98] R. Baptista, O. Zahm, and Y. Marzouk. “An adaptive transport framework for joint and conditional density estimation”. In: *arXiv preprint:2009.10303* (2020).
- [99] J.-C. Hütter and P. Rigollet. “Minimax estimation of smooth optimal transport maps”. In: *The Annals of Statistics* 49.2 (2021).
- [100] V. Divol, J. Niles-Weed, and A.-A. Pooladian. “Optimal transport map estimation in general function spaces”. In: *arXiv:2212.03722* (2022).
- [101] P. Ghosal and B. Sen. “Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing”. In: *The Annals of Statistics* 50.2 (2022), pp. 1012–1037.
- [102] N. Deb, P. Ghosal, and B. Sen. “Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29736–29753.
- [103] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. “Plugin estimation of smooth optimal transport maps”. In: *arXiv:2107.12364* (2021).
- [104] B. Muzellec, A. Vacher, F. Bach, F.-X. Vialard, and A. Rudi. “Near-optimal estimation of smooth transport maps with kernel sums-of-squares”. In: *arXiv:2112.01907* (2021).
- [105] A.-A. Pooladian and J. Niles-Weed. “Entropic estimation of optimal transport maps”. In: *arXiv:2109.12004* (2024).
- [106] R. Baptista, B. Hosseini, N. B. Kovachki, Y. M. Marzouk, and A. Sagiv. “An approximation theory framework for measure-transport sampling algorithms”. In: *arXiv:2302.13965* (2023).
- [107] J. Zech and Y. Marzouk. “Sparse Approximation of Triangular Transports, Part I: The Finite-Dimensional Case”. In: *Constructive Approximation* 55.3 (2022), pp. 919–986.
- [108] J. Zech and Y. Marzouk. “Sparse approximation of triangular transports, part II: The infinite-dimensional case”. In: *Constructive Approximation* 55.3 (2022), pp. 987–1036.

- [109] N. J. Irons, M. Scetbon, S. Pal, and Z. Harchaoui. “Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 10161–10195.
- [110] S. Wang and Y. Marzouk. “On minimax density estimation via measure transport”. In: *arXiv:2207.10231* (2022).
- [111] P. Jaini, I. Kobyzev, Y. Yu, and M. Brubaker. “Tails of Lipschitz triangular flows”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4673–4681.
- [112] J. Rabin, G. Peyré, and L. D. Cohen. “Geodesic shape retrieval via optimal mass transport”. In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part V 11*. Springer. 2010, pp. 771–784.
- [113] J. Rabin, G. Peyré, J. Delon, and M. Bernot. “Wasserstein barycenter and its application to texture mixing”. In: *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*. Springer. 2012, pp. 435–446.
- [114] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. “Sliced and Radon wasserstein barycenters of measures”. In: *Journal of Mathematical Imaging and Vision* 51 (2015), pp. 22–45.
- [115] G. Mahey, L. Chapel, G. Gasso, C. Bonet, and N. Courty. “Fast Optimal Transport through Sliced Wasserstein Generalized Geodesics”. In: *arXiv:2307.01770* (2023).
- [116] S. Li and C. Moosmueller. “Measure transfer via stochastic slicing and matching”. In: *arXiv:2307.05705* (2023).
- [117] S. Li and C. Moosmueller. “Approximation properties of slice-matching operators”. In: *arXiv:2310.10869* (2023).
- [118] L. Rüschendorf. “Bounds for distributions with multivariate marginals”. In: *Lecture Notes-Monograph Series* (1991), pp. 285–310.
- [119] P. Embrechts and G. Puccetti. “Bounds for the sum of dependent risks having overlapping marginals”. In: *Journal of Multivariate Analysis* 101.1 (2010), pp. 177–190.
- [120] Y. Fan, H. Park, and G. Xu. “Quantifying Distributional Model Risk in Marginal Problems via Optimal Transport”. In: *arXiv:2307.00779* (2023).
- [121] V. I. Bogachev. *Gaussian Measures*. AMS, Providence, 1998.
- [122] K. Yosida. *Functional Analysis*. 6th ed. New York: Springer-Verlag, 1980.
- [123] A. Grelaud, C. P. Robert, J.-M. Marin, F. Rodolphe, J.-F. Taly, et al. “ABC likelihood-free methods for model choice in Gibbs random fields”. In: *Bayesian Analysis* 4.2 (2009), pp. 317–335.
- [124] M. U. Gutmann and J. Corander. “Bayesian optimization for likelihood-free inference of simulator-based statistical models”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 4256–4302.

- [125] G. Papamakarios, D. Sterratt, and I. Murray. “Sequential Neural Likelihood: Fast Likelihood-free Inference with Autoregressive Flows”. In: *International Conference on Artificial Intelligence and Statistics*. 2019.
- [126] V. Chen, M. M. Dunlop, O. Papaspiliopoulos, and A. M. Stuart. “Dimension-robust MCMC in Bayesian inverse problems”. In: *arXiv preprint:1803.03344* (2018).
- [127] B. Sprungk. “On the local Lipschitz stability of Bayesian inverse problems”. In: *Inverse Problems* 36.5 (2020), p. 055015.
- [128] A. Garbuno-Inigo, T. Helin, F. Hoffmann, and B. Hosseini. “Bayesian Posterior Perturbation Analysis with Integral Probability Metrics”. In: *arXiv:2303.01512* (2023).
- [129] R. J. Berman. “Convergence rates for discretized Monge–Ampère equations and quantitative stability of optimal transport”. In: *Foundations of Computational Mathematics* 21.4 (2021), pp. 1099–1140.
- [130] Q. Mérigot, A. Delalande, and F. Chazal. “Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space”. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, Aug. 2020, pp. 3186–3196.
- [131] K. Bhattacharya, B. Hosseini, N. B. Kovachki, and A. M. Stuart. “Model Reduction And Neural Networks For Parametric PDEs”. In: *The SMAI journal of computational mathematics* 7 (2021), pp. 121–157.
- [132] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. “FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models”. In: *International Conference on Learning Representations*. 2018.
- [133] M. S. Alnaes, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. “The FEniCS Project Version 1.5”. In: *Archive of Numerical Software* 3 (2015).
- [134] M. S. Albergo and E. Vanden-Eijnden. “Building normalizing flows with stochastic interpolants”. In: *arXiv:2209.15571* (2022).
- [135] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. “Physics-informed machine learning”. In: *Nature Reviews Physics* 3.6 (2021), pp. 422–440.
- [136] N. Boullé and A. Townsend. “Chapter 3 - A mathematical guide to operator learning”. In: *Numerical Analysis Meets Machine Learning*. Ed. by S. Mishra and A. Townsend. Vol. 25. Handbook of Numerical Analysis. Elsevier, 2024, pp. 83–125.
- [137] S. L. Brunton, J. L. Proctor, and J. N. Kutz. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems”. In: *Proceedings of the National Academy of Sciences* 113.15 (2016), pp. 3932–3937.
- [138] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. “Data-driven discovery of partial differential equations”. In: *Science advances* 3.4 (2017), e1602614.

- [139] N. B. Kovachki, S. Lanthaler, and A. M. Stuart. “Chapter 9 - Operator learning: Algorithms and analysis”. In: *Numerical Analysis Meets Machine Learning*. Ed. by S. Mishra and A. Townsend. Vol. 25. Handbook of Numerical Analysis. Elsevier, 2024, pp. 419–467.
- [140] P. Batlle, M. Darcy, B. Hosseini, and H. Owhadi. “Kernel methods are competitive for operator learning”. In: *Journal of Computational Physics* 496 (2024), p. 112549.
- [141] M. Raissi, P. Perdikaris, and G. E. Karniadakis. “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 (2019), pp. 686–707.
- [142] Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart. “Solving and learning nonlinear PDEs with Gaussian processes”. In: *Journal of Computational Physics* 447 (2021), p. 110668.
- [143] J. Bongard and H. Lipson. “Automated reverse engineering of nonlinear dynamical systems”. In: *Proceedings of the National Academy of Sciences* 104.24 (2007), pp. 9943–9948.
- [144] M. Schmidt and H. Lipson. “Distilling free-form natural laws from experimental data”. In: *Science* 324.5923 (2009), pp. 81–85.
- [145] K. J. Åström and P. Eykhoff. “System identification—a survey”. In: *Automatica* 7.2 (1971), pp. 123–162.
- [146] L. Ljung. “Perspectives on system identification”. In: *Annual Reviews in Control* 34.1 (2010), pp. 1–12.
- [147] K. J. Keesman. *System identification: an introduction*. Springer Science & Business Media, 2011.
- [148] H. Schaeffer. “Learning partial differential equations via data discovery and sparse optimization”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473.2197 (2017), p. 20160446.
- [149] S. H. Kang, W. Liao, and Y. Liu. “Ident: Identifying differential equations with numerical time evolution”. In: *Journal of Scientific Computing* 87.1 (2021), pp. 1–27.
- [150] D. A. Messenger and D. M. Bortz. “Weak SINDy: Galerkin-based data-driven model selection”. In: *Multiscale Modeling & Simulation* 19.3 (2021), pp. 1474–1497.
- [151] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton. “Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control”. In: *Proceedings of the Royal Society A* 478.2260 (2022), p. 20210904.
- [152] S. Zhang and G. Lin. “Robust data-driven discovery of governing physical laws with error bars”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 474.2217 (2018), p. 20180305.

- [153] R. K. Niven, A. Mohammad-Djafari, L. Cordier, M. W. Abel, and M. Quade. “Bayesian identification of dynamical systems”. In: *39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2019)*. Vol. 33. 2019.
- [154] Y. Yang, M. Aziz Bhouri, and P. Perdikaris. “Bayesian differential programming for robust systems identification under uncertainty”. In: *Proceedings of the Royal Society A* 476.2243 (2020), p. 20200290.
- [155] S. M. Hirsh, D. A. Barajas-Solano, and J. N. Kutz. “Sparsifying priors for Bayesian uncertainty quantification in model discovery”. In: *Royal Society Open Science* 9.2 (2022), p. 211823.
- [156] J. S. North, C. K. Wikle, and E. M. Schliep. “A Bayesian approach for data-driven dynamic equation discovery”. In: *Journal of Agricultural, Biological and Environmental Statistics* 27.4 (2022), pp. 728–747.
- [157] J. S. North, C. K. Wikle, and E. M. Schliep. “A Bayesian Approach for Spatio-Temporal Data-Driven Dynamic Equation Discovery”. In: *Bayesian Analysis* 1.1 (2023), pp. 1–30.
- [158] J. S. North, C. K. Wikle, and E. M. Schliep. “A Review of Data-Driven Discovery for Dynamic Systems”. In: *International Statistical Review* 91.3 (2023), pp. 464–492.
- [159] F. Sun, Y. Liu, and H. Sun. “Physics-informed Spline Learning for Nonlinear Dynamics Discovery”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. Ed. by Z.-H. Zhou. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 2054–2061.
- [160] Z. Chen, Y. Liu, and H. Sun. “Physics-informed learning of governing equations from scarce data”. In: *Nature communications* 12.1 (2021), p. 6136.
- [161] H. Owhadi. “Bayesian numerical homogenization”. In: *Multiscale Modeling & Simulation* 13.3 (2015), pp. 812–828.
- [162] O. A. Chkrebtii, D. A. Campbell, B. Calderhead, and M. A. Girolami. “Bayesian Solution Uncertainty Quantification for Differential Equations”. In: *Bayesian Analysis* 11.4 (2016), pp. 1239–1267.
- [163] J. Cockayne, C. Oates, T. Sullivan, and M. Girolami. “Probabilistic numerical methods for PDE-constrained Bayesian inverse problems”. In: *AIP Conference Proceedings*. Vol. 1853. Author(s), 2017, p. 060001.
- [164] M. Raissi, P. Perdikaris, and G. E. Karniadakis. “Numerical Gaussian processes for time-dependent and nonlinear partial differential equations”. In: *SIAM Journal on Scientific Computing* 40.1 (2018), A172–A198.
- [165] H. Owhadi, C. Scovel, and F. Schäfer. “Statistical numerical approximation”. In: *Notices of the AMS* (2019).
- [166] H. Owhadi and C. Scovel. *Operator-Adapted Wavelets, Fast Solvers, and Numerical Homogenization: From a Game Theoretic Approach to Numerical Approximation and Algorithm Design*. Vol. 35. Cambridge University Press, 2019.

- [167] J. Cockayne, C. J. Oates, T. J. Sullivan, and M. Girolami. “Bayesian probabilistic numerical methods”. In: *SIAM Review* 61.4 (2019), pp. 756–789.
- [168] J. Wang, J. Cockayne, O. Chkrebtii, T. J. Sullivan, and C. J. Oates. “Bayesian numerical methods for nonlinear partial differential equations”. In: *Statistics and Computing* 31 (2021), pp. 1–20.
- [169] M. Raissi, P. Perdikaris, and G. E. Karniadakis. “Machine learning of linear differential equations using Gaussian processes”. In: *Journal of Computational Physics* 348 (2017), pp. 683–693.
- [170] K. Lahouel, M. Wells, V. Rielly, E. Lew, D. Lovitz, and B. M. Jedynak. “Learning nonparametric ordinary differential equations from noisy data”. In: *Journal of Computational Physics* 507 (2024), p. 112971.
- [171] B. Hamzi and H. Owhadi. “Learning dynamical systems from data: a simple cross-validation perspective, part I: parametric kernel flows”. In: *Physica D: Nonlinear Phenomena* 421 (2021), p. 132817.
- [172] M. Darcy, B. Hamzi, J. Susiluoto, A. Braverman, and H. Owhadi. *Learning dynamical systems from data: a simple cross-validation perspective, part II: nonparametric kernel flows*. 2021.
- [173] J. Lee, E. De Brouwer, B. Hamzi, and H. Owhadi. “Learning dynamical systems from data: A simple cross-validation perspective, Part III: Irregularly-sampled time series”. In: *Physica D: Nonlinear Phenomena* 443 (2023), p. 133546.
- [174] B. Hamzi, H. Owhadi, and Y. Kevrekidis. “Learning dynamical systems from data: A simple cross-validation perspective, part iv: case with partial observations”. In: *Physica D: Nonlinear Phenomena* (2023), p. 133853.
- [175] L. Yang, X. Sun, B. Hamzi, H. Owhadi, and N. Xie. “Learning dynamical systems from data: A simple cross-validation perspective, Part V: Sparse Kernel Flows for 132 chaotic dynamical systems”. In: *Physica D: Nonlinear Phenomena* 460 (2024), p. 134070.
- [176] L. Yang, B. Hamzi, Y. Kevrekidis, H. Owhadi, X. Sun, and N. Xie. *Learning Dynamical Systems from Data: A Simple Cross-Validation Perspective, Part VI: Hausdorff metric based training of kernels to learn attractors with application to 133 chaotic dynamical systems*. 2023.
- [177] J. A. Rosenfeld, B. P. Russo, R. Kamalapurkar, and T. T. Johnson. “The Occupation Kernel Method for Nonlinear System Identification”. In: *SIAM Journal on Control and Optimization* 62.3 (2024), pp. 1643–1668.
- [178] M. Wells, K. Lahouel, and B. Jedynak. “The Stochastic Occupation Kernel Method for System Identification”. In: *arXiv preprint arXiv:2406.15661* (2024).
- [179] M. Darcy, B. Hamzi, G. Livieri, H. Owhadi, and P. Tavallali. “One-shot learning of stochastic differential equations with data adapted kernels”. In: *Physica D: Nonlinear Phenomena* 444 (2023), p. 133583.

- [180] D. Long, N. Mrvaljević, S. Zhe, and B. Hosseini. “A kernel framework for learning differential equations and their solution operators”. In: *Physica D: Nonlinear Phenomena* 460 (2024), p. 134095.
- [181] Z. Long, Y. Lu, X. Ma, and B. Dong. “PDE-net: Learning PDEs from data”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3208–3216.
- [182] Z. Long, Y. Lu, and B. Dong. “PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network”. In: *Journal of Computational Physics* 399 (2019), p. 108925.
- [183] H. Xu. “DL-PDE: Deep-Learning Based Data-Driven Discovery of Partial Differential Equations from Discrete and Noisy Data”. In: *Communications in Computational Physics* 29.3 (2021), pp. 698–728.
- [184] H. Xu, H. Chang, and D. Zhang. “DLGA-PDE: Discovery of PDEs with incomplete candidate library via combination of deep learning and genetic algorithm”. In: *Journal of Computational Physics* 418 (2020), p. 109584.
- [185] H. Xu, D. Zhang, and J. Zeng. “Deep-learning of parametric partial differential equations from sparse and noisy data”. In: *Physics of Fluids* 33.3 (2021).
- [186] M. Raissi. “Deep hidden physics models: Deep learning of nonlinear partial differential equations”. In: *Journal of Machine Learning Research* 19.25 (2018), pp. 1–24.
- [187] G.-J. Both, S. Choudhury, P. Sens, and R. Kusters. “DeepMoD: Deep learning for model discovery in noisy data”. In: *Journal of Computational Physics* 428 (2021), p. 109985.
- [188] B. C. Koenig, S. Kim, and S. Deng. “KAN-ODEs: Kolmogorov–Arnold network ordinary differential equations for learning dynamical systems and hidden physics”. In: *Computer Methods in Applied Mechanics and Engineering* 432 (2024), p. 117397.
- [189] T. Berry and S. Das. “Limits of Learning Dynamical Systems”. In: *SIAM Review* 67.1 (2025), pp. 107–137.
- [190] L. Zhang and H. Schaeffer. “On the convergence of the SINDy algorithm”. In: *Multiscale Modeling & Simulation* 17.3 (2019), pp. 948–972.
- [191] H. Schaeffer, G. Tran, R. Ward, and L. Zhang. “Extracting structured dynamical systems using sparse optimization with very few samples”. In: *Multiscale Modeling & Simulation* 18.4 (2020), pp. 1435–1461.
- [192] P. Scholl, A. Bacho, H. Boche, and G. Kutyniok. “The uniqueness problem of physical law learning”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.
- [193] Y. He, H. Zhao, and Y. Zhong. “How much can one learn a partial differential equation from its solution?” In: *Foundations of Computational Mathematics* 24.5 (2024), pp. 1595–1641.
- [194] A. Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

- [195] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*. Vol. 160. Springer Science & Business Media, 2006.
- [196] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Vol. 375. Springer Science & Business Media, 1996.
- [197] S. I. Kabanikhin. *Inverse and Ill-posed Problems: Theory and Applications*. Vol. 55. Walter de Gruyter, 2011.
- [198] C. R. Vogel. *Computational methods for inverse problems*. SIAM, 2002.
- [199] E. Haber and D. Oldenburg. “Joint inversion: a structural approach”. In: *Inverse problems* 13.1 (1997), p. 63.
- [200] B. Kaltenbacher. “Regularization Based on All-At-Once Formulations for Inverse Problems”. In: *SIAM Journal on Numerical Analysis* 54.4 (2016), pp. 2594–2618.
- [201] B. Kaltenbacher. “All-at-once versus reduced iterative methods for time dependent inverse problems”. In: *Inverse Problems* 33.6 (2017), p. 064002.
- [202] K. Law, A. Stuart, and K. Zygalakis. “Data assimilation”. In: *Cham, Switzerland: Springer* 214 (2015), p. 52.
- [203] S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.
- [204] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan. *Estimation with applications to tracking and navigation: theory algorithms and software*. John Wiley & Sons, 2004.
- [205] J. Xiong. *An introduction to stochastic filtering theory*. Vol. 18. OUP Oxford, 2008.
- [206] T. Glad and L. Ljung. *Control theory*. CRC press, 2018.
- [207] G. F. Franklin, J. D. Powell, A. Emami-Naeini, and J. D. Powell. *Feedback control of dynamic systems*. Vol. 4. Prentice hall Upper Saddle River, 2002.
- [208] H. U. Voss, J. Timmer, and J. Kurths. “Nonlinear dynamical system identification from uncertain and indirect measurements”. In: *International Journal of Bifurcation and Chaos* 14.06 (2004), pp. 1905–1933.
- [209] J. Schoukens and L. Ljung. “Nonlinear system identification: A user-oriented road map”. In: *IEEE Control Systems Magazine* 39.6 (2019), pp. 28–99.
- [210] S. Särkkä and L. Svensson. *Bayesian filtering and smoothing*. Vol. 17. Cambridge university press, 2023.
- [211] Z. Ghahramani and S. Roweis. “Learning nonlinear dynamical systems using an EM algorithm”. In: *Advances in neural information processing systems* 11 (1998).
- [212] T. B. Schön, A. Wills, and B. Ninness. “System identification of nonlinear state-space models”. In: *Automatica* 47.1 (2011), pp. 39–49.
- [213] E. A. Wan and A. T. Nelson. “Dual extended Kalman filter methods”. In: *Kalman filtering and neural networks* (2001), pp. 123–173.
- [214] J. Wang, A. Hertzmann, and D. J. Fleet. “Gaussian process dynamical models”. In: *Advances in neural information processing systems* 18 (2005).

- [215] R. Frigola, Y. Chen, and C. E. Rasmussen. “Variational Gaussian process state-space models”. In: *Advances in neural information processing systems* 27 (2014).
- [216] J. Liu and M. West. “Combined parameter and state estimation in simulation-based filtering”. In: *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 197–223.
- [217] Z. Li, N. B. Kovachki, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, A. Anandkumar, et al. “Fourier Neural Operator for Parametric Partial Differential Equations”. In: *International Conference on Learning Representations*. 2020.
- [218] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis. “Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators”. In: *Nature Machine Intelligence* 3.3 (2021), pp. 218–229.
- [219] M. C. Kennedy and A. O’Hagan. “Bayesian calibration of computer models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464.
- [220] D. Xiu and G. E. Karniadakis. “The Wiener–Askey polynomial chaos for stochastic differential equations”. In: *SIAM journal on scientific computing* 24.2 (2002), pp. 619–644.
- [221] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach*. Dover Publications, 2003.
- [222] D. J. Lucia, P. S. Beran, and W. A. Silva. “Reduced-order modeling: new approaches for computational physics”. In: *Progress in aerospace sciences* 40.1-2 (2004), pp. 51–117.
- [223] W. H. Schilders, H. A. Van der Vorst, and J. Rommes. *Model order reduction: theory, research aspects and applications*. Vol. 13. Springer, 2008.
- [224] T. De Ryck and S. Mishra. “Generic bounds on the approximation error for physics-informed (and) operator learning”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 10945–10958.
- [225] S. Goswami, A. Bora, Y. Yu, and G. E. Karniadakis. “Physics-informed deep neural operator networks”. In: *Machine Learning in Modeling and Simulation: Methods and Applications*. Springer, 2023, pp. 219–254.
- [226] Z. Li, H. Zheng, N. Kovachki, D. Jin, H. Chen, B. Liu, K. Azizzadenesheli, and A. Anandkumar. “Physics-informed neural operator for learning partial differential equations”. In: *ACM/JMS Journal of Data Science* 1.3 (2024), pp. 1–27.
- [227] H. Owhadi. “Computational graph completion”. In: *Research in the Mathematical Sciences* 9.2 (2022), p. 27.
- [228] T. Bourdais, P. Battie, X. Yang, R. Baptista, N. Rouquette, and H. Owhadi. “Codiscovering graphical structure and functional relationships within data: A Gaussian Process framework for connecting the dots”. In: *Proceedings of the National Academy of Sciences* 121.32 (2024), e2403449121.

- [229] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- [230] I. Steinwart. *Support Vector Machines*. Springer, 2008.
- [231] C. Ma, L. Wu, et al. “The Barron space and the flow-induced function spaces for neural network models”. In: *Constructive Approximation* 55.1 (2022), pp. 369–406.
- [232] D. Voina, S. Brunton, and J. N. Kutz. “Deep Generative Modeling for Identification of Noisy, Non-Stationary Dynamical Systems”. In: *arXiv preprint arXiv:2410.02079* (2024).
- [233] D. A. Messenger and D. M. Bortz. “Weak SINDy for partial differential equations”. In: *Journal of Computational Physics* 443 (2021), p. 110525.
- [234] P. Batlle, Y. Chen, B. Hosseini, H. Owhadi, and A. M. Stuart. “Error analysis of kernel/GP methods for nonlinear and parametric PDEs”. In: *Journal of Computational Physics* 520 (2025), p. 113488.
- [235] C. Mora, A. Yousefpour, S. Hosseinmardi, H. Owhadi, and R. Bostanabad. “Operator learning with Gaussian processes”. In: *Computer Methods in Applied Mechanics and Engineering* 434 (2025), p. 117581.
- [236] F. Schäfer, M. Katzfuss, and H. Owhadi. “Sparse Cholesky Factorization by Kullback–Leibler Minimization”. In: *SIAM Journal on Scientific Computing* 43.3 (2021), A2019–A2046.
- [237] Y. Chen, E. N. Epperly, J. A. Tropp, and R. J. Webber. “Randomly pivoted Cholesky: Practical approximation of a kernel matrix with few entry evaluations”. In: *Communications on Pure and Applied Mathematics* (2023).
- [238] Y. Chen, H. Owhadi, and F. Schäfer. “Sparse Cholesky factorization for solving nonlinear PDEs via Gaussian processes”. In: *Mathematics of Computation* (2024).
- [239] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, 2018.
- [240] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.
- [241] H. Wendland and C. Rieger. “Approximate interpolation with applications to selecting smoothing parameters”. In: *Numerische Mathematik* 101 (2005), pp. 729–748.
- [242] F. J. Narcowich, J. D. Ward, and H. Wendland. “Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions”. In: *Constructive Approximation* 24 (2006), pp. 175–186.
- [243] Q. T. Le Gia, F. J. Narcowich, J. D. Ward, and H. Wendland. “Continuous and discrete least-squares approximation by radial basis functions on spheres”. In: *Journal of Approximation Theory* 143.1 (2006), pp. 124–133.
- [244] R. Arcangéli, M. C. López de Silanes, and J. J. Torrens. “An extension of a bound for functions in Sobolev spaces, with applications to (m, s) -spline interpolation and smoothing”. In: *Numerische Mathematik* 107.2 (2007), pp. 181–211.
- [245] C. Moriarty-Osborne and A. L. Teckentrup. “Convergence rates of non-stationary and deep Gaussian process regression”. In: *arXiv preprint arXiv:2312.07320* (2023).

- [246] G. Wynne, F.-X. Briol, and M. Girolami. “Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness”. In: *Journal of Machine Learning Research* 22.123 (2021), pp. 1–40.
- [247] A. Beck and A. Ben-Tal. “On the solution of the Tikhonov regularization of the total least squares problem”. In: *SIAM Journal on Optimization* 17.1 (2006), pp. 98–118.
- [248] A. Beck, A. Ben-Tal, and M. Teboulle. “Finding a global optimal solution for a quadratically constrained fractional quadratic problem with applications to the regularized total least squares”. In: *SIAM Journal on Matrix Analysis and Applications* 28.2 (2006), pp. 425–445.
- [249] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. “Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs”. In: *Journal of Machine Learning Research* 24.89 (2023), pp. 1–97.
- [250] S. Lanthaler, S. Mishra, and G. E. Karniadakis. “Error Estimates for DeepONets: A Deep Learning Framework in Infinite Dimensions”. In: *Transactions of Mathematics and Its Applications* 6.1 (2022), tnac001.
- [251] F. Schäfer and H. Owhadi. “Sparse recovery of elliptic solvers from matrix-vector products”. In: *SIAM Journal on Scientific Computing* 46.2 (2024), A998–A1025.
- [252] N. Boullé, D. Halikias, and A. Townsend. “Elliptic PDE learning is provably data-efficient”. In: *Proceedings of the National Academy of Sciences* 120.39 (2023), e2303904120.
- [253] B. P. Russo and M. P. Laiu. “Convergence of weak-SINDy surrogate models”. In: *SIAM Journal on Applied Dynamical Systems* 23.2 (2024), pp. 1017–1051.
- [254] N. B. Kovachki, S. Lanthaler, and A. M. Stuart. “Operator Learning: Algorithms and Analysis”. In: *arXiv:2402.15715* (2024).
- [255] R. A. Adams and J. J. Fournier. *Sobolev Spaces*. Elsevier, 2003.
- [256] L. C. Evans. *Partial differential equations*. American Mathematical Society, 2022.
- [257] F. Leibfried, V. Dutordoir, S. John, and N. Durrande. “A tutorial on sparse Gaussian processes and variational inference”. In: *arXiv preprint arXiv:2012.13962* (2020).
- [258] A. Y. Aravkin, R. Baraldi, and D. Orban. “A Levenberg–Marquardt method for nonsmooth regularized least squares”. In: *SIAM Journal on Scientific Computing* 46.4 (2024), A2557–A2581.
- [259] H. P. Gavin. “The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems”. In: *Department of Civil and Environmental Engineering Duke University August 3* (2019), pp. 1–23.
- [260] M. Kanagawa, P. Hennig, D. Sejdinovic, and B. K. Sriperumbudur. *Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences*. 2018. arXiv: 1807.02582 [stat.ML].
- [261] M. GEVERS. “A personal view of the development of system identification: A 30-year journey through an exciting field”. In: *IEEE Control Systems Magazine* 26.6 (Dec. 2006), pp. 93–105.

- [262] L. Ljung. *System identification (2nd ed.): theory for the user*. USA: Prentice Hall PTR, 1999.
- [263] G. L. Smith, S. F. Schmidt, and L. A. McGee. *Application of Statistical Filter Theory to the Optimal Estimation of Position and Velocity on Board a Circumlunar Vehicle*. Tech. rep. NASA-TR-R-135. NTRS Author Affiliations: NASA Ames Research Center NTRS Document ID: 20190002215 NTRS Research Center: Ames Research Center (ARC). Ames Research Center, Jan. 1, 1962.
- [264] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1970.
- [265] S. L. Brunton, J. L. Proctor, and J. Nathan Kutz. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems”. In: *Proceedings of the National Academy of Sciences* 113.15 (2016).
- [266] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, A. Ramadhan, and A. Edelman. *Universal differential equations for scientific machine learning*. Aug. 31, 2020.
- [267] M. Raissi. “Deep hidden physics models: Deep learning of nonlinear partial differential equations”. In: *Journal of Computational Physics* 357 (2018), pp. 125–141.
- [268] A. A. Kaptanoglu, B. M. de Silva, U. Fasel, et al. “PySINDy: A comprehensive Python package for robust sparse system identification”. In: *Journal of Open Source Software* 7.69 (2022), p. 3994.
- [269] L. Fung. *Overcoming error-in-variable problem in data-driven model discovery by orthogonal distance regression*. 2025. arXiv: 2507.23426 [stat.ME].
- [270] J. M. Hokanson, G. Iaccarino, and A. Doostan. “Simultaneous identification and denoising of dynamical systems”. en. In: *SIAM J. Sci. Comput.* 45.4 (4 Aug. 31, 2023), A1413–A1437.
- [271] W.-X. Wang, R. Yang, Y.-C. Lai, V. Kovanis, and C. Grebogi. “Predicting Catastrophes in Nonlinear Dynamical Systems by Compressive Sensing”. In: *Phys. Rev. Lett.* 106.15 (15 Apr. 2011), p. 154101.
- [272] D. L. Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.
- [273] E. J. Candès, J. Romberg, and T. Tao. “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”. In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509.
- [274] E. J. Candès and M. B. Wakin. “An introduction to compressive sampling”. In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 21–30.
- [275] R. Chartrand. “Numerical differentiation of noisy, nonsmooth data”. en. In: *ISRN Appl. Math.* 2011.1 (1 May 11, 2011), pp. 1–11.
- [276] J. M. Varah. “A spline least squares method for numerical parameter estimation in differential equations”. In: *SIAM J. Sci. Stat. Comput.* 3.1 (1 Mar. 1982), pp. 28–46.

- [277] J. M. Stevens-Haas, Y. Bhangale, J. Nathan Kutz, and A. Aravkin. “Learning Non-linear Dynamics Using Kalman Smoothing”. In: *IEEE Access* 12 (2024), pp. 138564–138574.
- [278] F. van Breugel, J. Nathan Kutz, and B. W. Brunton. “Numerical differentiation of noisy data: A unifying multi-objective optimization framework”. In: *IEEE Access* (2020).
- [279] H. Ribera, S. Shirman, A. V. Nguyen, and N. M. Mangan. “Model selection of chaotic systems from data with hidden variables using sparse data assimilation”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 32.6 (June 2022), p. 063101.
- [280] K. Lahouel. “Learning nonparametric ordinary differential equations from noisy data”. In: *Journal of Computational Physics* 479 (2024), p. 111019.
- [281] T. van Leeuwen and Y. Yang. “An analysis of constraint-relaxation in PDE-based inverse problems”. In: *Inverse Problems* 41.2 (2 Feb. 28, 2025), p. 025009.
- [282] T. v. Leeuwen and F. J. Herrmann. “A penalty method for PDE-constrained optimization in inverse problems”. In: *Inverse Problems* 32.1 (2016), p. 015007.
- [283] A. Y. Aravkin, D. Drusvyatskiy, and T. van Leeuwen. “Efficient Quadratic Penalization Through the Partial Minimization Technique”. In: *IEEE Transactions on Automatic Control* 63.7 (2018), pp. 2131–2138.
- [284] T. van Leeuwen and F. J. Herrmann. “Mitigating Local Minima in Full-Waveform Inversion by Expanding the Search Space”. In: *Geophysical Journal International* 195.1 (2013), pp. 661–667.
- [285] Z. Xu, D. Long, Y. Xu, G. Yang, S. Zhe, and H. Owhadi. “Toward Efficient Kernel-Based Solvers for Nonlinear PDEs”. In: *Proceedings of the 42nd International Conference on Machine Learning*. Ed. by A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu. Vol. 267. Proceedings of Machine Learning Research. PMLR, 13–19 Jul 2025, pp. 69526–69544.
- [286] D.-X. Zhou. “Derivative reproducing properties for kernel methods in learning theory”. en. In: *J. Comput. Appl. Math.* 220.1-2 (1-2 Oct. 2008), pp. 456–463.
- [287] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- [288] K. Levenberg. “A method for the solution of certain non-linear problems in least squares”. In: *Quarterly of applied mathematics* 2.2 (1944), pp. 164–168.
- [289] D. W. Marquardt. “An Algorithm for Least-Squares Estimation of Nonlinear Parameters”. In: *Journal of the Society for Industrial and Applied Mathematics* 11.2 (1963), pp. 431–441. eprint: <https://doi.org/10.1137/0111030>.
- [290] S. Wright, J. Nocedal, et al. “Numerical optimization”. In: *Springer Science* 35.67-68 (1999), p. 7.
- [291] R. Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), pp. 267–288.

- [292] P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin. “A Unified Framework for Sparse Relaxed Regularized Regression: SR3”. In: *IEEE Access* 7 (2019), pp. 1404–1423.
- [293] K. Champion, P. Zheng, A. Y. Aravkin, S. L. Brunton, and J. N. Kutz. “A unified sparse optimization framework to learn parsimonious physics-informed models from data”. In: *IEEE Access* 8 (2020), pp. 169259–169271.
- [294] D. Bertsimas and W. Gurnee. “Learning sparse nonlinear dynamics via mixed-integer optimization”. In: *Nonlinear Dynamics* 111.7 (Apr. 1, 2023), pp. 6585–6604.
- [295] L. M. Gao, U. Fasel, S. L. Brunton, and J. N. Kutz. *Convergence of uncertainty estimates in Ensemble and Bayesian sparse model discovery*. 2023. arXiv: 2301.12649 [cs.LG].
- [296] L. Mars Gao and J. Nathan Kutz. “Bayesian autoencoders for data-driven discovery of coordinates, governing equations and fundamental constants”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 480.2286 (Mar. 2024), p. 20230506. eprint: <https://royalsocietypublishing.org/rspa/article-pdf/doi/10.1098/rspa.2023.0506/1228812/rspa.2023.0506.pdf>.
- [297] K. Kaheman, S. L. Brunton, and J. Nathan Kutz. “Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data”. In: *Machine Learning: Science and Technology* 3.1 (Mar. 2022), p. 015031.
- [298] B. M. Bell. “The Iterated Kalman Smoother as a Gauss–Newton Method”. In: *SIAM Journal on Optimization* 4.3 (1994), pp. 626–636.
- [299] S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.
- [300] H. Owahdi. “Computational graph completion”. en. In: *Res. Math. Sci.* 9.2 (2 June 18, 2022), pp. 1–33.
- [301] R. S. Martin, C. M. Greve, C. E. Huerta, A. S. Wong, J. W. Koo, and D. Q. Eckhardt. “A robust time-delay selection criterion applied to convergent cross mapping”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 34.9 (Sept. 2024), p. 093110. eprint: https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/5.0209028/20143851/093110_1_5.0209028.pdf.
- [302] E. George, C. E. Chan, G. Dimand, R. M. Chakmak, C. Falcon, D. Eckhardt, and R. S. Martin. “Decomposing Signals from Dynamical Systems Using Shadow Manifold Interpolation”. In: *SIAM Journal on Applied Dynamical Systems* 20.4 (2021), pp. 2236–2260.
- [303] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz. “Chaos as an intermittently forced linear system”. In: *Nature Communications* 8.1 (May 30, 2017). Publisher: Nature Publishing Group, p. 19.
- [304] F. Takens. “Detecting strange attractors in turbulence”. In: *Dynamical Systems and Turbulence, Warwick 1980*. Ed. by D. Rand and L.-S. Young. Vol. 898. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, pp. 366–381.

- [305] J. Bakarji, K. Champion, J. Nathan Kutz, and S. L. Brunton. “Discovering governing equations from partial measurements with deep delay autoencoders”. en. In: *Proc. Math. Phys. Eng. Sci.* 479.2276 (2276 Aug. 30, 2023).
- [306] D. S. Broomhead and R. Jones. “Time-series analysis”. en. In: *Proc. R. Soc. Lond.* 423.1864 (1864 May 8, 1989), pp. 103–121.
- [307] S. Lee, F. Dietrich, G. E. Karniadakis, and I. G. Kevrekidis. “Linking Gaussian process regression with data-driven manifold embeddings for nonlinear data fusion”. en. In: *Interface Focus* 9.3 (3 June 6, 2019), p. 20180083.
- [308] S. Pan and K. Duraisamy. “On the structure of time-delay embedding in linear models of non-linear dynamical systems”. en. In: *Chaos* 30.7 (7 July 28, 2020), p. 073135.
- [309] G. Teschl. *Ordinary differential equations and dynamical systems*. Vol. 140. American Mathematical Soc., 2012.
- [310] L. Sun, D. Z. Huang, H. Sun, and J.-X. Wang. “Bayesian spline learning for equation discovery of nonlinear dynamics with quantified uncertainty”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., 2022.
- [311] S. H. Rudy, S. L. Brunton, and J. N. Kutz. “Smoothing and parameter estimation by soft-adherence to governing equations”. In: *Journal of Computational Physics* 398 (2019), p. 108860.
- [312] P. Kidger. “On Neural Differential Equations”. PhD thesis. University of Oxford, 2021.
- [313] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. “Neural Ordinary Differential Equations”. In: *Advances in Neural Information Processing Systems (NeurIPS 31)*. 2018, pp. 6572–6583.
- [314] D. Givoli. “A tutorial on the adjoint method for inverse problems”. en. In: *Comput. Methods Appl. Mech. Eng.* 380.113810 (113810 July 1, 2021), p. 113810.
- [315] J.-L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Springer-Verlag, 1971.
- [316] A. H. Ribeiro, K. Tiels, J. Umenberger, T. B. Schön, and L. A. Aguirre. “On the smoothness of nonlinear system identification”. In: *Automatica* 121 (2020), p. 109158.
- [317] D. Chakraborty, S. W. Chung, T. Arcomano, and R. Maulik. “Divide and conquer: Learning chaotic dynamical systems with multistep penalty neural ordinary differential equations”. In: *Computer Methods in Applied Mechanics and Engineering* 432 (2024), p. 117442.
- [318] F. Carbonell, Y. Iturria-Medina, and J. Jimenez. “Multiple Shooting-Local Linearization method for the identification of dynamical systems”. In: *Communications in Nonlinear Science and Numerical Simulation* 37 (2016), pp. 292–304.
- [319] S. W. Chung and J. B. Freund. “An optimization method for chaotic turbulent flow”. en. In: *J. Comput. Phys.* 457.111077 (111077 May 2022), p. 111077.

- [320] A. S. Householder. *On Prony's method of fitting exponential decay curves and multiple-hit survival curves*. Tech. rep. Oak Ridge National Laboratory, 1950.
- [321] A. Aravkin, J. V. Burke, L. Ljung, A. Lozano, and G. Pillonetto. “Generalized Kalman smoothing: Modeling and algorithms”. In: *Automatica* 86 (2017), pp. 63–86.
- [322] P. Hennig, M. A. Osborne, and H. P. Kersting. *Probabilistic numerics: Computation as machine learning*. en. Cambridge, England: Cambridge University Press, June 30, 2022. 410 pp.
- [323] Y. Saatçi. “Scalable inference for structured Gaussian process models”. PhD thesis. University of Cambridge, 2012.
- [324] D. P. O’Leary and B. W. Rust. “Variable projection for nonlinear least squares problems”. en. In: *Comput. Optim. Appl.* 54.3 (3 Apr. 2013), pp. 579–593.
- [325] G. H. Golub and V. Pereyra. “The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate”. en. In: *SIAM J. Numer. Anal.* 10.2 (2 Apr. 1973), pp. 413–432.
- [326] L. Fung, U. Fasel, and M. Juniper. “Rapid Bayesian identification of sparse nonlinear dynamics from scarce and noisy data”. en. In: *Proc. Math. Phys. Eng. Sci.* 481.2307 (2307 Feb. 2025).
- [327] A. A. Klishin, J. Bakarji, J. N. Kutz, and K. Manohar. “Statistical mechanics of dynamical system identification”. In: *Phys. Rev. Res.* 7.3 (3 Aug. 2025), p. 033181.
- [328] A. Argyriou and F. Dinuzzo. “A unifying view of representer theorems”. In: *ICML* 32.2 (2 June 21, 2014). Ed. by E. P. Xing and T. Jebara, pp. 748–756.
- [329] B. Schölkopf, R. Herbrich, and A. J. Smola. “A Generalized Representer Theorem”. In: *Computational Learning Theory*. Ed. by D. Helmbold and B. Williamson. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 416–426.
- [330] H. Wendland. *Scattered data approximation*. en. Cambridge monographs on applied and computational mathematics. Cambridge, England: Cambridge University Press, Feb. 11, 2010. 348 pp.
- [331] L. C. Evans. *Partial Differential Equations*. 2nd. Vol. 19. Graduate Studies in Mathematics. Providence, RI: American Mathematical Society, 2010.
- [332] A. Meurer, C. P. Smith, M. Paprocki, et al. “SymPy: symbolic computing in Python”. In: *PeerJ Computer Science* 3 (Jan. 2017), e103.
- [333] J. Bradbury, R. Frostig, P. Hawkins, et al. *JAX: Composable transformations of Python+NumPy programs*. <https://github.com/google/jax>. Version 0.6.0. 2018.
- [334] P. Kidger. *sympy2jax: Turn SymPy expressions into trainable JAX expressions*. <https://github.com/patrick-kidger/sympy2jax>. Version 0.0.7. Apache-2.0 license. 2025.
- [335] C. Tsitouras. “Runge–Kutta pairs of order 5 (4) satisfying only the first column simplifying assumption”. In: *Computers & Mathematics with Applications* 62.2 (2011), pp. 770–775.

- [336] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II Stiff and Differential-Algebraic Problems*. Second Revised Edition. Berlin: Springer, 2002.
- [337] G. Söderlind. “Automatic control and adaptive time-stepping”. In: *Numerical Algorithms* 31 (2002), pp. 281–310.
- [338] P. Kidger and C. Garcia. “Equinox: neural networks in JAX via callable PyTrees and filtered transformations”. In: *Differentiable Programming workshop at Neural Information Processing Systems 2021* (2021).
- [339] JSINDy. <https://github.com/AHsu98/jsindy>. 2025.
- [340] K. Burdzy and W. Werner. “A counterexample to the “Hot spots” conjecture”. In: *Annals of Mathematics* (1999), pp. 309–317.
- [341] S. Steinerberger. “An upper bound on the hot spots constant”. In: *Revista Matematica Iberoamericana* 39.4 (2023).
- [342] P. Mariano, H. Panzo, and J. Wang. “Improved upper bounds for the Hot Spots constant of Lipschitz domains”. In: *Potential Analysis* 59.2 (2023), pp. 771–787.
- [343] L. Brasco and A. Pratelli. “Sharp stability of some spectral inequalities”. In: *Geom. Funct. Anal.* 22.1 (2012), pp. 107–135.
- [344] G. Szegő. “Inequalities for Certain Eigenvalues of a Membrane of Given Area”. In: *Journal of Rational Mechanics and Analysis* 3 (1954), pp. 343–356.
- [345] H. F. Weinberger. “An Isoperimetric Inequality for the N -Dimensional Free Membrane Problem”. In: *Journal of Rational Mechanics and Analysis* 5 (1956), pp. 633–636.
- [346] J. Rauch. “Lecture #1. Five problems: An introduction to the qualitative theory of partial differential equations”. In: *Partial differential equations and related topics: Ford Foundation sponsored program at Tulane University, January to May, 1974*. Vol. 446. Lecture Notes in Mathematics. Springer, 1975, pp. 355–369.
- [347] K. Burdzy. “The hot spots problem in planar domains with one hole”. In: *Duke Mathematical Journal* 129 (2005), pp. 481–502.
- [348] J. d. Dios Pont. “Convex sets can have interior hot spots”. In: *arXiv preprint arXiv:2412.06344* (2024).
- [349] C. Judge and S. Mondal. “Euclidean triangles have no hot spots”. In: *Annals of Mathematics* 191.1 (2020), pp. 167–211.
- [350] C. Judge and S. Mondal. “Erratum: Euclidean triangles have no hot spots”. In: *Annals of Mathematics* 195.1 (2022), pp. 337–362.
- [351] R. Bañuelos and K. Burdzy. “On the hot spots conjecture of J. Rauch”. In: *Journal of Functional Analysis* 164.1 (1999), pp. 1–33.
- [352] R. Atar and K. Burdzy. “On Neumann eigenfunctions in lip domains”. In: *Journal of the American Mathematical Society* 17.2 (2004), pp. 243–265.
- [353] D. Krejčířík and M. Tušek. “Location of hot spots in thin curved strips”. In: *Journal of Differential Equations* 266.6 (2019), pp. 2953–2977.

- [354] D. Jerison and N. Nadirashvili. “The “hot spots” conjecture for domains with two axes of symmetry”. In: *Journal of the American Mathematical Society* 13.4 (2000), pp. 741–772.
- [355] M. Pascu. “Scaling coupling of reflecting Brownian motions and the hot spots problem”. In: *Transactions of the American Mathematical Society* 354.11 (2002), pp. 4681–4702.
- [356] B. Kawohl. *Rearrangements and Convexity of Level Sets in PDE*. Vol. 1150. Lecture Notes in Mathematics. Springer, 1985.
- [357] H. Chen, Y. Li, and L. Wang. “Monotone properties of the eigenfunction of Neumann problems”. In: *Journal de Mathématiques Pures et Appliquées* 130 (2019), pp. 112–129.
- [358] P.-f. Yang. “The hot spots conjecture on a class of domains in $\mathbb{R}^n$ with $n \geq 3$ ”. In: *Acta Mathematicae Applicatae Sinica, English Series* 27 (2011), pp. 639–646.
- [359] J. B. Kennedy and J. Rohleder. “On the hot spots conjecture in higher dimensions”. In: *arXiv preprint arXiv:2410.00816* (2024). arXiv: 2410.00816 [math.SP].
- [360] A. Kleefeld. “The hot spots conjecture can be false: Some numerical examples”. In: *Advances in Computational Mathematics* 47.6 (2021), p. 85.
- [361] M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and N. Nadirashvili. “The nodal line of the second eigenfunction of the Laplacian in $\mathbf{R}^2$ can be closed”. In: *Duke Math. J.* 90.3 (1997), pp. 631–640.
- [362] S. Fournais. “The nodal surface of the second eigenfunction of the Laplacian in $\mathbf{R}^D$ can be closed”. In: *J. Differential Equations* 173.1 (2001), pp. 145–159.
- [363] J. Dahne, J. Gómez-Serrano, and K. Hou. “A counterexample to Payne’s nodal line conjecture with few holes”. In: *Communications in Nonlinear Science and Numerical Simulation* 103 (2021), p. 105957.
- [364] J. B. Kennedy. “Closed nodal surfaces for simply connected domains in higher dimensions”. In: *Indiana University Mathematics Journal* (2013), pp. 785–798.
- [365] H. Berestycki, L. Nirenberg, and S. S. Varadhan. “The principal eigenvalue and maximum principle for second-order elliptic operators in general domains”. In: *Communications on Pure and Applied Mathematics* 47.1 (1994), pp. 47–92.
- [366] V. A. Marchenko and G. Suzikov. “The second boundary value problem in regions with a complex boundary”. In: *Matematicheskii Sbornik* 111.1 (1966), pp. 35–60.
- [367] A. Damlamian. “Le probleme de la passoire de Neumann”. In: *Rend. Sem. Mat. Univ. Politec. Torino* 43.3 (1985), pp. 427–450.
- [368] T. Del Vecchio. “The thick Neumann’s sieve”. In: *Ann. Mat. Pura Appl. (4)* 147 (1987), pp. 363–402.
- [369] A. Khrabustovskiy. “The Neumann sieve problem revisited”. In: *Journal of Mathematical Analysis and Applications* (2025), p. 129933.
- [370] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.