Probabilistic modeling and Bayesian inference via triangular transport

by

Ricardo Miguel Baptista
B.A.Sc., University of Toronto (2015)
S.M., Massachusetts Institute of Technology (2017)

Submitted to the Department of Aeronautics and Astronautics
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computational Science and Engineering
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
March 2022

Author .....
Department of Aeronautics and Astronautics
March 31, 2022

Certified by .....
Youssef Marzouk
Professor of Aeronautics and Astronautics
Thesis Supervisor

Certified by .....
Philippe Rigollet
Professor of Mathematics
Thesis Committee Member

Certified by .....
Alan Willsky
Edwin Sibley Webster Professor of Electrical Engineering
Thesis Committee Member

Accepted by .....
Jonathan P. How
R. C. Maclaurin Professor of Aeronautics and Astronautics
Chair, Graduate Program Committee

Accepted by .....
Nicolas Hadjiconstantinou
Professor of Mechanical Engineering
Co-Director, Center for Computational Science and Engineering

Probabilistic modeling and Bayesian inference via triangular transport

by

Ricardo Miguel Baptista

Submitted to the Department of Aeronautics and Astronautics
on March 31, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computational Science and Engineering

Abstract

Probabilistic modeling and Bayesian inference in non-Gaussian settings are pervasive challenges for science and engineering applications. Transportation of measure provides a principled framework for treating non-Gaussianity and for generalizing many methods that rest on Gaussian assumptions. A transport map deterministically couples a simple reference distribution (e.g., a standard Gaussian) to a complex target distribution via a bijective transformation. Finding such a map enables efficient sampling from the target distribution and immediate access to its density. Triangular maps comprise a general class of transports that are attractive from the perspectives of analysis, modeling, and computation. This thesis: (1) develops a general representation for monotone triangular maps, and adaptive methodologies for estimating such maps (and their associated pushforward densities) from samples; (2) uses triangular maps and their compositions to perform Bayesian computation in likelihood-free settings, including new ensemble methods for nonlinear filtering; and (3) proposes parameter and data dimension reduction techniques with error guarantees for high-dimensional inverse problems.

The first part of the thesis explores the use of triangular transport maps for density estimation and for learning probabilistic graphical models. To construct triangular maps, we represent monotone functions as smooth transformations of unconstrained (non-monotone) functions. We show how certain structural choices for these transformations lead to smooth optimization problems with no spurious local minima, i.e., where all local minima are global minima. Given samples, we then propose an adaptive algorithm that estimates maps with sparse variable dependence. We demonstrate how this framework enables joint and conditional density estimation across a range of sample sizes, and how it can explicitly learn the Markov properties of a continuous non-Gaussian distribution. To this end, we introduce a consistent estimator for the Markov structure based on integrated Hessian information from the log-density. We then propose an iterative algorithm for learning sparse graphical models by exploiting a corresponding sparsity structure in triangular maps.

A core advantage of triangular maps is that their components expose *conditionals*

of the target distribution. Hence, learning a map that depends on both parameters and observations enables efficient sampling from the posterior distribution in a Bayesian inference problem. Crucially, this can be done without evaluating the likelihood function, which is often inaccessible or computationally prohibitive in scientific applications (as with forward models given by stochastic partial differential equations, which we consider here). In the second part of this thesis, we propose and analyze a specific *composition* of transport maps that directly transforms prior samples into posterior samples. We show that this approach, termed the stochastic map (SM) algorithm, improves over other transport-based methods for conditional sampling by reducing the bias and variance of the associated posterior approximation. We then use the SM algorithm to sequentially estimate the state of a chaotic dynamical system given online observations, a nonlinear filtering problem known in geophysical applications as "data assimilation" (DA). We show that when the SM algorithm is restricted to linear maps, it reduces to the ensemble Kalman filter (EnKF), a workhorse algorithm for DA; with nonlinear updates, however, the SM algorithm substantially improves on the performance of the EnKF in challenging regimes.

Finally, we extend the use of transport for high-dimensional inference problems by developing a *joint* dimension reduction strategy for parameters and observations. We identify relevant low-dimensional projections of these variables by minimizing an information theoretic upper bound on the error in the posterior approximation. We show that this approach reduces to canonical correlation analysis in the linear–Gaussian setting, while outperforming standard dimension reduction strategies in a variety of nonlinear and non-Gaussian inference problems.

Thesis Supervisor: Youssef Marzouk
Title: Professor of Aeronautics and Astronautics

Thesis Committee Member: Philippe Rigollet
Title: Professor of Mathematics

Thesis Committee Member: Alan Willsky
Title: Edwin Sibley Webster Professor of Electrical Engineering

Acknowledgments

I am extremely grateful to my advisor, Youssef Marzouk, for his mentorship, advice and encouragement over the past few years. I will always cherish his academic guidance and the countless ways he has provided support on a personal level. His passion and perseverance towards answering meaningful scientific questions is an inspiration to me that I hope to carry forward to all of my future endeavors.

I would also like to thank the members of my thesis committee, Alan Willsky and Philippe Rigollet, and my thesis readers, Ann Lee and Dennis McLaughlin, who kindly volunteered their time to join me on this academic journey. I am very grateful for their insightful questions and invaluable feedback that greatly improved this thesis.

I owe a debt of gratitude to all of my collaborators and friends. I would like to sincerely thank Alessio Spantini, Olivier Zahm and Rebecca Morrison. This thesis would not have been possible without them. I would also like to thank Amir Sagiv, Bamdad Hosseini, Giulio Trigila, Muhammad Izzatullah, Jayanth Jagalur Mohan, Joshua Chen, Kevin Carlberg, Lianghao Cao, Mathieu Le Provost, Matthias Poloczek, Maximilian Ramgraber, Nikola Kovachki, Prasanth Nair, and Rahul Krishnan. I am very fortunate to have met all of them, and I appreciate the time and effort they expended to share their knowledge with me.

The Uncertainty Quantification (UQ) group and ACDL have been a home away from home over these years. I want to thank all past and present members, especially Andrea, Benjamin, Daniele, Elizabeth, Fengyi, Michael, Matthew, Nisha, Paul-Baptiste, and Sven. Thank you for the wonderful memories in between research. I would also like to thank Beth Marois and Jean Sofronas from AeroAstro, and Kate Nelson from CCSE for their kind and timely administrative support.

I am also sincerely grateful to the Air Force Office of Scientific Research, the Department of Energy AEOLUS center, and the Government of Canada NSERC PGS-D fellowship for generously supporting this research.

Finally, I would not be where I am today without the support of my family and friends back home. This thesis is dedicated to my parents for their unwavering love and encouragement to keep pursuing my goals.

Contents

Chapter 1

Introduction

1.1 Motivation

Probabilistic modeling and Bayesian inference are two problems at the core of many science and engineering studies. While modeling aims to characterize the probability distributions for observed data, Bayesian inference estimates model parameters and their uncertainty from noisy, and often indirect, measurements. Two features that are commonly present in both problems are high-dimensional variables and generic non-Gaussian structure. In many scientific problems, it is common to make various approximations to robustly characterize these distributions. For instance, Gaussian assumptions are commonly used for sequential inference of states with billions of degrees of freedom in numerical weather prediction [35].

Measure transport is one flexible and principled approach for representing non-Gaussianity and generalizing many Gaussian-based methods. Transport maps couple a complex target distribution with a simple reference distribution (e.g., a standard Gaussian) by seeking a transformation between the underlying random variables. Triangular transport maps are one special type of coupling that offers several computational advantages over other maps, such as those found via optimal transport [141]. In particular, triangular maps provide a tractable expression for the target density, and they can be estimated given a collection of i.i.d. samples. Furthermore, triangular maps inherit low-dimensional structure from the target distribution. For ex-

ample, [204] showed that triangular maps have sparse variable dependence when the random variables satisfy conditional independence, or Markov, properties.

Another key advantage of triangular maps is their components expose conditionals of the target distribution. This property enables simulating from and evaluating conditional densities, such as the posterior density for model parameters in a Bayesian inference problem. In doing so, we can solve Bayesian inference problems by learning a transport map given only samples from the joint distribution of parameters and observations. Most significantly, this learning process does not require evaluating the likelihood function that generated the observations, thereby making it a *likelihood-free* inference (LFI) method [47]. LFI is a popular technique in settings where the likelihood function is unavailable or computationally prohibitive. Key applications include parameter estimation in stochastic partial differential equations (PDEs) and in hidden Markov models with high-dimensional latent variables. Another class of related problems is to estimate the states of a dynamical systems as new observations arrive over time, which is known as "data assimilation" in geophysics.

A major obstacle to using measure transport techniques for these modeling and inference tasks, however, is to develop reliable estimators for high-dimensional distributions given limited samples. This is particularly important in scientific applications with computationally expensive forward models. This thesis sets forward the following goals: (1) to build a general framework for representing and optimizing monotone and triangular transport maps, (2) to develop sample-efficient transport map estimators that are tailored for probabilistic modeling and performing Bayesian computation, and (3) to identify low-dimensional structure in inference problems that can be exploited by transport-based inference algorithms.

The remainder of this thesis is organized as follows. Chapter 2 provides background material on measure transport and triangular maps, and can be skipped by a familiar reader. Chapter 3 analyzes and develops new methodologies for approximating transport maps and estimating the target density. Chapter 4 shows how to exploit sparse structure in transport maps to learn undirected graphical models that encode Markov properties. Chapters 5 and 6 propose novel transport map estimators for

solving likelihood-free inference and data assimilation problems, respectively. Chapter 7 proposes linear dimension reduction strategies for high-dimensional parameters and observations in Bayesian inverse problems. Lastly, concluding remarks are offered in Chapter 8. Proofs of the main results and additional details on each chapter are presented in Appendices A-F.

1.2 Thesis contributions and roadmap

This thesis covers several topics linked to measure transport. Here we preview the main contributions of each chapter.

Approximating triangular transport maps: Chapter 3 focuses on estimating monotone triangular maps S that push forward a target density π to the standard Gaussian density given i.i.d. samples $\{\boldsymbol{X}^i\}_{i=1}^n \sim \pi$. Popular parametric representations of monotone functions (e.g., normalizing flows) constrain the form of the map [161], thereby limiting their expressiveness, or use parameterizations that result in non-convex and non-smooth optimization problems. Here, we propose a novel representation for monotone functions where the kth map component $S_k(\boldsymbol{x}) = \mathcal{R}_k(f)(\boldsymbol{x})$ is written as the transformation of a (non-monotone) function f through a bijective operator $\mathcal{R}_k$. We provide conditions on $\mathcal{R}_k$ so that the resulting optimization problem for f is well-behaved, i.e., the objective is continuous and smooth, and there are no spurious local minima. This permits us to reliably identify such maps in practice using deterministic optimization techniques. We then propose a greedy procedure named Adaptive Transport Maps (`ATM`) to construct a basis (e.g., using polynomials or wavelets) for f that is adapted to each target density. We demonstrate the advantages of using `ATM` over non-adaptive procedures for several benchmark (conditional) density estimation problems and datasets.

Learning non-Gaussian graphical models: Chapter 4 is focused on learning the structure of undirected probabilistic graphical models. These graphs represent the

conditional independence properties of a collection of random variables. While many structure learning methods prescribe a parametric form for the joint or conditional density (e.g., as Gaussian or in a exponential family), these approaches do not consistently learn the graph given data generated from a model outside of this class. Furthermore, measures for assessing conditional independence in non-Gaussian distributions are often difficult to compute, especially in high-dimensions [206, 207]. This chapter introduces a new score matrix Ω that encodes pairwise conditional independencies of a continuous and non-Gaussian distribution. Each entry of the score matrix $\Omega_{ij} = \int |\partial_i \partial_j \log \pi(\boldsymbol{x})|^2 \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ is based on Hessian information of the log-density π and we prove that it bounds the conditional mutual information (between pairs of variables) for densities that satisfy a log-Sobolev inequality. We propose sample-based estimators for Ω based on iteratively learning a transport map given $\{\boldsymbol{X}^i\}_{i=1}^n \sim \pi$, and we show that adaptive threshold estimators for Ω are consistent for recovering the Markov structure of π as the number of samples $n \to \infty$.

Likelihood-free Bayesian inference via couplings: Measure-transport approaches for LFI learn a triangular map S depending on both parameters $\mathbf{X}$ and observations $\mathbf{Y}$ that expose the conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}$. We begin Chapter 5 by using the `ATM` algorithm to learn such maps for probabilistic model calibration of a polymer science model described by a stochastic PDE. Next, we propose a transformation T that directly pushes forward samples from the prior density $\pi_{\mathbf{X}}$ to the posterior $\pi_{\mathbf{X}|\mathbf{Y}}$ by using a composition of transport maps. We demonstrate that using T instead of S reduces the bias and variance of the resulting posterior approximation. Furthermore, we show that the composed map generalizes regression adjustment, a technique that is used to correct the mean and variance of approximate posterior samples [26]. We then propose a new information-theoretic optimization problem that is tailored for finding the map S when it used to build T. We conclude by generalizing triangular maps to monotone block-triangular maps for conditional sampling. Block-triangular maps are less affected by variable ordering, and we show that they converge to an optimal transport map that minimizes a transportation cost for moving samples. We

demonstrate the numerical performance of block-triangular maps to infer a high-dimensional random field for subsurface permeability.

Data assimilation via couplings: In Chapter 6 we consider the problem of sequential Bayesian inference in nonlinear and non-Gaussian state space models. In practice, sequential state estimation (i.e., filtering) is often performed using algorithms such as the ensemble Kalman filter (EnKF), which are inconsistent for recovering the Bayesian solution in the large-sample limit. Here, we develop nonlinear ensemble filtering algorithms based on the compositions of triangular transport maps from Chapter 5. These algorithm reduce to the EnKF when constraining the maps to be linear. We numerically demonstrate the advantage of sparse nonlinear maps over the stochastic EnKF for improved tracking and posterior moment estimation in representative chaotic dynamical systems. The algorithms in this chapter were also recently applied to estimate the turbulent flow downstream of an airfoil [123] given pressure measurements on the airfoil's surface. Using nonlinear filters resulted in higher fidelity recovery of the flow than the EnKF, even when using simplified physics-based models for the predictions.

Parameter and data dimension reduction for Bayesian inference: Lastly, we address of problem of constructing maps as functions of high, or possibly infinite, dimensional parameters and observations in Bayesian inverse problems. In Chapter 7 we propose a dimension reduction strategy that finds linear projections of the parameters and of the observations. This is particularly beneficial for statistical models where the observations are noisy and indirectly related to the parameters by a smoothing forward operator. In these cases, it has been shown that part of the parameters are uninformed by the observations, and the observations often contain redundant information. This permits us to closely approximate the posterior distribution by updating only the informed part of the parameters using a reduced set of informative observations [205, 50]. We identify the low-dimensional subspaces for the parameters and observations by minimizing a tractable upper bound for the

Kullback-Leibler divergence of the posterior approximation in expectation over the observations. This upper bound depends on integrated gradient information of the likelihood function and the basis vectors for the subspaces can be computed as the solution to (generalized) eigenvalue problems. We show that the solutions reduce to canonical correlation analysis (CCA) for inverse problems with linear forward models, and we demonstrate the numerical benefit of our approach over traditional techniques, like principal component analysis and CCA, for nonlinear forward models.

1.3 Published material and preprints

This thesis includes the following material that is published or in review:

1. Chapter 3 is based in part on: Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. "An adaptive transport framework for joint and conditional density estimation". In: *arXiv:2009.10303* (2020).

2. Chapter 4 is based in part on: Ricardo Baptista, Youssef Marzouk, Rebecca E Morrison, and Olivier Zahm. "Learning non-Gaussian graphical models via Hessian scores and triangular transport". In: *arXiv:2101.03093* (2021).

3. Chapter 5 is based in part on: Ricardo Baptista, Lianghao Cao, Joshua Chen, Omar Ghattas, Fengyi Li, Youssef Marzouk, and Tinsley Oden. "Statistical learning of models from diblock co-polymer images: Likelihood-free inference and information gain estimation via measure transport". In: *In preparation* (2022), and on: Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. "Conditional Sampling With Monotone GANs". In: *arXiv:2006.06755* (2020).

4. Chapter 6 is based in part on: Alessio Spantini, Ricardo Baptista, and Youssef Marzouk. "Coupling techniques for nonlinear ensemble filtering". In: *SIAM Review* (2022, In Press).

Other work by the author that is not presented in this thesis, but is related to its main themes, includes:

1. Rebecca Morrison, Ricardo Baptista, and Youssef Marzouk. "Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 2359–2369

2. Rebecca E Morrison, Ricardo Baptista, and Estelle L Basor. "Diagonal Nonlinear Transformations Preserve Structure in Covariance and Precision Matrices". In: *Journal of Multivariate Analysis* (2022, In Press)

3. Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff Eldredge. "A low-rank nonlinear ensemble filter for vortex models of aerodynamic flows". In: *AIAA Scitech 2021 Forum*. 2021, p. 1937

4. Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff D Eldredge. "A low-rank ensemble Kalman filter for elliptic observations". In: *arXiv:2203.05120* (2022)

Chapter 2

Background on measure transport

For a pair of measures $\boldsymbol{\nu}_\pi$ and $\boldsymbol{\nu}_\eta$ defined on $\mathbb{R}^d$, with densities π and η, respectively, a *coupling* is a pair of random variables $(\mathbf{X}, \mathbf{Z})$ which admit π and η as marginal densities [222]. One special kind of coupling is a deterministic coupling defined by a measurable function $S\colon \mathbb{R}^d \to \mathbb{R}^d$ that satisfies $\mathbf{Z} = S(\mathbf{X})$ in distribution [222]. We call the map S that satisfies this property a *transport map* because it transports mass from $\boldsymbol{\nu}_\pi$ to $\boldsymbol{\nu}_\eta$. This map is said to *push forward* $\boldsymbol{\nu}_\pi$ to $\boldsymbol{\nu}_\eta$, which is denoted by $S_\sharp \boldsymbol{\nu}_\pi = \boldsymbol{\nu}_\eta$. This means that $\boldsymbol{\nu}_\eta(A) = \boldsymbol{\nu}_\pi(S^{-1}(A))$ for any set A, or equivalently for any $\boldsymbol{\nu}_\eta$ integrable function ψ we have

$$\int \psi(\boldsymbol{z}) d\boldsymbol{\nu}_\eta(\boldsymbol{z}) = \int \psi(S(\boldsymbol{x})) d\boldsymbol{\nu}_\pi(\boldsymbol{x}).$$

We denote the push-forward condition in terms of densities as $S_\sharp \pi = \eta$.

One main feature of deterministic maps is that we can use them to easily generate samples from the two distributions. If $\{\boldsymbol{X}^i\}_{i=1}^n$ are independent and identically distributed (i.i.d.) samples from π, then $\{S(\boldsymbol{X}^i)\}_{i=1}^n$ are i.i.d. samples from η. In practice, we consider η to be a reference distribution that is easy to sample from (e.g., a standard Gaussian) and seek an invertible map S so that we can sample from π by applying the inverse map S^{-1} to samples from η, as seen in Figure 2-1.

Given two general measures, deterministic couplings are not always guaranteed to exist. For example, if $\boldsymbol{\nu}_\eta$ is a Dirac mass and $\boldsymbol{\nu}_\pi$ is not, then we can not find

a map that splits the probability mass. Under the condition that $\boldsymbol{\nu}_\pi$ and $\boldsymbol{\nu}_\eta$ are absolutely continuous with respect to the Lebesgue measure such that their densities exist, then there always exists a transport map S such that $\eta = S_\sharp \pi$ [189]. We define $S^\sharp \eta := (S^{-1})_\sharp \eta$ as the *pullback* density of η under the map S (see Figure 2-2). For diffeomorphisms S, the pullback density is given by the change of variables formula

$$S^\sharp \eta(\boldsymbol{x}) = \eta \circ S(\boldsymbol{x}) |\det \nabla S(\boldsymbol{x})|. \tag{2.1}$$

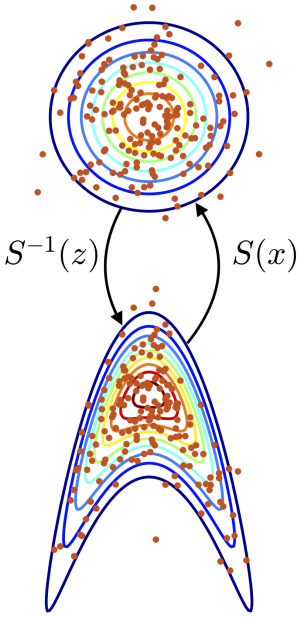


Figure 2-1: Mapping between samples

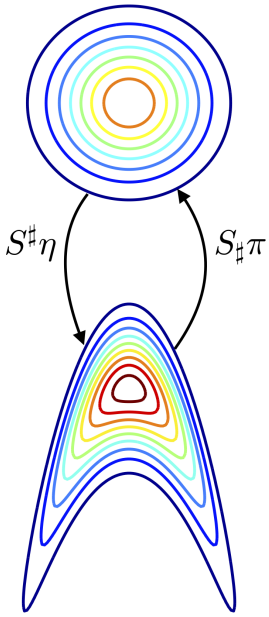


Figure 2-2: Mapping between densities

In addition to existence of the map, there may be infinitely many transport maps S that couple two probability distributions. For example, the Monge map is a transformation that solves the following problem:

$$\min_S \left\{ \int_{\mathbb{R}^d} c(\boldsymbol{x}, S(\boldsymbol{x})) \mathrm{d}\boldsymbol{\nu}_\pi(\boldsymbol{x}), \text{ s.t. } S_\sharp \boldsymbol{\nu}_\pi = \boldsymbol{\nu}_\eta \right\}, \tag{2.2}$$

where $c(\boldsymbol{x}, \boldsymbol{z})$ measures the cost of transporting one unit of mass from $\boldsymbol{x}$ to $\boldsymbol{z}$. This has served as the starting point for the field of optimal transport which characterizes these maps and their properties (e.g., their regularity). Many numerical and computational approaches have been developed to find the Monge map or to solve a relaxation

of problem (2.2), known as the Kantorovich problem (see [165] for comprehensive overview). For instance, [142] showed that the well-known Brienier map minimizes the quadratic cost $c(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{z}\|_2^2$ and is (uniquely) given by the gradient of a convex potential. We refer the reader to [222] for a description of other well-known transport maps.

In this thesis, we focus on a particular transport map between smooth and strictly positive densities π and η on $\mathbb{R}^d$[1]. While many deterministic maps satisfy the properties above, [185, 110] showed that there exists a lower triangular map $S\colon \mathbb{R}^d \to \mathbb{R}^d$ known as the Knothe-Rosenblatt (KR) rearrangement that pushes forward π to η. A monotone and lower triangular map is a multivariate function of the form

$$S(\boldsymbol{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ \vdots \\ S_d(x_1, x_2, \ldots, x_d) \end{bmatrix}, \tag{2.3}$$

where $\xi \to S_k(x_1, \ldots, x_{k-1}, \xi)$, i.e., the restriction of the kth map component to its last variable, is monotone increasing for all $(x_1, \ldots, x_{k-1}) \in \mathbb{R}^{k-1}$. Each component S_k in the KR rearrangement is defined using an iterative process of disintegrating the two measures. First, let $\boldsymbol{\nu}_{\pi_1}(\mathrm{d}x_1)$ and $\boldsymbol{\nu}_{\eta_1}(\mathrm{d}z_1)$ denote the marginals on the first variable. Then $z_1 = S_1(x_1)$ is defined as the increasing rearrangement that pushes forward $\boldsymbol{\nu}_{\pi_1}$ to $\boldsymbol{\nu}_{\eta_1}$. Next, for $k = 2, \ldots, d$, we define $z_k = S_k(x_1, \ldots, x_k)$ as the increasing rearrangement that pushes forward the marginal conditional $\boldsymbol{\nu}_{\pi_k}(\mathrm{d}x_k|x_1, \ldots, x_{k-1})$ to $\boldsymbol{\nu}_{\eta_k}(\mathrm{d}z_k|z_1, \ldots, z_{k-1})$. [27] showed that such a map is unique up to subsets of $\boldsymbol{\nu}_\pi$-measure zero, provided that the conditionals of $\boldsymbol{\nu}_\eta$ do not contain atoms. We refer to [27] for an exhaustive discussion on the regularity properties of the KR rearrangement. Recently it was shown in [33] that the KR rearrangement also corresponds to the limit in L^2_π of optimal transport maps S_ϵ as a parameter $\epsilon \to 0$. Each map S_ϵ is optimal with respect to the weighted quadratic cost $c_\epsilon(\boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^d \lambda_k(\epsilon)(x_k - z_k)^2$ where $\lambda_k(\epsilon)$ are positive scalars that satisfy $\lambda_k(\epsilon)/\lambda_{k+1}(\epsilon) \to 0$

[1] More generally, we only need the measure $\boldsymbol{\nu}_\pi$ to be absolute continuous with respect to $\boldsymbol{\nu}_\eta$.

for all $k \in \{1, \dots, d-1\}$.

Working with a *triangular* map such as the KR rearrangement confers several advantages. First, these maps can be easily inverted to sample $\boldsymbol{X} \sim \pi$ given a sample $\boldsymbol{Z} \sim \eta$ by solving a sequence of one-dimensional root-finding problems[2]. Second, for differentiable maps S, we can easily evaluate the Jacobian determinant of these transformations from the product of the partial derivatives of each component with respect to the last variable, i.e., $\det \nabla S(\boldsymbol{x}) = \prod_{k=1}^{d} \partial_k S_k(\boldsymbol{x}_{\leq k})$ where $\boldsymbol{x}_{\leq k} \coloneqq (x_1, \dots, x_k)$. The tractable evaluations of the log-determinant enables the pullback density in (2.1) to be easily evaluated without requiring an $\mathcal{O}(d^3)$ computation for each sample. Lastly, but most importantly for several applications in this thesis, triangular maps provide immediate access to the marginal conditional densities of π. For a tensor product reference measure $\boldsymbol{\nu}_\eta$—i.e., with a reference density that can be written as $\eta(\boldsymbol{x}) = \prod_{k=1}^{d} \eta_k(x_k)$—the kth component S_k pushes forward the kth marginal conditional density $\pi_{X_k|\mathbf{X}_{<k}}(x_k|\boldsymbol{x}_{<k})$ to the reference marginal $\eta_k(x_k)$ for all $\boldsymbol{x}_{<k} \coloneqq (x_1, \dots, x_{k-1}) \in \mathbb{R}^{k-1}$. As a result, we can equivalently express the marginal conditional density as

$$\pi_{X_k|\mathbf{X}_{<k}}(x_k|\boldsymbol{x}_{<k}) = \eta_k \circ S_k(\boldsymbol{x}_{\leq k}) |\partial_k S_k(\boldsymbol{x}_{\leq k})|. \tag{2.4}$$

We will use this last property in Chapter 3 to perform conditional density estimation and in Chapters 5 and 6 to characterize the posterior densities in inference problems.

One drawback of triangular maps is they are only unique up to a lexicographic ordering of the input variables $\boldsymbol{x}$. This order determines which marginal conditional measures are pushed forward by each component of S. Different choices for the ordering can affect the approximation and statistical estimation of the map components. For instance, some non-Gaussian measures have a low-dimensional non-Gaussian marginal that depends on $m \ll d$ variables while the remaining variables are conditionally Gaussian. Ordering the m non-Gaussian variables first yield's a

[2]For any $\boldsymbol{z} \in \mathbb{R}^d$, $\boldsymbol{x} = S^{-1}(\boldsymbol{z})$ can be computed recursively by $x_k = T_k(\boldsymbol{x}_{<k}, z_k)$ for $k = 1, \dots, d$, where $T_k(\boldsymbol{x}_{<k}, \cdot)$ is the inverse function of $x_k \mapsto S_k(\boldsymbol{x}_{<k}, x_k)$. Each root-finding problem has a unique solution from the bijectivity of map component S_k with respect to x_k.

map that is affine for the last $d - m$ components. If a natural ordering for the variables is unknown *a-priori*, Spantini, Bigoni, and Marzouk [204] select the ordering to maximize the sparsity of the map's variable dependence. In general, finding this ordering is equivalent to finding a permutation $\sigma\colon [d] \to [d]$, which is a NP-complete problem. In practice, it is common to use heuristics to find permutations or to select the ordering based on the sequential or spatial properties of the variables. We will not address this question, and will assume such an ordering is provided in this thesis.

We conclude this section with a canonical example of triangular maps.

Example 1 (Gaussian case)**.** *Suppose that* $\boldsymbol{X} \sim \pi = \mathcal{N}(\mathbf{0}, \Sigma)$ *is a Gaussian vector with non-singular covariance and* $\boldsymbol{Z} \sim \eta = \mathcal{N}(\mathbf{0}, I_d)$*. Let* $L^T L = \Sigma^{-1}$ *be the Cholesky decomposition of the inverse covariance matrix of* $\boldsymbol{X}$*. Then,* L *is a linear operator that maps samples* $\mathbf{X}^i \sim \pi$ *from the target density to samples* $\mathbf{Z}^i \sim \eta$ *from the reference density, and similarly,* L^{-1} *maps* $\mathbf{Z}^i$ *to* $\mathbf{X}^i$*. That is,*

$$L\mathbf{X}^i = \mathbf{Z}^i, \qquad L^{-1}\mathbf{Z}^i = \mathbf{X}^i.$$

Thus, $L\boldsymbol{x}$ *is an example of a linear lower triangular transport map* $S(\boldsymbol{x})$*. We note that other non-triangular transformations can also map* $\mathbf{X}$ *to* $\mathbf{Z}$ *(e.g., any inverse square root of* Σ^{-1}*), but* L *is the unique lower triangular map.*

More generally, affine transport maps are sufficient to represent Gaussian target distributions. On the other hand, nonlinear triangular maps S can be interpreted as the generalization of the Cholesky decomposition for a covariance matrix. These nonlinear maps represent arbitrary continuous and non-Gaussian target distributions. We use this generalization in the following chapters to develop algorithms that handle non-Gaussianity for density estimation (Chapter 3), learning graphical models (Chapter 4), likelihood-free Bayesian inference (Chapter 5) and sequential inference (Chapter 6).

Chapter 3

Approximating triangular transport maps

3.1 Introduction

In this chapter we study the approximation of monotone triangular transport maps S between continuous densities on $\mathbb{R}^d$. As compared to non-triangular maps, such as those found via optimal transport, triangular maps can be more easily parameterized and identified using maximum likelihood estimation given i.i.d. observations from π. In recent years, several parametric representations have been proposed for triangular maps. These include maps based on polynomials [141, 99], radial basis functions [211], and neural networks of varying capacity [56, 109]. Furthermore, multiple layers of triangular maps have been composed to define complex transformations known as normalizing flows [111, 161]. These transformations have been successfully used in several applications including density estimation [162]; variational inference [181, 28]; generation of images, video, and other structured objects [159, 109]; and likelihood-free inference [160].

In practice, a core challenge of approximating triangular transport maps is satisfying the monotonicity constraint. In [163], the constraint was enforced at a finite collection of samples from π. This approach results in a convex optimization problem under a linear expansion for the map components S_k. In a setting with few samples,

however, this is not sufficient to guarantee that S is monotone over the full support of π. Alternatively, many parametric forms for triangular transport maps or normalizing flows enforce monotonicity by making structural choices for each map component. For instance, [162] considered map components that have an affine dependence on the last variable, i.e., $S_k(\boldsymbol{x}_{\leq k}) = \alpha(\boldsymbol{x}_{<k}) + \exp(\beta(\boldsymbol{x}_{<k}))x_k$, where α and β are defined using neural networks. While S is guaranteed to be a monotone in this case, each map of this form can only represent densities π that are products of Gaussian marginal conditionals. To increase the "expressiveness" of these transport maps, several authors [227, 95, 62] have considered general parametric representations of monotone functions S_k at the trade-off of having to solve higher-dimensional and more complex optimization problems for the map parameters.

Despite the empirical success of different parameterizations, little work has addressed the properties of the optimization problems for learning transport maps given samples from π. Our contributions are as follows. We present a framework for representing and learning monotone triangular transports. Our approach relies on an operator that transforms broad classes of smooth functions into monotone functions. From a theoretical perspective, we show that the associated optimization problem is *smooth* under appropriate tail conditions on the target density; we also demonstrate that it has *no spurious local minima*, i.e., all local minima are global minima. Algorithmically, we then present an adaptive procedure for selecting appropriate smooth functions given a hierarchical basis of the corresponding function space. The procedure produces map representations that are *sparse* and *interpretable*—in particular, it yields map components S_k that are functions of a subset of the inputs $\boldsymbol{x}_{<k}$. As a result, the conditional densities $\pi_{X_k|\mathbf{X}_{<k}}(x_k|\boldsymbol{x}_{<k})$ in (2.4) don't depend on all inputs $\boldsymbol{x}_{<k}$. Hence, this procedure also exploits and implicitly discovers conditional independence. We use these learned maps for *density estimation* given only i.i.d. observations $\{\boldsymbol{X}^i\}$ drawn from π. Maintaining a strict triangular structure also exposes marginal conditionals of the target density, immediately enabling *conditional density estimation*. Our numerical experiments show that the algorithm provides robust performance at small-to-moderate sample sizes, and constitutes a *semiparametric* approach that links

map complexity to the size of the data.

The remainder of this chapter is organized as follows. Section 3.2 provides background information on triangular transport maps and the density estimation problem. Section 3.3 introduces our representation for monotone triangular maps and analyzes the resulting optimization problem. Section 3.4 presents the adaptive approximation algorithm for learning the map, and Section 3.5 provides numerical results from applying our algorithm on a series of test problems. Lastly, the appendices contain proofs of the theoretical results.

3.2 Triangular maps for density estimation

We consider the unsupervised learning problem of approximating a target probability density function π defined on $\mathbb{R}^d$ using i.i.d. samples from π. Our objective is to construct a sufficiently smooth and invertible map $S\colon \mathbb{R}^d \to \mathbb{R}^d$ such that the pullback density

$$S^\sharp\eta(\boldsymbol{x}) = \eta \circ S(\boldsymbol{x})|\det \nabla S(\boldsymbol{x})|, \tag{3.1}$$

is an approximation to π. Here, we choose $\eta(\boldsymbol{x}) \propto \exp(-\|\boldsymbol{x}\|^2/2)$, i.e., the probability density function (PDF) of the standard normal distribution on $\mathbb{R}^d$, where $\|\cdot\|$ is the canonical norm of $\mathbb{R}^d$. To ensure S is invertible, we constrain the map to be an increasing lower triangular function of the form in (2.3). We denote the *Knothe–Rosenblatt* (KR) rearrangement as the increasing lower triangular map S_{KR} such that

$$\pi(\boldsymbol{x}) = S_{\mathrm{KR}}^\sharp\eta(\boldsymbol{x}).$$

We measure the error between π and its approximation $S^\sharp\eta$ using the Kullback–Leibler (KL) divergence $\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta) = \int \log(\pi/S^\sharp\eta)\mathrm{d}\pi$. The following inequality is a direct consequence of Corollary 3.10 in [27] and the proof is given in Appendix A.1: for any map $S : \mathbb{R}^d \to \mathbb{R}^d$ as in (2.3) we have

$$\int \|S_{\mathrm{KR}}(\boldsymbol{x}) - S(\boldsymbol{x})\|^2 \mathrm{d}\pi(\boldsymbol{x}) \leq 2\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta). \tag{3.2}$$

This shows that convergence in the KL sense $\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta) \to 0$ implies convergence of S towards S_{KR} in the L^2_π sense. Given that the standard normal PDF η is a product of its marginal densities, the KL divergence decomposes as

$$\mathcal{D}_{\mathrm{KL}}(\pi||S^\sharp\eta) = \sum_{k=1}^{d} \mathcal{J}_k(S_k) - \mathcal{J}_k(S_{\mathrm{KR},k}), \tag{3.3}$$

where the objective functions $\mathcal{J}_1, \ldots, \mathcal{J}_d$ are given by

$$\mathcal{J}_k(s) = \int \left(\frac{1}{2} s(\boldsymbol{x}_{\leq k})^2 - \log |\partial_k s(\boldsymbol{x}_{\leq k})| \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{3.4}$$

The decomposition of (3.3) into d objective functions allows the components S_k to be computed independently, and hence in parallel [141]. This embarrassing parallelization was also exploited for Cholesky factorization via KL minimization in [191]. In addition, minimizing $s \mapsto \mathcal{J}_k(s)$ over functions $s \colon \mathbb{R}^k \to \mathbb{R}$ that are strictly increasing in the last variable is a convex optimization problem; see Appendix A.2 for details.

Given n i.i.d. samples $\{\boldsymbol{X}^i\}_{i=1}^n$ from $\pi(\boldsymbol{x})$, we replace the expectation in (3.4) by a sample average, which yields the empirical objective

$$\widehat{\mathcal{J}}_k(s) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} s(\boldsymbol{X}^i_{\leq k})^2 - \log \left| \partial_k s(\boldsymbol{X}^i_{\leq k}) \right| \right). \tag{3.5}$$

Minimizing (3.5) under the constraint $\partial_k s(\boldsymbol{x}_{\leq k}) > 0$ produces an estimator $\widehat{S}_k$ of $S_{\mathrm{KR},k}$. The collection of all components defines an estimator for the KR rearrangement $\widehat{S}$ that approximates the density $\pi(\boldsymbol{x})$ as $\widehat{\pi}(\boldsymbol{x}) \coloneqq \widehat{S}^\sharp\eta(\boldsymbol{x})$. Furthermore, convexity is conserved when replacing the expectation in (3.4) by a sample average.

A core challenge for minimizing (3.7) is how to parameterize a rich class of monotone map components S_k while allowing a computational efficient solution of the optimization problem. For instance, while map components with affine dependence on x_k can be found by the solution of a least-squares problem (see [203, Appendix A]), these structural forms for the map are limited to representing Gaussian marginal conditional densities. On the other hand, more complex forms for S_k often result in

optimization problems for the map parameters with lots of local minima, and thus hard to minimize. This is typically the case when using representations of S with nonlinear parameter dependence such as neural networks. In Section 3.3, we propose a representation for monotone functions that yields a smooth objective function with well-characterized local minima. We conclude this section by discussing how learning triangular maps are also useful to estimate and simulate from conditional densities.

Conditional density estimation One benefit of the triangle structure in (2.3) is that each component represents one marginal conditional density of π. More precisely, $S_{\mathrm{KR},k}$ pulls back the k-th marginal of the reference density $\eta_k(x_k)$ to the marginal conditional $\pi_k(x_k|\boldsymbol{x}_{<k})$. We use this property for the supervised learning problem of approximating the conditional probability density function $\pi(\boldsymbol{x}|\boldsymbol{y})$ given i.i.d. samples $\{(\mathbf{X}^i, \mathbf{Y}^i)\}_{i=1}^n$ from the joint density $\pi(\boldsymbol{x}, \boldsymbol{y})$. Here, $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^m$. For this task, we use an increasing lower triangular map $S\colon \mathbb{R}^{m+d} \to \mathbb{R}^{m+d}$ with the following structure

$$S(\boldsymbol{y}, \boldsymbol{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\boldsymbol{y}) \\ S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}) \end{bmatrix}, \tag{3.6}$$

where $S^{\mathcal{Y}}\colon \mathbb{R}^m \to \mathbb{R}^m$ and $S^{\mathcal{X}}(\boldsymbol{y}, \cdot)\colon \mathbb{R}^d \to \mathbb{R}^d$ are increasing lower triangular maps for any $\boldsymbol{y} \in \mathbb{R}^m$. Recall that the reference density $\eta(\boldsymbol{y}, \boldsymbol{x})$ is standard normal, and thus factorizes as $\eta(\boldsymbol{y}, \boldsymbol{x}) = \eta_1(\boldsymbol{y})\eta_2(\boldsymbol{x})$. The corresponding KR rearrangement $S_{\mathrm{KR}}(\boldsymbol{y}, \boldsymbol{x})$ as in (3.6) allows writing the marginal density as $\pi(\boldsymbol{y}) = (S^{\mathcal{Y}}_{\mathrm{KR}})^\sharp \eta_1(\boldsymbol{y})$ and, more interestingly, it exposes the conditional density as $\pi(\boldsymbol{x}|\boldsymbol{y}) = S^{\mathcal{X}}_{\mathrm{KR}}(\boldsymbol{y}, \cdot)^\sharp \eta_2(\boldsymbol{x})$.

The last d components of the KR rearrangement, $S^{\mathcal{X},k}_{\mathrm{KR}}(\boldsymbol{y}, \boldsymbol{x}_{\le k}), 1 \le k \le d$, can be estimated given independent samples from $\pi(\boldsymbol{x}, \boldsymbol{y})$ by minimizing

$$\widehat{\mathcal{J}}_k(s) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} s(\mathbf{Y}^i, \mathbf{X}^i_{\le k})^2 - \log \left| \partial_{m+k} s(\mathbf{Y}^i, \mathbf{X}^i_{\le k}) \right| \right), \tag{3.7}$$

under the constraints $\partial_{m+k} s(\boldsymbol{y}, \boldsymbol{x}_{\le k}) > 0$. This produces an estimator $\widehat{S}^{\mathcal{X}}$ of $S^{\mathcal{X}}_{\mathrm{KR}}$ which in turn approximates the conditional density $\pi(\boldsymbol{x}|\boldsymbol{y})$ as $\widehat{\pi}(\boldsymbol{x}|\boldsymbol{y}) \coloneqq \widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \cdot)^\sharp \eta_2(\boldsymbol{x})$. This property has been used to perform conditional density estimation (CDE) in [160,

113, 48].

One particular application of CDE is likelihood-free inference where $\mathbf{X}$ represents an unknown parameter and $\mathbf{Y}$ are data drawn from a likelihood function for $\mathbf{X}$ that is computationally expensive to evaluate (rendering standard likelihood-based inference methods inaccessible). Given a collection of joint samples $\{(\boldsymbol{Y}^i, \boldsymbol{X}^i)\}_{i=1}^n$, we can estimate the map component $S^{\mathcal{X}}$ and use it to simulate from the conditional density $\pi(\boldsymbol{x}|\boldsymbol{y}^*)$ given a specific realization of the data $\boldsymbol{y}^*$ by sampling $\boldsymbol{Z}^i \sim \eta_2$ and solving the system $S^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{X}^i) = \boldsymbol{Z}^i$ for each $\boldsymbol{X}^i$. Furthermore, we can sample from or evaluate the estimated conditional densities for multiple realizations of the data $\boldsymbol{y}^*$, that are not necessarily present in the dataset $\{\boldsymbol{Y}^i\}_{i=1}^n$. Thus, learning a single map $S^{\mathcal{X}}$ parameterized by $\boldsymbol{y}$ is said to *amortize* the cost of conditional sampling over the data. More details on likelihood-free inference using triangular maps will be presented in Chapter 5.

3.3 Representation and identification of triangular map components

In this section we define a general representation of monotone triangular maps. Each map component S_k is expressed as a nonlinear transformation $S_k = \mathcal{R}_k(f)$ of a smooth function f where $\mathcal{R}_k$ is an operator that enforces the monotonicity constraint by construction. We then find the map component by solving the re-parameterized problem

$$\min_{f \in V_k} \mathcal{L}_k(f) = \mathcal{J}_k(\mathcal{R}_k(f)), \tag{3.8}$$

where V_k is a linear space of functions in which we seek f. Using this transformation, we lose convexity of the constrained problem $\min_{\{s:\, \partial_k s > 0\}} \mathcal{J}_k(s)$, but we obtain an unconstrained minimization problem. In Section 3.3.1, we define the operator $\mathcal{R}_k$ and in Section 3.3.2 we discuss the regularity of the objective function $\mathcal{L}_k$ given a choice of the function space V_k. In Section 3.3.3 we discuss when solving (3.8) exactly recovers the KR rearrangement.

3.3.1 Representing continuous monotone functions

For any sufficiently smooth $f\colon \mathbb{R}^k \to \mathbb{R}$ we let $\mathcal{R}_k(f)\colon \mathbb{R}^k \to \mathbb{R}$ be the function defined by

$$\mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = f(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big)\mathrm{d}t, \tag{3.9}$$

where $g\colon \mathbb{R} \to \mathbb{R}_{>0}$ is a positive function. We call the operator $\mathcal{R}_k\colon f \mapsto \mathcal{R}_k(f)$ a *rectifier* because it transforms any f into a function which is increasing in the k-th variable, i.e., $\partial_k \mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = g(\partial_k f(\boldsymbol{x}_{\leq k})) > 0$. For instance, functions of the form $f(\boldsymbol{x}_{\leq k}) = \alpha(\boldsymbol{x}_{<k}) + \beta(\boldsymbol{x}_{<k})x_k$ are rectified into $\mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = \alpha(\boldsymbol{x}_{<k}) + g(\beta(\boldsymbol{x}_{<k}))x_k$. See Figure 3-1 for a numerical example of applying the rectifier to a non-monotone function f where the map S corresponds to the function that pushes forward the one-dimensional mixture of Gaussians density $\pi(x) = 0.5\mathcal{N}(x; -1, 1) + 0.5\mathcal{N}(x; 1, 1)$ into a standard Gaussian reference density η.

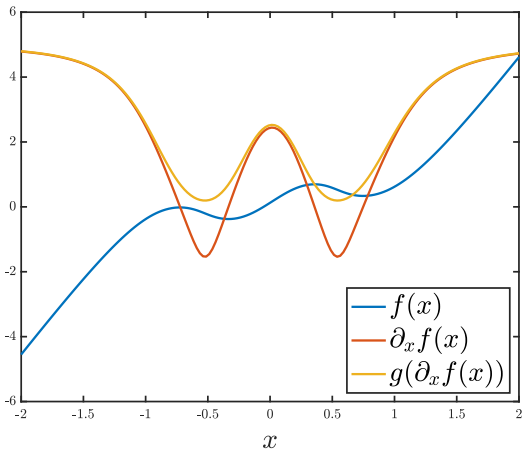


Figure 3-1: Function f and its derivative

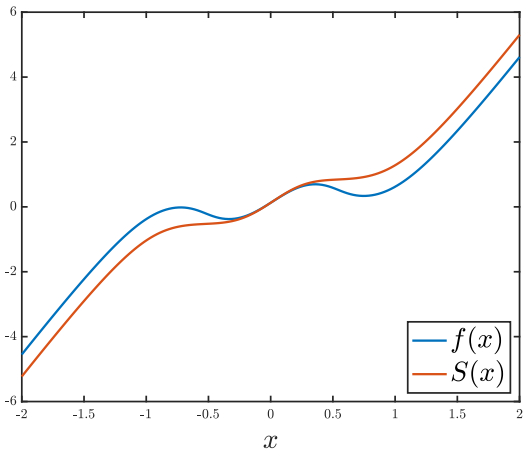


Figure 3-2: Functions f and $S = \mathcal{R}(f)$

The choice of the function g in (3.9) is essential regarding the optimization properties of (3.8). One possible choice, proposed in [99, 28], is the square function $g(\xi) = \xi^2$. While this choice permits the closed form computation of the integral in (3.9) when f is polynomial, it yields an optimization problem (3.8) which possesses many local minima, see Figure 3-3. This can be explained by the fact that this g is not bijective. Instead, we let g be a bijective function from $\mathbb{R}$ to $\mathbb{R}_{>0}$ such as the

soft-plus function

$$g(\xi) = \log(1 + \exp(\xi)), \tag{3.10}$$

whose inverse is $g^{-1}(\xi) = \log(\exp(\xi) - 1)$. Another example of a bijective function that was considered in [227] is the shifted exponential linear unit (ELU)

$$g(\xi) = \begin{cases} \exp(\xi) & \xi < 0 \\ \xi + 1 & \xi \geq 0 \end{cases}, \tag{3.11}$$

whose inverse is $g^{-1}(\xi) = \xi - 1$ if $\xi \geq 1$ and $g^{-1}(\xi) = \log(\xi)$ otherwise.

As a consequence of g being bijective, the inverse of the rectifier $\mathcal{R}_k^{-1}(s)$ exists for any sufficiently smooth $s\colon \mathbb{R}^k \to \mathbb{R}$ with $\partial_k s(\boldsymbol{x}_{\leq k}) > 0$ and

$$\mathcal{R}_k^{-1}(s)(\boldsymbol{x}_{\leq k}) = s(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g^{-1}\big(\partial_k s(\boldsymbol{x}_{<k}, t)\big)\mathrm{d}t. \tag{3.12}$$

More importantly, the fact that g is invertible yields an objective function $\mathcal{L}_k$ that is far better behaved in terms of the optimization procedure for f compared to $g(\xi) = \xi^2$, which is not invertible; see the numerical illustration in Figure 3-3. We observe that $\mathcal{L}_k = \mathcal{J}_k \circ \mathcal{R}_k$, though non-convex in general, has no local minima nor saddle points. In the next sections, we will analyze the properties (smoothness, uniqueness of the minimizer, etc.) of the optimization problem (3.8) depending on the choices made for g and V_k.

Remark. *It can be easily shown that composing a smooth (and strictly) convex objective function with a $\mathcal{C}^1$-diffeomorphic map preserves the (unique) global minima of the objective. Such diffeomorphic maps have been explored for accelerating optimization on finite-dimensional manifolds, e.g. see [127]. Establishing that $\mathcal{R}_k$ in (3.9) is $\mathcal{C}^1$-diffeomorphic is, however, non-trivial. We will show in Section 3.3.3 that under some additional assumptions, our reparameterization yields an optimization problem where local minima correspond to global minima.*

Before showing the desirable properties of the optimization problem above, we provide an example when the objective function for the map is non-convex.

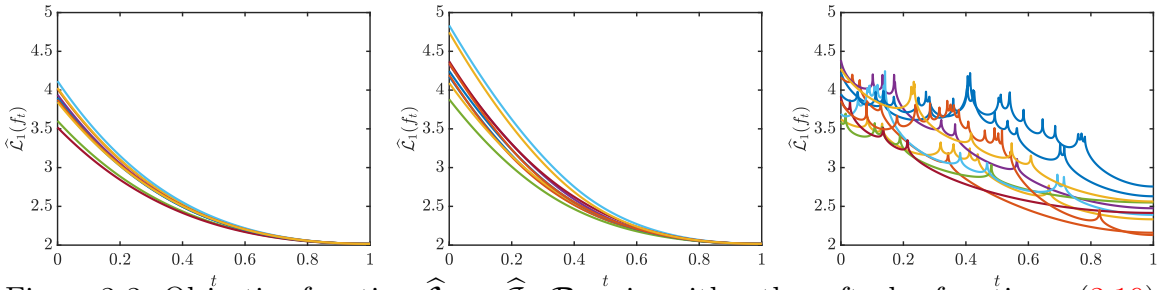


Figure 3-3: Objective function $\widehat{\mathcal{L}}_1 = \widehat{\mathcal{J}}_1 \circ \mathcal{R}_1$ using either the soft-plus function g (3.10) (*left*), the shifted exponential linear unit (*middle*), or the square function $g(\xi) = \xi^2$ (*right*). Here, $\pi(x) = 1/2\mathcal{N}(x; -2, 0.5) + 1/2\mathcal{N}(x; 2, 2)$ is a one-dimensional Gaussian mixture and we use $n = 50$ samples to estimate $\mathcal{J}_1$ with f_1 represented using a linear combination of Hermite functions up to degree 10. The objective is evaluated along line segments that interpolate between random initial maps ($t = 0$) and critical points resulting from a gradient-based optimization method ($t = 1$). Observe also that with the bijective functions g (*left* and *middle*) the algorithm always arrives at the same optimal value (at $t = 1$), whereas with the square function g (*right*) the algorithm gets stuck in local minima and rarely attains the optimal value.

Example 2. *Let π be a one-dimensional target density and consider the optimization problem for the map component $S(x) = \mathcal{R}_1(f)(x)$ where $f\colon \mathbb{R} \to \mathbb{R}$ is an affine function of the form $f(x) = \alpha_0 + g(\alpha_1)x$ with parameters $\alpha_0, \alpha_1 \in \mathbb{R}$. In this case, the objective function in (3.4) for the parameters is*

$$\mathcal{L}_1(\alpha_0, \alpha_1) = \mathbb{E}_\pi \left[\frac{1}{2} \left(\alpha_0 + g(\alpha_1)x \right)^2 - \log |g(\alpha_1)| \right] \tag{3.13}$$

To examine the convexity of the objective function, we check under which conditions the Hessian matrix $\nabla^2_{\boldsymbol{\alpha}} \mathcal{L}_1$ is positive definite for all parameter values. The second partial derivatives of the objective function are given by

$$\begin{aligned} \partial^2_{\alpha_0} \mathcal{L}_1 &= \mathbb{E}_\pi [1] = 1 \\ \partial_{\alpha_0, \alpha_1} \mathcal{L}_1 &= \mathbb{E}_\pi [g'(\alpha_1)x] \\ \partial^2_{\alpha_1} \mathcal{L}_1 &= \mathbb{E}_\pi \left[(g'(\alpha_1)x)(g'(\alpha_1)x) + (\alpha_0 + g(\alpha_1)x)(g''(\alpha_1)x) \right. \\ &\qquad \left. - \frac{g''(\alpha_1)}{|g(\alpha_1)|} + \frac{(g'(\alpha_1))^2}{|g(\alpha_1)|^2} \right] \end{aligned} \tag{3.14}$$

Sylvester's criterion provides a necessary and sufficient condition for the Hessian to be positive definitive by checking that all principal minors are positive. While the first

principal matrix $\partial^2_{\alpha_0}\mathcal{L} = 1$ *is positive, the determinant of the second principal matrix is given by*

$$\begin{aligned}
\det(\nabla^2_{\boldsymbol{\alpha}}\mathcal{L}_1) &= \partial^2_{\alpha_1}\mathcal{L}[\alpha_0,\alpha_1]\partial^2_{\alpha_0}\mathcal{L}[\alpha_0,\alpha_1] - (\partial_{\alpha_0,\alpha_1}\mathcal{L}[\alpha_0,\alpha_1])^2 \\
&= \mathbb{E}_\pi\left[(g'(\alpha_1)x)(g'(\alpha_1)x) + (\alpha_0 + g(\alpha_1)x)(g''(\alpha_1)x)\right] \\
&\quad - \mathbb{E}_\pi\left[g'(\alpha_1)x\right]^2 + \mathbb{E}_\pi\left[-\frac{g''(\alpha_1)}{|g(\alpha_1)|} + \frac{(g'(\alpha_1))^2}{|g(\alpha_1)|^2}\right] \\
&= Cov_\pi[g'(\alpha_1)x] + \alpha_0 g''(\alpha_1)\mathbb{E}_\pi[x] \\
&\quad + g(\alpha_1)g''(\alpha_1)\mathbb{E}_\pi[x^2] + \mathbb{E}_\pi\left[-\frac{g''(\alpha_1)}{|g(\alpha_1)|} + \frac{(g'(\alpha_1))^2}{|g(\alpha_1)|^2}\right].
\end{aligned} \tag{3.15}$$

For the soft-plus function $g(x) = \log(1+\exp(x))$*, the third and fourth terms in* (3.15) *satisfy*

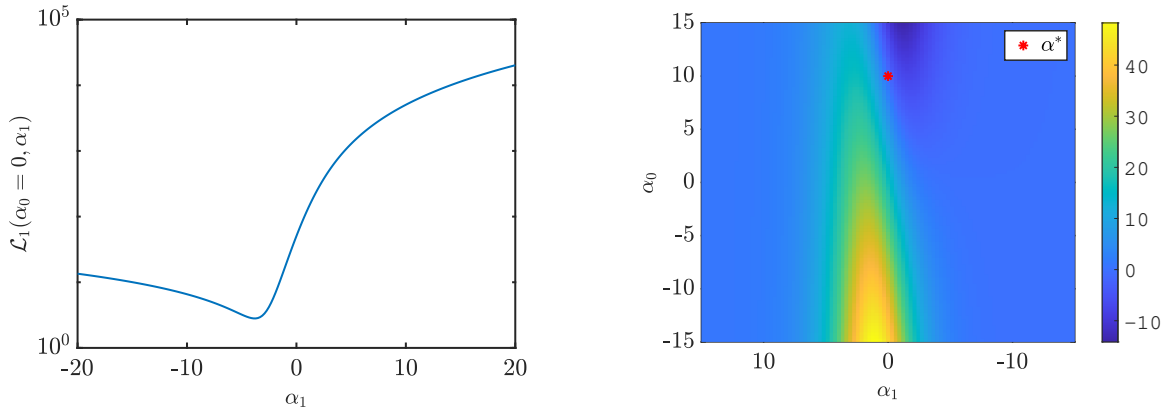
$$g(\alpha_1)g''(\alpha_1)\mathbb{E}_\pi[x^2] = \log(1+e^{\alpha_1})\frac{e^{-\alpha_1}}{(1+e^{-\alpha_1})^2}\mathbb{E}_\pi[x^2] > 0.$$

$$\frac{1}{|g(\alpha_1)|^2}\left[g'(\alpha_1))^2 - g''(\alpha_1)g(\alpha_1)\right] = \frac{1}{|g(\alpha_1)|^2}\left[\frac{1 - e^{-\alpha_1}\log(1+e^{\alpha_1})}{(1+e^{-\alpha_1})^2}\right] > 0.$$

Given that a covariance is always positive, the determinant in (3.15) *is guaranteed to be positive if* $\alpha_0 g''(\alpha_1)\mathbb{E}_\pi[x] > 0$ *for all values of* α_0 *and* α_1*. If* $\mathbb{E}_\pi[x] \neq 0$*, however, the principal matrix can be negative for a sufficiently large value for* $|\alpha_0|$*. As an example, Figure 3-4 plot a slice of the non-convex objective function* $\mathcal{L}_1$ *vs.* α_1 *and the determinant of the Hessian* $\nabla^2_{\boldsymbol{\alpha}}\mathcal{L}_1$ *for the Gaussian target density* $\pi(x) = \mathcal{N}(x; -10, 1)$*.*

3.3.2 Smoothness of the optimization problem

In this section we give sufficient conditions on the function g which guarantee that the objective function is smooth over an appropriate function space for f. We introduce the function space V_k and show the rectifier is continuous with respect to functions $f \in V_k$ before stating our main result in Proposition 2.



(a) (b)

Figure 3-4: (a) Slice of objective function $\mathcal{L}_1(0, \alpha_1)$ and (b) determinant of the Hessian matrix $\nabla^2_{\boldsymbol{\alpha}} \mathcal{L}_1$ for target density $\pi(x) = \mathcal{N}(x; -10, 1)$. Certain parameters yield a negative value for the determinant of the Hessian matrix and result in an objective function for the map that is non-convex.

We begin by defining the weighted Sobolev space

$$V_k = \left\{ f \colon \mathbb{R}^k \to \mathbb{R} \text{ such that } \|f\|^2_{V_k} := \int \Big(|f(\boldsymbol{x})|^2 + |\partial_k f(\boldsymbol{x})|^2 \Big) \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} < +\infty \right\}, \tag{3.16}$$

where $\eta_{\leq k}$ is the product of the first k marginals of η, i.e., the standard normal density on $\mathbb{R}^k$. By [114, Theorem 1.11], this space is complete and thus is a Hilbert space. The space V_k has the sufficient regularity for $\partial_k f$ to exist, but also to permit the pointwise evaluation $f(\boldsymbol{x}_{<k}, 0)$, as required in the definition (3.9) of the rectifier. This property is formalized by the following trace theorem which shows that, for any $f \in V_k$, the function $\boldsymbol{x}_{<k} \mapsto f(\boldsymbol{x}_{<k}, 0)$ is a function in $L^2_{\eta_{<k}}$, the weighted space of square integrable functions.

Theorem 3.3.1. *There exists a constant $C_T < \infty$ such that for any $f \in V_k$*

$$\int f(\boldsymbol{x}_{<k}, 0)^2 \, \eta_{<k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq C_T \|f\|^2_{V_k}. \tag{3.17}$$

Proof. See Appendix A.3. □

Remark. *Notice that $H^1_{\eta_{\leq k}} \subset V_k \subset L^2_{\eta_{\leq k}}$, where $L^2_{\eta_{\leq k}} = \{f \colon \mathbb{R}^k \to \mathbb{R} : \|f\|^2_{L^2_{\eta_{\leq k}}} :=$*

$\int f^2 \mathrm{d}\eta_{\leq k} < \infty\}$ *and* $H^1_{\eta_{\leq k}} = \{f \colon \mathbb{R}^k \to \mathbb{R} : \|f\|^2_{H^1_{\eta_{\leq k}}} := \int f^2 + \|\nabla f\|^2 \mathrm{d}\eta_{\leq k} < \infty\}$ *are the standard weighted Sobolev spaces. Because the standard normal density factorizes as* $\eta_{\leq k}(\boldsymbol{x}) = \prod_{i=1}^k \eta_i(x_i)$*, the space* V_k *admits the following tensor product structure*

$$V_k = L^2_{\eta_1} \otimes \ldots \otimes L^2_{\eta_{k-1}} \otimes H^1_{\eta_k}, \tag{3.18}$$

and the norm $\|\cdot\|_{V_k}$ *is a product norm*[1]*. This tensor product structure will be used later on in Section 3.4 to elaborate the numerical scheme for approximating the map components* S_k*.*

The following proposition shows that, under mild assumptions on g, the rectifier $\mathcal{R}_k$ is a Lipschitz continuous operator from V_k to V_k. The proof relies on the Hardy inequality [153] and on the Trace theorem 3.3.1.

Proposition 1. *Let* $g \colon \mathbb{R} \to \mathbb{R}_{>0}$ *be a Lipschitz function, i.e., there exists a constant* $L < \infty$ *so that*

$$|g(\xi) - g(\xi')| \leq L|\xi - \xi'|, \tag{3.19}$$

holds for any $\xi, \xi' \in \mathbb{R}$*. Then* $\mathcal{R}_k(f) \in V_k$ *for any* $f \in V_k$*, where* $\mathcal{R}_k(f)$ *is defined as in (3.9). Furthermore there exists a constant* $C < \infty$ *such that*

$$\|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{V_k} \leq C\|f_1 - f_2\|_{V_k}. \tag{3.20}$$

holds for any $f_1, f_2 \in V_k$*.*

Proof. See Appendix A.4. □

The next proposition shows that the objective function $\mathcal{L}_k$ in (3.8), seen as a function from V_k to $\mathbb{R}$, is well-defined, continuous and differentiable.

[1] That is $\|v_1 \otimes \cdots \otimes v_k\|_{V_k} = \|v_1\|_{L^2_{\eta_1}} \|v_2\|_{L^2_{\eta_2}} \cdots \|v_{k-1}\|_{L^2_{\eta_{k-1}}} \|v_k\|_{H^1_{\eta_k}}$ for any $v_j \in L^2_{\eta_k}$ if $j < k$ and $v_k \in H^1_{\eta_k}$.

Proposition 2. *Let π be a probability density function on $\mathbb{R}^d$ such that there exists a constant $C_\pi < \infty$ so that*

$$\pi(\boldsymbol{x}) \leq C_\pi \eta(\boldsymbol{x}), \tag{3.21}$$

for all $\boldsymbol{x} \in \mathbb{R}^d$. Furthermore, let $g\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ be an increasing function such that there exists a constant $L < \infty$ so that

$$|g(\xi) - g(\xi')| \leq L|\xi - \xi'|, \tag{3.22}$$

$$|\log \circ g(\xi) - \log \circ g(\xi')| \leq L|\xi - \xi'|, \tag{3.23}$$

holds for any $\xi, \xi' \in \mathbb{R}$. Then

$$\mathcal{L}_k(f) \coloneqq \mathcal{J}_k(\mathcal{R}_k(f)) < \infty,$$

for any $f \in V_k$, where $\mathcal{J}_k(s) = \int \left(\frac{1}{2} s(\boldsymbol{x}_{\leq k})^2 - \log|\partial_k s(\boldsymbol{x}_{\leq k})|\right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ as in (3.4) and where $\mathcal{R}_k(f)$ is defined as in (3.9). Moreover, the function $\mathcal{L}_k\colon V_k \to \mathbb{R}$ is continuous and differentiable at any $f \in V_k$ and its gradient $\nabla \mathcal{L}_k(f) \in V_k$ is given by

$$\begin{aligned}\langle \nabla \mathcal{L}_k(f), \varepsilon \rangle_{V_k} = &\int \mathcal{R}_k(f)(\boldsymbol{x}) \left(\varepsilon(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &- \int (\log \circ g)'(\partial_k f(\boldsymbol{x})) \partial_k \varepsilon(\boldsymbol{x}) \ \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3.24}$$

for any $\varepsilon \in V_k$, where $\langle \cdot, \cdot \rangle_{V_k}$ is the scalar product in V_k.

Proof. See Appendix A.5. □

It is useful to discuss some implications of Proposition 2. A smooth objective function permits us to use gradient-based first or second-order optimization algorithms. Furthermore, in Section 3.4 we exploit the gradients of $\mathcal{J}_k$ in (3.24) to construct an adaptive polynomial (or wavelet) basis for V_k. We note that several functions g satisfy the conditions (3.22) and (3.23) in the proposition. These include the soft-plus function in (3.10) and the shifted ELU function (3.11) considered in Figure 3-3.

Remark (Discussion of Assumption (3.21))**.** *The assumption (3.21) implies that*

$\mathbb{P}_{\mathbf{X}\sim\pi}(\|\mathbf{X}\|_2 > t) \leq C_\pi \mathbb{P}_{\mathbf{Z}\sim\eta}(\|\mathbf{Z}\|_2 > t)$ *holds for any* $t \geq 0$, *and hence that* π *has sub-Gaussian tails [220]. The reverse, however, does not hold in general. For instance, a compactly supported random variable with density* $\pi(x) \propto 1/\sqrt{|x|}\mathbb{1}_{\{x\in[-1,1]\}}$ *is sub-Gaussian, but the density is unbounded at* $x = 0$ *and thus does not satisfy* (3.21). *We also note that* (3.21) *can be relaxed with the (less-interpretable) assumption that there exists a constant* $C_\pi < \infty$ *and diagonal scalings* $d_k > 0$ *so that* $\pi(\boldsymbol{x}) \leq C_\pi \prod_{k=1}^d \eta_k(d_k x_k)$ *for all* $\boldsymbol{x} \in \mathbb{R}^d$. *In practice, we can always normalize the marginals of* π *to match the mean and variance of a standard Gaussian marginal reference. Thus, without loss of generality, we will use the assumption above for the remainder of this article.*

Remark. *Under additional assumptions on* g, *the gradient* $f \mapsto \nabla\mathcal{L}_k(f)$ *is locally Lipschitz from* $\overline{V}_k$ *to* V_k, *where* $\overline{V}_k = \{f \in V_k, \partial_k f \in L^\infty\}$ *is the space endowed with the norm* $\|f\|_{\overline{V}_k} = \|f\|_{V_k} + \|\partial_k f\|_{L^\infty}$. *More specifically, if* g' *and* $(\log\circ g)'$ *are Lipschitz, then there exists a constant* $M < \infty$ *such that*

$$\|\nabla\mathcal{L}_k(f_1) - \nabla\mathcal{L}_k(f_2)\|_{V_k} \leq M(1 + \|f_2\|_{V_k})\|f_1 - f_2\|_{\overline{V}_k}, \tag{3.25}$$

holds for any $f_1, f_2 \in \overline{V}_k$. *See the derivation of this result in Appendix A.6. Such local Lipschitz regularity is useful to analyze the convergence of backtracking gradient descent procedures, i.e., gradient descent with an inexact line search such as Armijo's rule, to a stationary point; see Theorem 2.1 in [218] and Proposition 2.1.1 in [19]. We leave the extension of such optimization guarantees for solving* $\min_{f\in V_k} \mathcal{L}_k(f)$ *to future work.*

3.3.3 Existence and uniqueness of solutions

In this section, we show that the optimization problem (3.8) does not admit any spurious local minima, meaning that local minimizers are in fact global minimizers. We also show that problem (3.8) admits a unique global minimizer which permits us to recover the KR rearrangement. To prove those results, we will need the following proposition which states sufficient conditions that ensure the inverse rectifier $\mathcal{R}_k^{-1}$ is

continuous.

Proposition 3. *Let g be a bijective function from $\mathbb{R}$ to $\mathbb{R}_{>0}$ such that for any $c > 0$ there exists a constant $L_c < \infty$ so that*

$$|g^{-1}(\xi) - g^{-1}(\xi')| \leq L_c|\xi - \xi'|, \tag{3.26}$$

holds for any $\xi, \xi' \geq c$. Then, for any $s \in V_k$ such that $\operatorname{ess\,inf} \partial_k s > 0$, we have $\mathcal{R}_k^{-1}(s) \in V_k$ and $\operatorname{ess\,inf} \partial_k \mathcal{R}_k^{-1}(s) > -\infty$. Furthermore for any $c > 0$, there exists a constant $C_c < \infty$ such that

$$\|\mathcal{R}_k^{-1}(s_1) - \mathcal{R}_k^{-1}(s_2)\|_{V_k} \leq C_c\|s_1 - s_2\|_{V_k}, \tag{3.27}$$

holds for any $s_1, s_2 \in V_k$ such that $\operatorname{ess\,inf} \partial_k s_i \geq c$.

Proof. See Appendix A.7. □

Let us remark that the softplus function (3.10) and the shifted exponential linear unit (3.11) satisfy (3.26) with $L_c = (1 - e^c)^{-1}$ and $L_c = \max\{1/c, 1\}$, respectively.

Local minima are global minima

The following proposition shows that, under certain conditions on g in (3.9), the image set $\mathcal{R}_k(V_k) = \{\mathcal{R}_k(f), f \in V_k\}$ is convex.

Proposition 4. *Let $g\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ be an increasing function such that $\xi \mapsto (g^{-1}(\xi))^2$ is a convex function. Let $\mathcal{R}_k(f)$ be defined as in (3.9) and V_k as in (3.16). Then $\mathcal{R}_k(V_k) = \{\mathcal{R}_k(f), f \in V_k\}$ is convex.*

Proof. See Appendix A.8. □

When g is the softplus function (3.10) or the exponential linear unit function (3.11) we have $\xi \mapsto (g^{-1}(\xi))^2$ is convex. However, the exponential function $g(\xi) = \exp(\xi)$ that has been used to parameterize monotone maps in [141, 174] does not satisfy this property, so there is no guarantee for $\mathcal{R}_k(V_k)$ to be convex in this case.

An important consequence of Proposition 4 is that the constrained optimization problem $\min_{s\in\mathcal{R}_k(V_k)} \mathcal{J}_k(s)$ remains convex. Hence, this constrained problem has a unique global minimizer by the strict convexity of $\mathcal{J}_k$ (see Appendix A.2). The next proposition shows that the local minima of the unconstrained problem in (3.8) are in fact global minima.

Proposition 5. *Under the assumptions of Proposition 4, let $f^* \in V_k$ be a local minimum of $f \mapsto \mathcal{J}_k(\mathcal{R}_k(f))$, meaning that there exists $\rho > 0$ such that*

$$\mathcal{J}_k(\mathcal{R}_k(f^*)) \leq \mathcal{J}_k(\mathcal{R}_k(f)), \quad \forall f \in V_k \text{ such that } \|f - f^*\|_{V_k} \leq \rho. \tag{3.28}$$

If $\operatorname{ess\,inf} \partial_k f^* > -\infty$ then f^* is a global minimum, meaning that

$$\mathcal{J}_k(\mathcal{R}_k(f^*)) \leq \mathcal{J}_k(\mathcal{R}_k(f)), \quad \forall f \in V_k.$$

Proof. Let $f \in V_k$. For any $0 < t < 1$ we let $s_t = t\mathcal{R}_k(f) + (1-t)\mathcal{R}_k(f^*)$. By Proposition 4, $\mathcal{R}_k(V_k)$ is convex so that $s_t \in \mathcal{R}_k(V_k)$. By convexity of $\mathcal{J}_k$ (*c.f.* Appendix A.2), we can write $\mathcal{J}_k(s_t) \leq t\mathcal{J}_k(\mathcal{R}_k(f)) + (1-t)\mathcal{J}_k(\mathcal{R}_k(f^*))$, or equivalently

$$\mathcal{J}_k(s_t) - \mathcal{J}_k(\mathcal{R}_k(f^*)) \leq t\Big(\mathcal{J}_k(\mathcal{R}_k(f)) - \mathcal{J}_k(\mathcal{R}_k(f^*))\Big). \tag{3.29}$$

Next we show that there exists a sufficiently small $t > 0$ such that the above left-hand side is positive. As a consequence, the right hand side of (3.29) will be positive, which will conclude the proof.

Let $M = \operatorname{ess\,inf} \partial_k f^*$ and $c = g(M)/2 > 0$. For any $t \leq 1/2$, as have $\partial_k s_t = tg(\partial_k f) + (1-t)g(\partial_k f^*) \geq 1/2 g(\partial_k f^*)$ so that $\operatorname{ess\,inf} \partial_k s_t \geq c$. In addition, we have $\operatorname{ess\,inf} \partial_k \mathcal{R}_k(f^*) = \operatorname{ess\,inf} g(\partial_k f^*) \geq c$. Thus, by Proposition 3, there exists a constant

$C_c < \infty$ such that

$$\begin{aligned}\|\mathcal{R}_k^{-1}(s_t) - f^*\|_{V_k} &= \|\mathcal{R}_k^{-1}(s_t) - \mathcal{R}_k^{-1}(\mathcal{R}_k(f^*))\|_{V_k} \\ &\leq C_c\|s_t - \mathcal{R}_k(f^*)\|_{V_k} \\ &= tC_c\|\mathcal{R}_k(f) - \mathcal{R}_k(f^*)\|_{V_k}.\end{aligned}$$

Thus, by letting $t = \rho/(C_c\|\mathcal{R}_k(f) - \mathcal{R}_k(f^*)\|_{V_k})$, we have $\|\mathcal{R}_k^{-1}(s_t) - f^*\|_{V_k} \leq \rho$. Therefore (3.28) with $f \leftarrow \mathcal{R}_k^{-1}(s_t)$ ensures that $0 \leq \mathcal{J}_k(s_t) - \mathcal{J}_k(\mathcal{R}_k(f^*))$. □

Uniqueness of the global minimizer and recovery of the KR rearrangement

As discussed earlier in Section 3.2, the Knothe–Rosenblatt rearrangement S_{KR} is the unique lower triangular and monotone map such that $D_{\mathrm{KL}}(\pi||S_{\mathrm{KR}}^\sharp\eta) = 0$; see [27]. The decomposition (3.3) of the KL divergence $D_{\mathrm{KL}}(\pi||S^\sharp\eta)$ thus permits us to write

$$\mathcal{J}_k(S_{\mathrm{KR},k}) \leq \mathcal{J}_k(\mathcal{R}_k(f)), \tag{3.30}$$

for any $f \in V_k$ and for any $1 \leq k \leq d$. Indeed, by letting S be the map such that $S_k = \mathcal{R}_k(f)$ and $S_i = S_{\mathrm{KR},i}$ for $i \neq k$, equation (3.3) gives (3.30). Thus, if there exists a function $f_{\mathrm{KR},k} \in V_k$ such that $S_{\mathrm{KR},k} = \mathcal{R}(f_{\mathrm{KR},k})$, then $f_{\mathrm{KR},k}$ is a global minimizer of $f \mapsto \mathcal{J}_k(\mathcal{R}_k(f))$ over $f \in V_k$. To show this, we first need the following intermediate result.

Proposition 6. *Let π be a probability density function on $\mathbb{R}^d$ such that there exists constants $0 < c, C < \infty$ so that*

$$c\eta(\boldsymbol{x}) \leq \pi(\boldsymbol{x}) \leq C\eta(\boldsymbol{x}), \tag{3.31}$$

for all $\boldsymbol{x} \in \mathbb{R}^d$. Then, for all $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$ and $k = 1, \ldots, d$, $S_{KR,k}(\boldsymbol{x}_{<k}, x_k) = \mathcal{O}(x_k)$ and $\partial_k S_{KR,k}(\boldsymbol{x}_{<k}, x_k) = \mathcal{O}(1)$ as $|x_k| \to \infty$. Furthermore, we have $S_{KR,k} \in V_k$ and $\operatorname{ess\,inf} \partial_k S_{KR,k} > 0$ for all $k = 1, \ldots, d$.

Proof. See Appendix A.9. □

Let us remark on the condition in (3.31). While Section 3.3.2 assumes only an upper bound on π, here we have both an upper and lower bound on the joint density. The upper and lower bounds together imply that any marginal conditional for $1 \leq k \leq d$ satisfies $c/C\eta_k(x_k) \leq \pi_k(x_k|\boldsymbol{x}_{<k}) \leq C/c\eta_k(x_k)$ for all $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. This condition implies the tails of the target density are Gaussian; rather than just being equal to or lighter than Gaussian under condition (3.31). Hence, the tails of the map components $S_{\mathrm{KR},k}$ are linear. It will be interesting to relax the lower bound in (3.31) while maintaining $S_{\mathrm{KR},k} \in V_k$ in future work.

We now combine the earlier propositions to show that under the assumption (3.31) on the target density π, we have $f_{\mathrm{KR},k} \in V_k$ and the existence of a unique solution to $\min_{f \in V_k} \mathcal{J}_k(\mathcal{R}_k(f))$.

Corollary 1. *Under the assumptions on π in Proposition 6, let g be Lipschitz bijection from $\mathbb{R}$ to $\mathbb{R}_{\geq 0}$ such that $\xi \mapsto (g^{-1}(\xi))^2$ is convex and such that, for any $c > 0$ there exists a constant $L_c < \infty$ such that $|g^{-1}(\xi) - g^{-1}(\xi')| \leq L_c|\xi - \xi'|$ for any $\xi, \xi' \geq c$. Then, for any $1 \leq k \leq d$, there exists a unique function $f_{KR,k} \in V_k$ that satisfies $\operatorname{ess\,inf} \partial_k f_{KR,k} > -\infty$ such that $\mathcal{R}_k(f_{KR,k})$ is the k-th component of the KR rearrangement S_{KR}. As a consequence, $\min_{f \in V_k} \mathcal{J}_k(\mathcal{R}_k(f))$ admits a unique solution and*

$$\mathcal{J}_k(\mathcal{R}_k(f_{KR,k})) = \min_{f \in V_k} \mathcal{J}_k(\mathcal{R}_k(f)).$$

Proof. Combining Proposition 6 and Proposition 3, there exists a function $f_{\mathrm{KR},k} \in V_k$ that satisfies $\operatorname{ess\,inf} \partial_k f_{\mathrm{KR},k} > -\infty$ such that $\mathcal{R}_k(f_{\mathrm{KR},k}) = S_{\mathrm{KR},k}$. Then from Proposition 5 and inequality (3.30), we have that $f_{\mathrm{KR},k}$ is a global minimizer of $f \mapsto \mathcal{J}_k(\mathcal{R}_k(f))$ over $f \in V_k$. From the uniqueness of the KR rearrangement $S_{\mathrm{KR},k}$ among monotone triangular maps and the injectivity of $\mathcal{R}_k^{-1}$, $f_{\mathrm{KR},k} = \mathcal{R}_k^{-1}(S_{\mathrm{KR},k})$ is unique. □

Lastly, we comment that if $f_{\mathrm{KR},k}$ is *not* contained in the function space V_k we consider in (3.16), then it is not possible to exactly recover the KR rearrangement by seeking $f \in V_k$. However, this does not imply that we cannot obtain a good approximation for π by solving the problem in (3.8); rather, it may affect the perfor-

mance of the approximation algorithm used to recover the minimizer f^*, which will be described in the next section.

3.4 Adaptive parameterization of transport maps

Given n i.i.d. samples $\{\boldsymbol{X}^i\}_{i=1}^n \sim \pi$, we now propose an adaptive algorithm to build $f \in V_k$ that minimizes the empirical objective $\widehat{\mathcal{L}}_k \coloneqq \widehat{\mathcal{J}}_k \circ \mathcal{R}_k(f)$, where $\widehat{\mathcal{J}}_k$ is as in (3.5). We construct f as a m-term expansion

$$f(\boldsymbol{x}_{\le k}) = \sum_{\boldsymbol{\alpha}\in\Lambda} c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\le k}), \tag{3.32}$$

where Λ is a set of indices with $\#\Lambda = m$ and where $\psi_{\boldsymbol{\alpha}} \colon \mathbb{R}^k \to \mathbb{R}$ are basis functions for V_k constructed as products of univariate functions

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}_{\le k}) = \prod_{j=1}^{k} \psi_{\alpha_j}^j(x_j).$$

Here, $\{\psi_\alpha^j\}_\alpha$ is a basis of either $L^2_{\eta_j}$ if $j < k$ or of $H^1_{\eta_k}$ if $j = k$. Because V_k possesses a tensor product structure (3.18) and because its norm $\|\cdot\|_{V_k}$ is a product norm, the basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}}$ is orthonormal if $\{\psi_\alpha^j\}_\alpha$ is an orthonormal basis for all $j \le k$. Given that f depends linearly on the coefficients $c_{\boldsymbol{\alpha}} \in \mathbb{R}$, the smoothness properties of the objective $\mathcal{L}_k$ derived in Section 3.3.2 (and demonstrated in Figure 3-3) transfer to the objective function over the coefficients $c_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \Lambda$.

Sections 3.4.1 and 3.4.2 describe two choices of basis functions, polynomials and wavelets, respectively. Then in Section 3.4.3 we present an greedy algorithm for building the multi-index set Λ. In addition, we propose a cross validation procedure in order to determine when to stop enriching Λ to avoid the approximation from over-fitting to the samples.

3.4.1 Polynomial space

Probabilist Hermite polynomials $\{\varphi_\alpha\}_{\alpha\in\mathbb{N}}$ form an orthogonal basis for L^2_η where $\eta(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ is the standard Gaussian density. The polynomial φ_α is of degree $\alpha \geq 0$ and is defined as

$$\varphi_\alpha(x) = (-1)^\alpha \exp(x^2/2)\frac{\mathrm{d}^\alpha}{\mathrm{d}x^\alpha}\exp(-x^2/2).$$

Furthermore, we have $\langle\varphi_\alpha,\varphi_\beta\rangle_{L^2_\eta} = \alpha!\,\delta_{\alpha,\beta}$ so that $\{\varphi_\alpha/\sqrt{\alpha!}\}_{\alpha\in\mathbb{N}_0}$ forms an orthonormal basis for L^2_η. Similarly, $\langle\varphi_\alpha,\varphi_\beta\rangle_{H^1_\eta} = (\alpha+1)!\,\delta_{\alpha,\beta}$ so that $\{\varphi_\alpha/\sqrt{(\alpha+1)!}\}_{\alpha\in\mathbb{N}_0}$ forms a basis for H^1_η; see [192, Proposition 1.3].

In practice, we modify the univariate Hermite polynomials to be linear outside of an arbitrary compact domain. Following [163], we let

$$\psi_{\alpha_j}(x_j) = \frac{1}{\sqrt{(\alpha_j+1)!}}\begin{cases}\varphi_{\alpha_j}(x_j^a) + \varphi'_{\alpha_j}(x_j^a)(x_j - x_j^a) & \text{if } x_j < x_j^a\\ \varphi_{\alpha_j}(x_j) & \text{if } x_j^a \leq x \leq x_j^b\\ \varphi_{\alpha_j}(x_j^b) + \varphi'_{\alpha_j}(x_j^b)(x_j - x_j^b) & \text{if } x_j > x_j^b\end{cases} \tag{3.33}$$

for each $j = 1,\ldots,k$. In our numerical experiments, we set x_j^a and x_j^b to be the 0.01 and 0.99 empirical quantiles of the one-dimensional samples $\{X_j^i\}_{i=1}^n$, respectively. The basis $\{\psi_{\alpha_j}\}_{\alpha_j\in\mathbb{N}_0}$, albeit not being exactly orthonormal, is close to being orthonormal for sufficiently small x_j^a and large x_j^b. Figure 3-5 display the univariate Hermite polynomials and linearized Hermite polynomials for $\alpha_1 \leq 10$.

For any $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)$ where

$$\alpha_j \in \mathbb{N}_0,$$

are non-negative integers, we let $\psi_{\boldsymbol{\alpha}} = \otimes_{j=1}^k \psi_{\alpha_j}$ be the tensorization of $\psi_{\alpha_1},\ldots,\psi_{\alpha_k}$. The basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}}$ generated by the piecewise polynomials (3.33) has numerical and structural advantages over Hermite polynomials. First, these functions avoid the fast growth rate of standard polynomials as $\|\boldsymbol{x}\| \to \infty$. This improves the numerical stability of map computations. Second, each map component $S_k = \mathcal{R}_k(f)$ for any

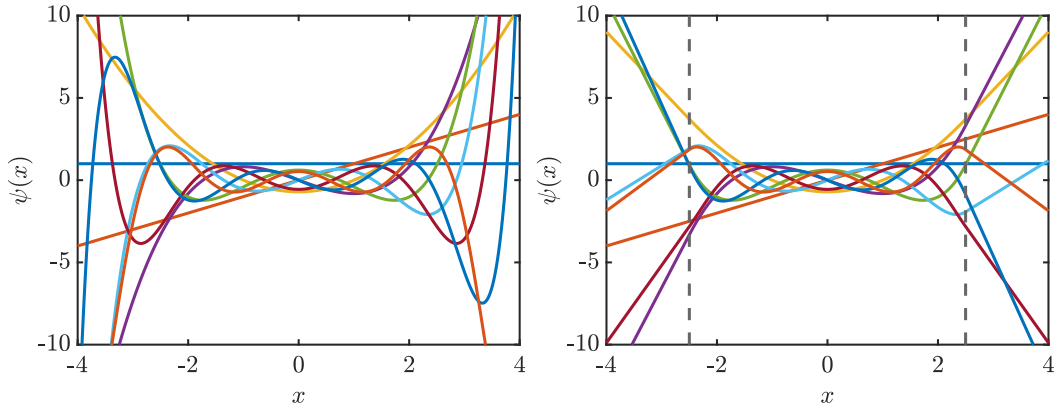


Figure 3-5: Hermite polynomials (a) and truncated Hermite polynomials (b) with linear extensions outside of $[-2.5, 2.5]$

function $f \in \text{span}\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Lambda\}$ behaves linearly as $x_k \to \infty$. The asymptotic linear growth yields a pullback density $S^\sharp \eta$ that has the same tail behavior as the Gaussian reference η. Thus, these functions can be seen as imposing a mild assumption on the tails of the target density π, which is consistent with the condition in (3.31) to guarantee there is no spurious local minima in the optimization problem for f. Furthermore, this assumption is necessary when given a finite set of samples that, in practice, contain limited information on the tails of the target density π. The linear growth assumption constrains the tails of the pullback density outside of the bounded domain containing the available samples[2].

3.4.2 Wavelet space

Wavelets are a popular tool for approximating functions that are not uniformly smooth [42, 221]. These techniques define a multi-resolution approximation to a function that can better capture local features than, for instance, global polynomial

[2]An alternative to linearized Hermite polynomials is Hermite functions that decay asymptotically to 0 as $x \to \infty$. [213] showed that L^2 approximations of functions defined on unbounded domains using a random set of design points is more stable with Hermite functions than standard Hermite polynomials that have unbounded growth. Furthermore, the authors showed that the condition numbers of the design matrices containing Hermite function evaluations are better conditioned. In our numerical experiments, we observed that Hermite functions tend to require adding more basis functions than linearized Hermite polynomials due to their behaviour of overfitting to individual samples.

basis functions. Given a compactly supported function $\psi\colon \mathbb{R} \to \mathbb{R}$ called the mother wavelet and $(m, l) \in \mathbb{Z}^2$, the function

$$\psi_{(l,m)} : x \mapsto 2^{l/2}\psi(2^l x - m),$$

is the mth wavelet at the level l. Common choices for the mother wavelet include the Haar, the Daubechies, and the Meyer wavelets. Given a mother wavelet $\psi \in L^2(\mathbb{R})$ whose Fourier transform satisfies some summability conditions described in [86] and in [97], it can be shown that $\{\psi_{(l,m)}\}_{(l,m)\in\mathbb{Z}^2}$ forms a basis for the weighted Sobolev space H^1_η (see Chapter 6 of [86] for unweighted Sobolev spaces).

In our numerical experiments, we use the continuous Mexican hat mother wavelet

$$\psi(x) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}}(1 - (x/\sigma)^2)\exp(-x^2/(2\sigma^2)), \tag{3.34}$$

with scale parameter $\sigma = 1$ [138], that is obtained from the second derivative of the univariate Gaussian density. The Mexican hat function is also commonly referred to as the Ricker wavelet in the geophysics community. We treat this function as having essentially compact support on $[-6, 6]$, up to numerical precision.

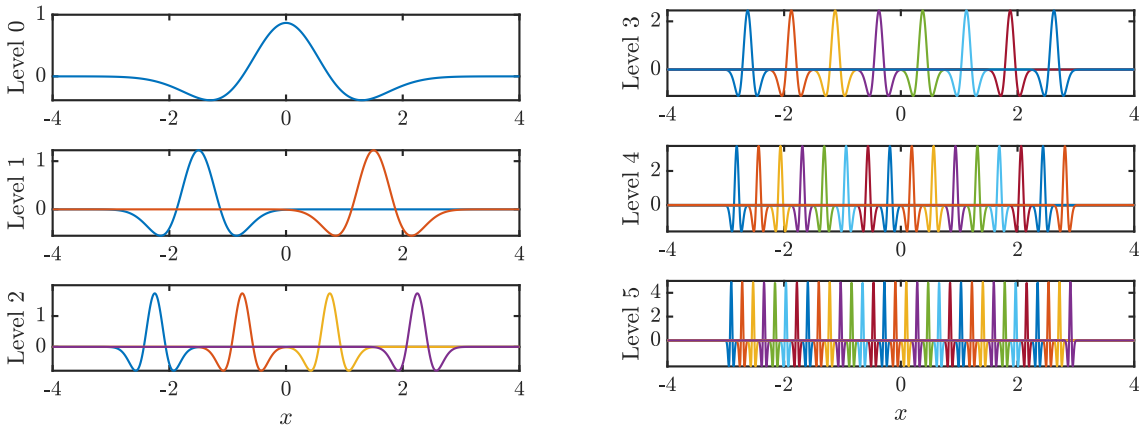
Tensorizing k wavelet bases yields a (not necessarily orthonormal) basis for V_k where each $\boldsymbol{\alpha}$ in the expansion (3.32) of f is a list of k tuples $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ containing two parameters

$$\alpha_j = (l_j, m_j) \in \mathbb{Z}^2. \tag{3.35}$$

In practice, we consider a truncated set of wavelets by first rescaling the mother wavelet ψ to have the same support as the interval between the 0.01 and 0.99 empirical quantiles of the marginal samples $\{X_j^i\}_{i=1}^n$. Then, we lower bound $l_j \geq 0$ and consider $l_j = 0$ to be the coarsest level of the approximation. Next, for any l_j, the translation parameter m_j is bounded as $0 \leq m_j \leq 2^{l_j} - 1$. Thus, we can consider

$$\alpha_j = (l_j, m_j) \in \mathbb{N}_0^2.$$

rather than (3.35). Figure 3-6 plots all Ricker wavelets for levels $1 \leq l_j \leq 6$.



(a) (b)

Figure 3-6: Ricker wavelets for levels $0 \leq l_j \leq 6$ with the mother wavelet rescaled to the domain $[-4, 4]$.

Lastly, let us notice that any function f in the form (3.32) expanded with compactly supported wavelets will decay to 0 as $|x_k| \to \infty$. Thus, the map $S_k = \mathcal{R}_k(f)$ will have asymptotic linear growth as $|x_k| \to \infty$, similarly to the linearized Hermite polynomial expansions.

3.4.3 Adaptive transport map algorithm

We propose a greedy construction of the multi-index set $\Lambda = \Lambda_t$ in (3.32). For simplicity, we consider only the case of single indices $\alpha_j \in \mathbb{N}_0$. The extension to the case $\alpha_j \in \mathbb{N}_0^2$ (encountered with the wavelet basis) is described in Appendix B.1. At each greedy iteration, we add one multi-index $\boldsymbol{\alpha}_t^*$ to Λ_t which best reduces the value of the objective function. Starting with $\Lambda_0 = \emptyset$ we let

$$\Lambda_{t+1} = \Lambda_t \cup \{\boldsymbol{\alpha}_t^*\}, \tag{3.36}$$

where $\boldsymbol{\alpha}_t^* \notin \Lambda_t$ is a multi-index which yields the best improvement of $\widehat{\mathcal{L}}_k$ in a sense to be defined later on. Borrowing ideas from [147, 146], we constrain the sets $\{\Lambda_t\}_{t \geq 0}$

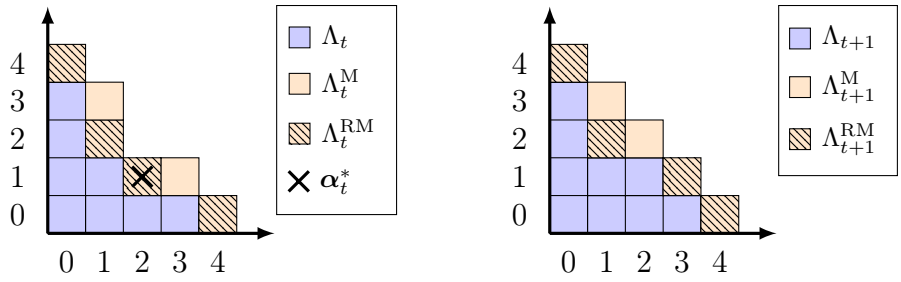


Figure 3-7: A $k = 2$ dimensional downward-closed active set of multi-indices Λ_t with its margin Λ_t^{M} and reduced margin Λ_t^{RM}. The margin and reduced margins are plotted before (*left*) and after (*right*) adding to $\boldsymbol{\alpha}_t^* = (2, 1)$, that is denoted with a cross, to the active set.

to be *downward-closed* [39, 43], meaning that they satisfy the property

$$\boldsymbol{\alpha} \in \Lambda_t \text{ and } \boldsymbol{\alpha}' \leq \boldsymbol{\alpha} \;\Rightarrow\; \boldsymbol{\alpha}' \in \Lambda_t, \tag{3.37}$$

where $\boldsymbol{\alpha}' \leq \boldsymbol{\alpha}$ means $\alpha_i' \leq \alpha_i$ for all $1 \leq i \leq k$. Intuitively, (3.37) means that Λ_t has a pyramidal shape that contains no hole. Downward-closed sets are known to preserve good approximation properties and permits a tractable construction of Λ_t, see [43]. Indeed, Λ_{t+1} remains downward-closed if and only if the multi-index $\boldsymbol{\alpha}_t^*$ is searched in the so-called *reduced margin* of Λ_t defined by

$$\Lambda_t^{\mathrm{RM}} = \{\boldsymbol{\alpha} \notin \Lambda_t \text{ such that } \boldsymbol{\alpha} - \mathbf{e}_i \in \Lambda_t \text{ for all } 1 \leq i \leq k \text{ with } \alpha_i \neq 0\},$$

where $\mathbf{e}_i$ denotes the i-th canonical vector of $\mathbb{N}^k$. The reduced margin is a subset of the margin set Λ_t^{M} (i.e., multi-indices $\boldsymbol{\alpha} \notin \Lambda_t$ such that $\exists\, i > 0$ where $\boldsymbol{\alpha} - \mathbf{e}_i \in \Lambda_t$); see Figure 3-7. The reduced margin typically grows more slowly with respect to the dimension k than the margin itself. For instance, if Λ_t contains all multi-indices in a hypercube of width p in dimension k, the margin has cardinality $(p+1)^k - p^k$, while the reduced margin has cardinality k.

Denoting by f_t the minimizer of $f \mapsto \widehat{\mathcal{L}}_k(f)$ over $f \in \mathrm{span}\{\psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \Lambda_t\}$, we select

$\boldsymbol{\alpha}_t^*$ in the reduced margin Λ_t^{RM} with the following heuristic

$$\boldsymbol{\alpha}_t^* \in \underset{\boldsymbol{\alpha} \in \Lambda_t^{\mathrm{RM}}}{\arg\max} \, |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)|. \tag{3.38}$$

Here, the notation $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)$ denotes the derivative of $c_{\boldsymbol{\alpha}} \mapsto \widehat{\mathcal{L}}_k(f_t + c_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}})$ evaluated at $c_{\boldsymbol{\alpha}} = 0$. In other words, we select the multi-index $\boldsymbol{\alpha}_t^*$ by choosing the largest functional derivative of $\widehat{\mathcal{L}}_k$ at f_t. This procedure for learning each map component is presented in detail in Algorithm 1.

Algorithm 1: Estimate map component S_k

1: **Input**: Training samples $\{\boldsymbol{X}_{\leq k}^i\}_{i=1}^n$, Maximum cardinality m for Λ_t
2: Initialize $\Lambda_0 = \emptyset$, $f_0 = 0$
3: **for** $t = 0, \ldots, m-1$ **do**
4: Construct the reduced margin: Λ_t^{RM}
5: Select the new multi-index: $\boldsymbol{\alpha}_t^* \in \arg\max_{\boldsymbol{\alpha} \in \Lambda_t^{\mathrm{RM}}} |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)|$
6: Update the active set: $\Lambda_{t+1} = \Lambda_t \cup \{\boldsymbol{\alpha}_t^*\}$
7: Update the approximation: $f_{t+1} = \arg\min_{f \in \mathrm{span}\{\psi_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \Lambda_{t+1}\}} \widehat{\mathcal{L}}_k(f)$
8: **end for**
9: **Output**: $\widehat{S}_k = \mathcal{R}_k(f_m)$

Remark. *If the objective function $\widehat{\mathcal{L}}_k$ is the linear least-squares loss, then Algorithm 1 corresponds to orthogonal matching pursuit for sparse linear regression with normalized basis functions. An alternative heuristic for selecting $\boldsymbol{\alpha}_t^*$ that requires second-order information of the objective is $\boldsymbol{\alpha}_t^* \in \arg\max_{\boldsymbol{\alpha} \in \Lambda_t^{RM}} |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}(f_t)|^2 / |\nabla_{\boldsymbol{\alpha}}^2 \widehat{\mathcal{L}}(f_t)|$. We found this criteria to perform similarly to (3.38) for the polynomial space, but it lead to faster convergence of the objective when working with the wavelet space. Both of these criteria can also be combined with an occasional backwards greedy step following step 7 in Algorithm 1 to remove multi-indices that do not reduce the combined bias and variance of the estimated map. See [241] for an adaptive forward-backward greedy algorithm in the context of sparse regression.*

Remark. *For convex objective functions with higher-order smoothness properties, [214] showed that a greedy algorithm using the criteria in (3.38) will identify the global minimizer, while avoiding a combinatorial optimization problem to find the optimal k-term*

approximation.

Remark. *The drawback of Algorithm 1 is that the greedy enrichment procedure does not see behind the reduced margin. For instance if a relevant multi-index is located far beyond the reduced margin and the gradient* $\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)$ *vanishes for all* $\boldsymbol{\alpha} \in \Lambda_t^{RM}$*, then the algorithm will be stuck. [146] suggested a safeguard mechanism to avoid this behaviour: arbitrarily activate the most ancient index from the reduced margin at every* l*th iteration for some* $l > 1$*. This modification, however, was not needed in our numerical experiments.*

Remark. *As pointed out in [43, 147, 146], adding multiple multi-indices at each greedy iteration could also yield better performance compared to adding only one multi-index at a time. The so-called "bulk-chasing" procedure identifies a subset* $\lambda_t^* \subset \Lambda_t^{RM}$ *of multi-indices that capture a fraction of the* L_2*-norm of the gradient along the reduced margin.*

Lastly, we use ν-fold cross-validation as in [226, 22] in order to determine the maximal cardinality of Λ_t. For each fold, we run Algorithm 1 with $\nu - 1$ folds of the training data for $m = n$ iterations and evaluate the objective function in (3.5) on the held-out test set. We then select the optimal number of terms $m^* \leq n$ in the expansion (3.32) that minimizes the average test errors over the ν folds. We call the complete procedure for learning each map component S_k with cross-validation, the Adaptive Transport Map (`ATM`) algorithm. The complete procedure is presented in Algorithm 2. In practice, we also stop the training on each fold early if the log-likelihood on the held-out test set does not continue to increase for more than 20 iterations.

The cross-validation procedure produces an approximation for f with an expansion whose cardinality is adapted to the number of training samples n. With more samples, we observe that the cross-validation procedure reliably adds more parameters, when needed, to reduce the bias in the approximation while controlling the variance of the estimated map parameters. Thus, we consider `ATM` a *semi-parametric* approach for approximating the map. The adaptation of the map complexity to the sample size n

will be demonstrated in the numerical results in the following section.

Algorithm 2: Adaptive Transport Map (`ATM`) algorithm for component S_k

Input: Training samples $\boldsymbol{\chi} = \{\boldsymbol{X}_{\leq k}^i\}_{i=1}^n$, Number of folds ν
Partition data into ν folds of equal size
for $j = 1, \ldots, \nu$ **do**
  Partition $\boldsymbol{\chi}$: $\boldsymbol{\chi}_j^{\text{test}}$ is the jth subset of $\boldsymbol{\chi}$ and $\boldsymbol{\chi}_j^{\text{train}} = \boldsymbol{\chi} \setminus \boldsymbol{\chi}_j^{\text{test}}$
  Estimate f_t using Algorithm 1 for $m = |\boldsymbol{\chi}_j^{\text{train}}|$ iterations with $\boldsymbol{\chi}_j^{\text{train}}$ samples
  Store iterates $\widehat{S}_k^{j,1} := \mathcal{R}_k(f_1), \ldots, \widehat{S}_k^{j,m} := \mathcal{R}_k(f_m)$
  Evaluate log-likelihood $\log(\widehat{S}_k^{j,t})^\sharp \eta_k$ for iterations $t = 1, \ldots, m$ with $\boldsymbol{\chi}_j^{\text{test}}$ samples
end for
Define m^* as the minimizer of the negative log-likelihood $t \mapsto \sum_{j=1}^{\nu} -\log(\widehat{S}_k^{j,t})^\sharp \eta_k(\boldsymbol{\chi}_j^{\text{test}})$
Estimate map f^* using Algorithm 1 for m^* iterations with $\boldsymbol{\chi}$ samples
Output: $\widehat{S}_k = \mathcal{R}_k(f^*)$

3.5 Experimental results

In this section, we evaluate the performance of the `ATM` algorithm on several joint and conditional density estimation problems. Sections 3.5.1–3.5.2 visualize the approximation power of the proposed method on various one and two-dimensional problems; Sections 3.5.3–3.5.4 illustrate the benefits of adaptivity and demonstrate how the estimated maps reveal and exploit structure in the target density; Section 3.5.5 presents density estimation results on a suite of UCI datasets. The code for reproducing the following numerical results is available online[3].

In all of the numerical experiments, we pre-process the data by subtracting the empirical mean and dividing each variable by its empirical standard deviation. While several functions (such as (3.10) and (3.11)) meet the conditions required on g in Propositions 2, 3, and 4 to guarantee the objective is smooth and there are no spurious local minima, we employ the modified soft-plus function $g(\xi) = \log(1 + 2^\xi)/\log(2)$ in our numerical experiments. This function satisfies $g(0) = 1$ so that $f(\boldsymbol{x}_{\leq k}) = 0$ is rectified into $\mathcal{R}_k(f)(\boldsymbol{x}_{\leq k}) = x_k$. We evaluate the integral in (3.9) numerically using

[3] https://github.com/baptistar/ATM

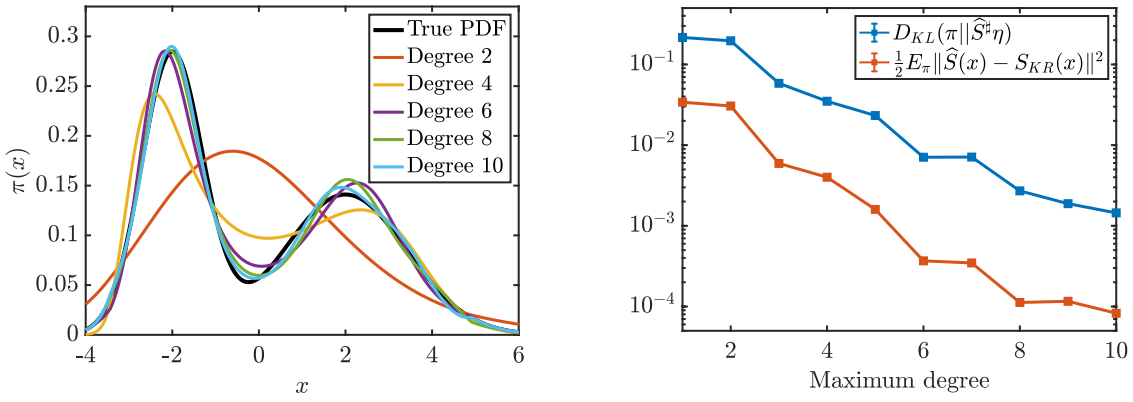
an adaptive quadrature method based on Clenshaw-Curtis points with a relative error of 10^{-3}. At each iteration of the `ATM` algorithm, we optimize the parameters $c_{\boldsymbol{\alpha}}$ for $\boldsymbol{\alpha} \in \Lambda_t$ using a BFGS quasi-Newton method [155].

3.5.1 One-Dimensional

We first consider a one-dimensional distribution defined as a mixture of Gaussians with density $\pi(x) = 0.5\mathcal{N}(x; -2, 0.5) + 0.5\mathcal{N}(x; 2, 2)$. To estimate π, we generate $n = 10^4$ training samples $\{X^i\}_{i=1}^n$ and use them to estimate the map S_{KR} that pushes forward π to η. In one-dimension, the KR rearrangement (which, in one-dimension, is also the optimal transport map [189]) is given by $S_{\text{KR}} = F_\eta^{-1} \circ F_\pi$, where F_η and F_π denote the cumulative density functions of the reference and target distributions, respectively. We approximate S as the transformation of f through $\mathcal{R}_1$ by looking for f in a finite-dimensional subspace V_k^p of V_k that is spanned by the polynomials in (3.33) of degree at most p. Figure 3-8a plots the approximate pullback densities $\widehat{\pi} = \widehat{S}^\sharp \eta$ when increasing the maximum polynomial degree p, or equivalently the number of terms $\#\Lambda_t$ in f (3.32). Figure 3-8b shows the convergence of the approximation as measured in the KL divergence from enlarging the finite-dimensional subspace of maps that we consider to approximate S_{KR}. We also observe convergence in the map towards the KR rearrangement in the L^2_π norm, as expected from the upper bound in (3.2).

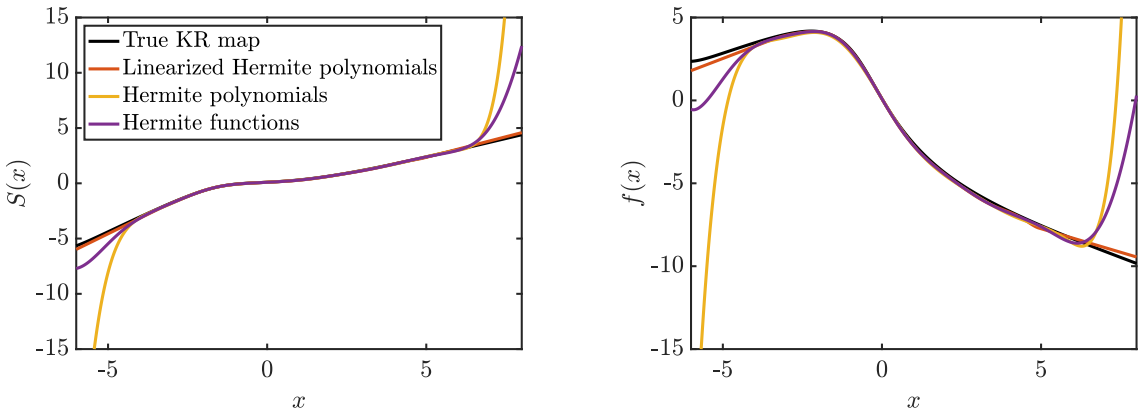
Figure 3-9 compares the approximation to the map S and the non-monotone function f with different basis ψ_i. Here we observe that approximating f using linearized Hermite polynomials closely tracks the KR rearrangement S_{KR} and the function f_{KR} in the tail regions, as compared to using unbounded Hermite polynomials or Hermite functions. Enforcing that the map has linear tail behavior ensures that the approximation has Gaussian tails in regions with few samples.

Next we consider a Gaussian mixture with density $\pi(x) = 0.5\mathcal{N}(x; 0, 1)+0.5\mathcal{N}(x; 0, \sqrt{0.05})$. This mixture is a common test case for sampling methods to exhibit their adaptability to densities with multiple scales [197]. Given $n = 10^4$ samples, we compute the approximate map using either the linearized Hermite polynomials or the Mexican hat



(a) (b)

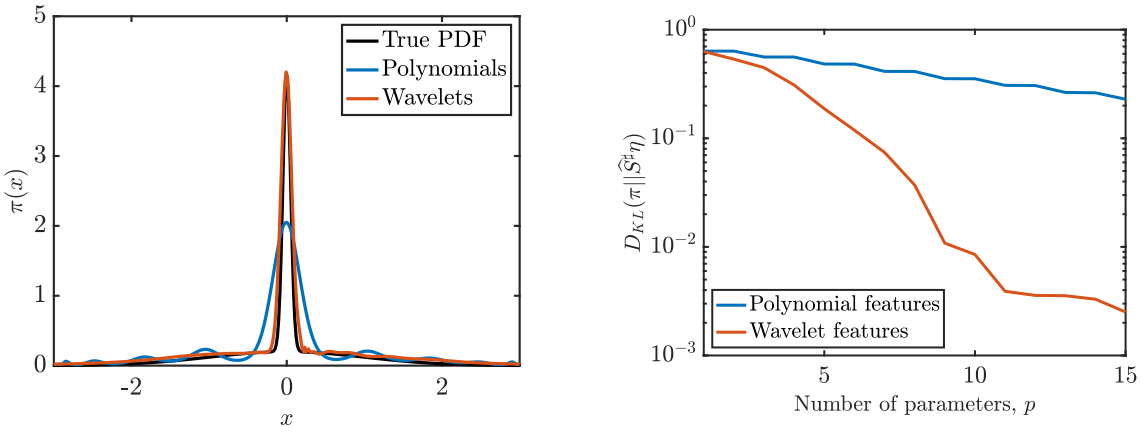
Figure 3-8: (a) The pullback density $\widehat{S}^\sharp\eta$ approaches the Gaussian mixture density π when increasing the maximum polynomial degree p for the function space $V_k^p \subset V_k$. (b) The pullback density converges to π in the KL divergence and estimated map converges to S_{KR} in L^2_π with increasing p.



(a) (b)

Figure 3-9: (a) The approximated transport maps S and (b) the non-monotonic functions f with different choices of basis ψ_i for the Gaussian mixture density. The linearized Hermite polynomials display the closest approximation to S_{KR} and f_{KR}.

wavelets. Figure 3-10 plots the approximation to the target density using a $m = 15$ term expansion in (3.32) for f, and the convergence in KL divergence for both polynomial and wavelet basis. We observe that the polynomial-based approximation suffers from oscillations in the approximate density, while the wavelets better capture localized features. This results in a much faster convergence in the KL divergence with wavelets compared with polynomials.



(a) (b)

Figure 3-10: Approximation of Gaussian mixture with overlapping components using Hermite polynomials or Ricker wavelet expansions. (a) A wavelet expansion with 15 basis functions better approximates the target density than a polynomial expansion of maximum degree 15, and (b) result in lower KL divergence for this example.

3.5.2 2D Datasets

To demonstrate the expressiveness of the (truncated) Hermite polynomial basis, we apply the `ATM` algorithm to approximate the Knothe–Rosenblatt rearrangement on several two dimensional distributions with various geometries. These include the Banana, Funnel, Cosine, Mixture of Gaussians (MoG), and Ring densities π that were considered in [228, 99]. For each density π, we generate $n = 10^4$ i.i.d. samples and apply a random rotation to the data using a uniformly distributed angle in $[0, 90]^\circ$. Figure 3-11 plots the true densities π and the approximate densities $\widehat{\pi} \coloneqq \widehat{S}^\sharp\eta$ found using `ATM`. We use 5-fold cross-validation to identify the optimal number of elements in the multi-index set for each map component: S_1 and S_2. Figure 3-11 includes the total number of parameters $\#\Lambda_t$ in both map components. We observe that the non-Gaussian densities are accurately captured with a small set of basis functions in each map component.

3.5.3 Mixture of Gaussians

In this example, we evaluate the stability and quality of the `ATM` algorithm for learning transport maps in small sample regimes. We consider a 3-dimensional distribution

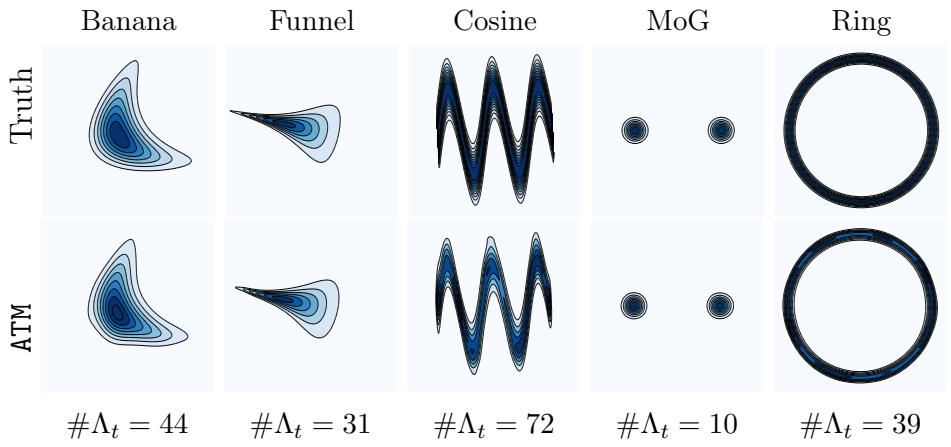
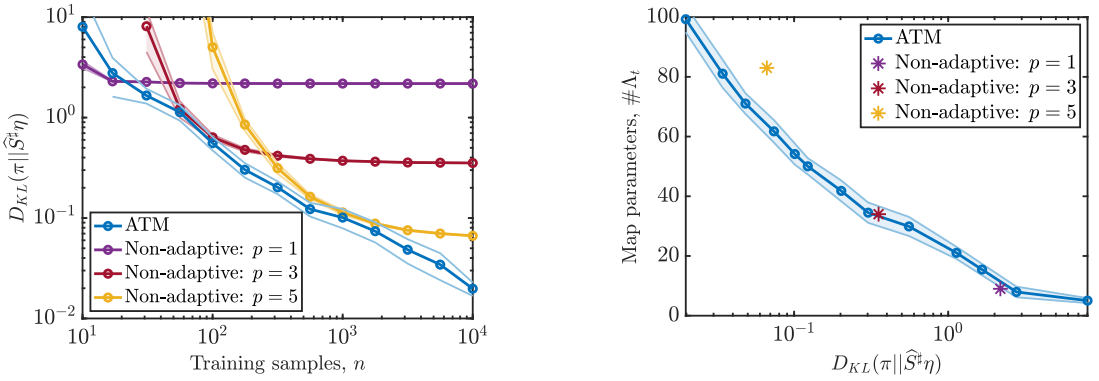


Figure 3-11: Five different 2D densities π (top row) and their approximations $\widehat{\pi}$ (bottom row) built using the `ATM` algorithm with the Hermite polynomials and with $n = 10^4$ samples from π. The total number of map parameters $\#\Lambda_t$ in both map components is included below each density.

π defined as a mixture of Gaussians centered at the 8 vertices of the hypercube $[-4, 4]^3$, each with covariance matrix I_d. The weights of the mixture components were randomly sampled from a uniform distribution $\mathcal{U}([0, 1])$ and then normalized so that $\int_{\mathbb{R}^3} \mathrm{d}\pi = 1$. To estimate π, we generate n training samples $\{\mathbf{X}^i\}_{i=1}^n \sim \pi$ and use the `ATM` algorithm to estimate the KR rearrangement that pushes forward π to η using 5-fold cross validation.

Figure 3-12a plots the KL divergence from $\widehat{\pi}$ to π averaged over 10 experiments. Here, the training sets used to build $\widehat{\pi} = \widehat{S}^\sharp \eta$ are of varying size $10^1 \leq n \leq 10^4$ and the reported KL divergence is computed on a common test set of 10^4 samples. Figure 3-12b plots the total number of parameters $\#\Lambda_t$ identified by the `ATM` algorithm using $\nu = 5$-fold cross validation. The performance of `ATM` is compared to a non-adaptive method where $\Lambda = \Lambda(p) = \{\boldsymbol{\alpha} \in \mathbb{N}_0^k, \|\boldsymbol{\alpha}\|_1 \leq p\}$ is arbitrarily fixed with $p = 1, 3,$ and 5. Note that $\Lambda(p)$ corresponds to a polynomial f with total-degree p. For each sample size n, `ATM` consistently finds a better estimator of π (with respect to the KL divergence) than a non-adaptive method with a fixed degree p by identifying the necessary basis $\{\psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \Lambda_t}$ to represent the map components. In addition, the `ATM` estimator achieves an improved KL divergence with a lower number of total map parameters, as seen in Figure 3-12b.



(a) (b)

Figure 3-12: (a) The KL divergence over 10 sets of training samples is lower for `ATM` than non-adaptive expansions. (b) The trade-off between the approximation quality and number of coefficients using the adaptive algorithm for different n. For each number of map parameters, the `ATM` algorithm finds a representation with similar or lower KL divergence than the lowest achievable error of each non-adaptive approximation (indicated with stars in plot (b)).

3.5.4 Stochastic volatility

Next we consider a dataset from a Markov process that describes the log-volatility for the return of a financial asset over time. The model has two hyper-parameters, μ and ϕ, and T states ($\mathbf{Z}_k$) for the log-volatility at times $k = 1, \ldots, T$. The two hyper-parameters are drawn from the distributions $\mu \sim \mathcal{N}(0, 1)$ and $\phi = 2\exp(\phi^*)/(1 + \exp(\phi^*))$ for $\phi^* \sim \mathcal{N}(3, 1)$. The states for the log-volatility follow the order one autoregressive process $Z_{k+1}|Z_k, \mu, \phi = \mu + \phi(Z_k - \mu) + \epsilon_k$ for all $k > 1$ where $\epsilon_k \sim \mathcal{N}(0, 1)$ is independent of all other variables and $Z_0|\mu, \phi \sim \mathcal{N}(\mu, \frac{1}{1-\phi^2})$. While the states are Gaussian conditioned on the hyper-parameters, the joint distribution of

$$\mathbf{X} = (\mu, \phi, \mathbf{Z}),$$

is non-Gaussian. In this example, the dimension of $\mathbf{X}$ is made arbitrarily large by increasing T.

A key property of triangular transport maps is that they inherit sparsity from the conditional independence structure of π. [204] showed that the Markov structure associated with π yields a lower bound on the sparsity of the map S (i.e., the functional

dependence of each component S_k on the input variables $\boldsymbol{x}_{<k}$). This sparsity was exploited to learn the structure of undirected probabilistic graph models from data in [150]. From the conditional independence properties of the stochastic volatility model, we know the KR rearrangement S_{KR} for the joint distribution of parameters and states is sparse. Furthermore, the sparsity of S_{KR} can be derived from Theorem 5.1 in [204]. We now show that the `ATM` algorithm is capable of detecting and exploiting such a structure without knowing it exists in advance.

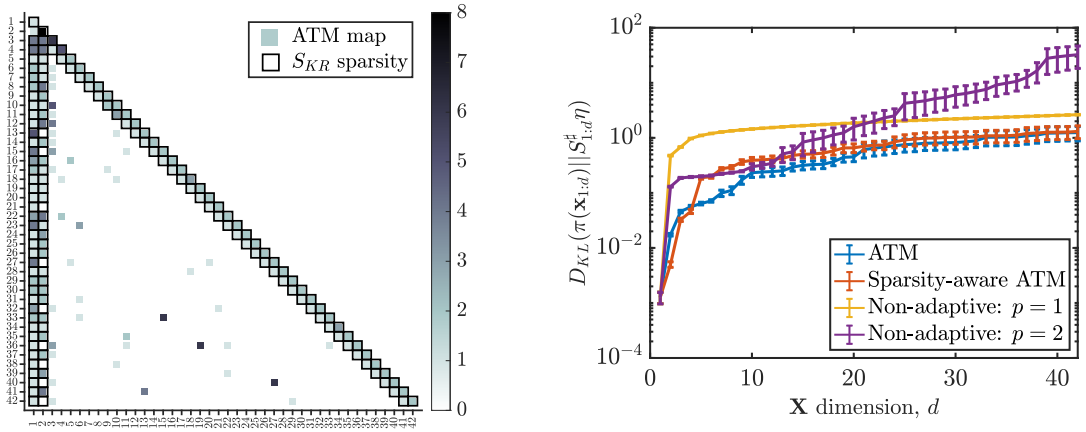
Figure 3-13a compares the variable dependence of the KR rearrangement S_{KR} and the map $\widehat{S}$ learned by the `ATM` algorithm for a distribution with $T = 40$ states using $n = 10^3$ samples from π; a non-filled (j, k) entry in the plot indicates that k-th map component does not depend on variable x_j. The dependence of component S_k on (x_k, x_{k-1}) shows that each state Z_k strongly depends on the previous state in time. Most of the map components also show dependence on the hyper-parameters (μ, ϕ). The estimated sparsity closely matches the exact sparsity of the KR rearrangement that is derived from the Markov process. Furthermore, the sparse variable dependence of the kth map component $\widehat{S}_k$ on parent nodes $\text{Pa}(k) \subseteq \{1, \ldots, k-1\}$ produces an approximation to the kth marginal conditional density given by

$$\widehat{\pi}_k(x_k|\boldsymbol{x}_{<k}) = \widehat{\pi}_k(x_k|\boldsymbol{x}_{\text{Pa}(k)}).$$

By identifying the parent nodes $\text{Pa}(k)$ of each variable k, we also learn a sparse Bayesian network or directed acyclic graphical (DAG) model representing the target distribution [112]. As a result, we can see the `ATM` algorithm as a technique for learning DAGs from samples, given a prescribed variable ordering.

Next, we consider the approximation of the S_{KR} map for Markov models of increasing state dimension T, and hence map dimension d. Figure 3-13b plots the KL divergence of approximations found with `ATM` and non-adaptive methods for marginal distributions of $\mathbf{X} \in \mathbb{R}^d$ as a function of the dimension using $n = 2000$ samples. We compare `ATM` with a variant of `ATM` where the exact sparsity pattern of the the KR rearrangement is known in advance. This algorithm, denoted "sparsity-aware `ATM`" in

Figure 3-13b, differs from the ATM algorithm in that it only activates multi-indices $\boldsymbol{\alpha}_t^*$ which match the sparsity of S_{KR} (meaning of the form $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, 0, \ldots, 0, \alpha_{k-1}, \alpha_k)$ for $k \geq 3$). Furthermore, we compare ATM with non-adaptive maps of degree $p = 1$ and $p = 2$ that do not exploit the conditional independence structure of π and depend on all input variables. While degree $p = 2$ maps can better capture the non-Gaussian conditional distributions for small-dimensional marginals as compared to $p = 1$ linear maps, the larger number of coefficients in these maps results in higher-variance estimators and a larger KL divergence. Instead, ATM achieves a KL divergence that grows slowly with d and performs similarly to the sparsity-aware ATM.



(a) (b)

Figure 3-13: (a) The sparsity of an ATM map $\widehat{S}$ and the KR rearrangement S_{KR} for the stochastic volatility model with $T = 40$ time steps. (b) The KL divergence using ATM for marginal distributions with increasing dimension d in comparison to adaptive map estimators with known sparsity and non-adaptive maps estimators.

3.5.5 Tabular Datasets

We now evaluate ATM's performance for density estimation on a suite of UCI datasets with a decreasing ratio of sample size n to dimension d. These datasets have dimensionalities between $d = 10$ and 15 and sample-sizes between $n = 506$ and 5875. We pre-process each dataset to eliminate discrete-valued variables and one variable in every pair that has Pearson correlation coefficient greater than 0.98, following the procedure in [219]. We consider 10 splits of the data. For each split, we use one fold (i.e.,

10% of the data) as a *test set* and the remaining 9 folds (i.e., 90% of the data) as the *training set* to build $\widehat{S}$. To assess the quality of our estimated map $\widehat{S}$, we evaluate the negative log-likelihood of the pullback density $\widehat{\pi} = \widehat{S}^\sharp \eta$ on the test set. The negative log-likelihood is an empirical estimator for $-\mathbb{E}_\pi[\log \widehat{S}^\sharp \eta] = \mathcal{D}_{\mathrm{KL}}(\pi || \widehat{S}^\sharp \eta) - \int \log \pi \mathrm{d}\pi$ and is, up to the unknown constant $-\int \log \pi \mathrm{d}\pi$, the Kullback-Leibler divergence from $\widehat{S}^\sharp \eta$ to π. Table 3.1 presents the mean of the negative log-likelihoods over the 10 splits and a 95% confidence interval for the mean. For all datasets, we observe an improvement using `ATM` to non-adaptive maps of maximum degree $p = 2$ with a lower number of total parameters.

Table 3.1: Mean negative log-likelihood for UCI datasets over 10 sets of training samples. The estimator with best performance (lowest negative log-likelihood) is highlighted in bold.

Dataset	(d, n)	Gaussian	`ATM`	Coefficients	$p = 2$ maps	Coefficients
Parkinsons	$(15, 5875)$	10.8 ± 0.4	$\mathbf{2.8 \pm 0.4}$	783 ± 17	5.1 ± 0.4	815
White Wine	$(11, 4898)$	13.2 ± 0.5	$\mathbf{11.0 \pm 0.2}$	342 ± 26	12.0 ± 1.0	363
Red Wine	$(11, 1599)$	13.2 ± 0.3	$\mathbf{9.8 \pm 0.4}$	289 ± 9	10.5 ± 0.2	363
Boston	$(10, 506)$	11.3 ± 0.5	$\mathbf{3.1 \pm 0.6}$	228 ± 7	6.5 ± 0.4	285

Lastly, we evaluate `ATM`'s performance for conditional density estimation on a suite of UCI datasets. We follow a similar procedure as above to pre-process each dataset. Each dataset has one-dimensional predictor variables X_d and covariates $\mathbf{X}_{<d}$ of varying dimension. To approximate the conditional density $\pi_d(x_d|\boldsymbol{x}_{<d})$, we estimate one map component S_d as a function of the predictor and the covariates using joint samples $\{(\mathbf{X}^i_{<d}, X^i_d)\}_{i=1}^n$. Table 3.2 presents the mean negative conditional log-likelihoods over 10 splits of the training data. The negative conditional log-likelihood is an empirical estimator for $-\mathbb{E}_\pi[\log \widehat{S}^\sharp_d \eta] = \mathbb{E}_{\pi(\boldsymbol{x}_{<d})}[D_{\mathrm{KL}}(\pi_d(x_d|\boldsymbol{x}_{<d})||\widehat{S}^\sharp_d \eta)] - \int \log \pi_d(x_d|\boldsymbol{x}_{<d}) \mathrm{d}\pi_d$ and is, up to the unknown constant $-\int \log \pi_d(x_d|\boldsymbol{x}_{<d}) \mathrm{d}\pi_d$, the expected Kullback-Leibler divergence from $\widehat{S}^\sharp_d \eta$ to the marginal conditional $\pi_d(x_d|\boldsymbol{x}_{<d})$.

We compare `ATM` to conditional kernel density estimation (`CKDE`) [195], ϵ-neighborhood kernel density estimation (`NKDE`) and kernel mixture networks (`KMN`) [2] using the implementation provided by [186]. We also include results for high-capacity parametric CDE methods that rely on neural-networks: mixture density networks (`MDN`) [23], and

conditional normalizing flows (NF) [217]; implementation details of these methods are provided in Appendix B.2. For all datasets, we observe comparable performance of ATM to neural-network based methods and improved performance on the Yacht dataset with our semiparametric method in comparison to all other approaches.

Table 3.2: Mean negative conditional log-likelihood for UCI datasets over 10 sets of training samples. The method with best performance from both categories is highlighted in bold.

Dataset	(d, n)	ATM	CKDE	NKDE	MDN	KMN	NF
Concrete	(9,1030)	$\mathbf{3.1 \pm 0.1}$	3.2 ± 0.1	3.9 ± 0.1	$\mathbf{2.9 \pm 0.1}$	3.5 ± 0.1	3.2 ± 0.2
Energy	(10,768)	1.5 ± 0.1	$\mathbf{1.0 \pm 0.1}$	2.1 ± 0.2	$\mathbf{1.2 \pm 0.1}$	1.7 ± 0.1	1.7 ± 0.3
Yacht	(7,308)	$\mathbf{0.5 \pm 0.2}$	1.1 ± 0.3	3.8 ± 0.2	$\mathbf{0.7 \pm 0.2}$	1.8 ± 0.2	1.3 ± 0.5
Boston	(12,506)	$\mathbf{2.6 \pm 0.2}$	$\mathbf{2.6 \pm 0.2}$	3.1 ± 0.2	$\mathbf{2.4 \pm 0.2}$	2.7 ± 0.2	$\mathbf{2.4 \pm 0.1}$

3.6 Discussion and extensions

This paper presented a functional analysis framework for approximating monotone triangular transport maps. The framework is based on representing monotone functions as the transformation of smooth unconstrained functions via a bijective operator. Particular choices for this operator coupled with an appropriate function space for its input results in an unconstrained optimization problem for learning the map components with nice properties. Theoretically, we show that these problems have smooth objectives, permitting us to use deterministic first or second-order methods to find the minimizers. Furthermore, they have no spurious local minima, thereby making the results robust to the choice of initial condition and other parameters of the optimization algorithms.

The functional analysis framework also enables the development of novel transport map estimators given only samples from a target density. In this work, we propose an adaptive algorithm for selecting a sparse set of basis functions to represent the map components. This procedure is semiparametric and automatically adapts the number of parameters to the sample size, thereby providing maps that capture complex structure with sufficient data, but that are also robust in settings with limited

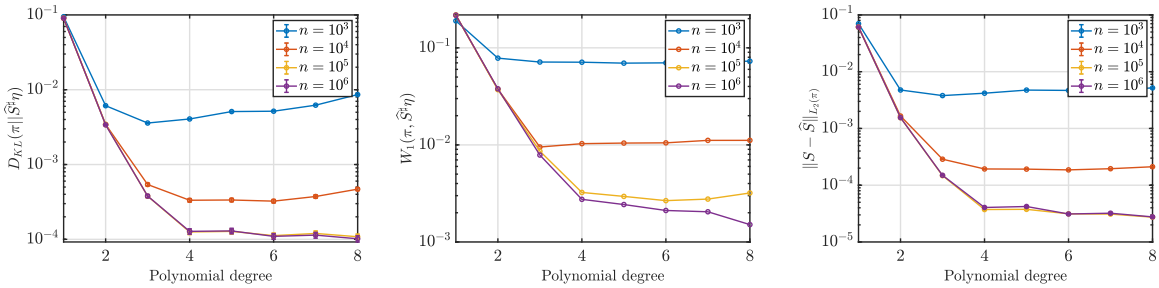
observations. In this work, we demonstrated the performance of this algorithm for various joint and conditional density estimation tasks.

Further understanding of the statistical and computational properties of transport map estimators is crucial to deploy these algorithms within large-scale applications, such as sequence inference (i.e., data assimilation) methods that require computing triangular transport maps; see [203]. Towards this goal, we outline some important future research directions below.

Approximation theory and statistical guarantees. An open research topic, to the best of our knowledge, is analyzing the finite-dimensional approximation of triangular transports on unbounded domains (see [240] for the case of analytic densities π, η on bounded domains). These results would show how approximate maps converge to the KR rearrangement with different parameterizations (e.g., polynomials with higher-order terms or neural networks of increasing widths). A parallel line of work is to develop non-asymptotic statistical convergence results for triangular transports (e.g., in the context of density estimation) with increasing sample size n. Combining these two results can provide lower bounds on the minimum number of samples required to learn maps of a given complexity as well as indicate how to optimally select the parameters to minimize the bias and variance of any map estimator given finite samples. The improvement in characterizing π with more parameters can also be traded-off with the greater computational resources that are required to solve higher-dimensional optimization problems for the map.

We studied the map approximation numerically using the monotone map representation introduced in Section 3.3. To reduce the randomness from the finite samples used in the empirical objective function, we performed this study with an increasing number of training samples $n \in \{10^3, 10^4, 10^5, 10^6\}$. For each sample size, we optimized the coefficients in f with the linearized Hermite polynomial expansion (3.33) for maximum polynomial degrees from 1 to 8. Figure 3-14 presents the KL divergence, Wasserstein-1 distance, and the L^2_π error in the map when approximating a one-dimensional Gumbel target density π. These errors were estimated on an inde-

pendent set of 10^7 i.i.d. samples from π. With a sufficiently large number of training samples, we observe an exponential decay in the error with respect to the polynomial degree. Future work will rigorously quantify and relate the convergence rates in various error metrics.



(a) (b) (c)

Figure 3-14: Convergence in (a) KL divergence (b) Wasserstein-1 distance, and (c) weighted L^2_π error in the map for increasing sample size and maximum polynomial degree

Map ordering. One disadvantage of triangular maps is that the approximation depends on the choice of variable ordering. In particular, each variable ordering yields a different factorization of the target density and Knothe-Rosenblatt rearrangement S_{KR}. Thus, it is of interest to develop variable ordering algorithms that minimizes the approximation complexity finite-sample error of the estimated transport map $\widehat{S}$, the estimated pullback density $\widehat{\pi}$, or other goal-oriented metrics. For target distributions that induce sparse variable dependence in the map, one approach is to find the ordering that maximizes the sparsity of $\widehat{S}$. This is equivalent to finding the sparsest Bayesian network or directed acyclic graphical (DAG) model for a distribution from samples. While this problem is in general NP-complete, [175] proposed an algorithm for finding the sparsest DAG. This algorithm reduces to finding a sparse maximum likelihood-estimator for the Cholesky factor of the inverse covariance matrix in the Gaussian setting. Given that a linear transport map corresponds to finding a Cholesky factor for Gaussian distributions (see [15, Section 3]), we expect the approach can be applied in non-Gaussian settings by seeking a nonlinear map using the `ATM` algorithm.

Nonparametric methods. Instead of finding a particular finite-dimensional basis for the map components $S_k = \mathcal{R}_k(f)$, nonparametric methods do not limit the functional form of f (e.g., as linear combinations of multivariate polynomials). For instance, one popular nonparametric method looks for f in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with reproducing kernel $K\colon \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$. Common examples of symmetric and positive definite kernels include the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|_2^2/\sigma^2)$ for $\sigma > 0$ and the polynomial kernel $K(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + 1)^p$ for some maximum degree $p \in \mathbb{N}$. The choice of kernel induces different smoothness properties on the functions in the corresponding RKHS. We refer the reader to [164] for an introduction to the theory of RKHSs.

One primary advantage of kernel methods is they result in infinite-dimensional feature expansions for f (with certain kernels) by solving finite-dimensional optimization problems. From Mercer's theorem, we can extract the features represented by a particular choice of kernel from the eigenfunctions of the operator $\mathcal{T}f(\boldsymbol{x}) \coloneqq \int_{\mathbb{R}^k} K(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}$. As an example, an orthonormal basis for the RKHS with the Gaussian kernel is the set of normalized Hermite functions.

Let ℓ be a loss function depending on n sample evaluations $(\boldsymbol{X}^i, f(\boldsymbol{X}^i)) \in \mathbb{R}^k \times \mathbb{R}$, such as $\ell((\boldsymbol{X}^1, f(\boldsymbol{X}^1)), \ldots, (\boldsymbol{X}^n, f(\boldsymbol{X}^n))) = \sum_{i=1}^n \frac{1}{2} \mathcal{R}_k(f(\boldsymbol{X}^i))^2 - \log |\partial_k \mathcal{R}_k(f(\boldsymbol{X}^i))|$. Let f^* be the solution to the minimization problem

$$\arg\min_{f \in \mathcal{H}} \ell(f) + \omega(\|f\|_{\mathcal{H}}), \tag{3.39}$$

where $\omega\colon [0, +\infty) \to \mathbb{R}$ is any strictly increasing function that encourages selecting functions with small RKHS norm. Then, the following theorem due to Schölkopf et al., describes the form of the solution. We refer the reader to [193] for a proof of this result.

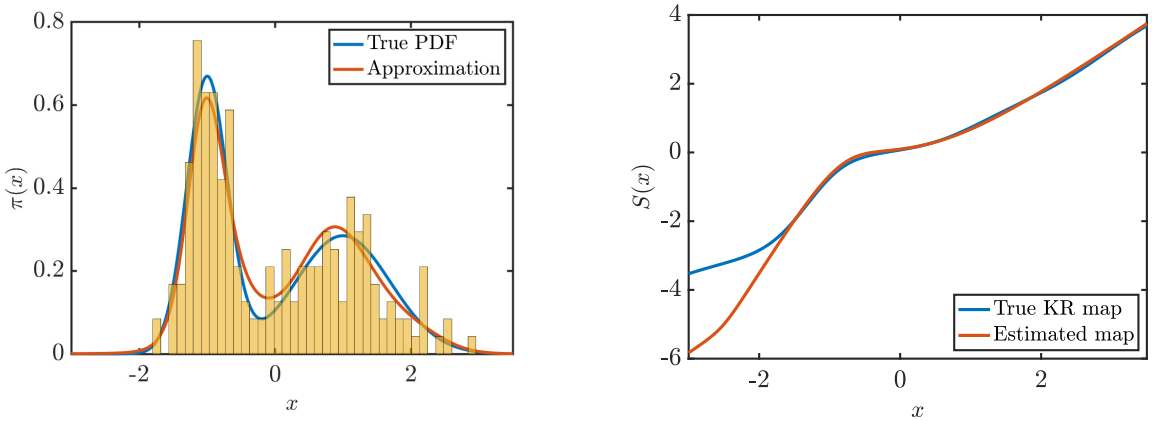
Theorem 3.6.1. *The minimizer f^* of the regularized empirical risk function in 3.39 admits a representation of the form $f^*(\boldsymbol{x}) = \sum_{i=1}^n \alpha_i K(\boldsymbol{x}, \boldsymbol{X}^i)$ with coefficients $\alpha_i \in \mathbb{R}$ for all $1 \leq i \leq n$.*

By introducing the optimal representation for f in (3.39), we can approximate f

by solving a finite-dimensional optimization problem for the n kernel coefficients (α_i). This problem has the form

$$\min_{(\alpha_i)} \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2}\mathcal{R}_k\left(\sum_{i=1}^{n}\alpha_i K(\boldsymbol{X}^j,\boldsymbol{X}^i)\right)^2 - \log\left|g(\sum_{i=1}^{n}\alpha_i \partial_k K(\boldsymbol{X}^j,\boldsymbol{X}^i))\right| + \omega(\|f\|_{\mathcal{H}}). \tag{3.40}$$

To demonstrate the value of kernel approximations, we consider a one-dimensional mixture of Gaussians target density $\pi(x) = 0.5\mathcal{N}(x,-1,0.3) + 0.5\mathcal{N}(x,1,0.7)$ and approximate the map $S = \mathcal{R}_1(f)$ by looking for f in an RKHS with a Gaussian kernel where ω is the squared function, i.e., $\omega(\|f\|_{\mathcal{H}}) = \lambda\|f\|_{L^2}^2$ with regularization parameter $\lambda > 0$. The approximate density and map using $n = 200$ samples with cross-validated $\lambda = 0.006$ and $\sigma = 0.77$ are plotted in Figures 3-15a and 3-15b, respectively. We observe that the map and density are well approximated within the bulk of π, while the tails of the estimated map $\widehat{S}$ are not well captured with localized kernel features centered at the training samples. Instead, $\widehat{S}(x)$ reverts to the identity function in the tails outside of the support of the training samples.



(a) (b)

Figure 3-15: (a) The mixture of Gaussians target density, samples, and its approximation. (b) Map $S = \mathcal{R}_1(f)$ with a non-parametric approximation of f

While it is not necessary to greedily select the basis expansions here, it is crucial to properly set the kernel parameters (i.e., the bandwidth σ) and the regularization parameter λ for the experiments. For instance, a small σ will lead the estimator to interpolate the data, while a large σ and/or λ will tend f to a constant. One drawback

of kernel-based methods is they are known to suffer from the *curse of dimensionality*. This is manifested in the number of samples n having to grow exponentially with the dimension k to obtain a bounded estimation error (see [84] for these convergence rates in the context of kernel regression). We note that this rate improves, howevever, for smoother functions with an appropriate choice of λ and σ. It will be interesting to compare the finite-sample performance of these nonparametric estimators to the semiparametric procedure proposed in this chapter and a related approach that uses Gaussian process expansions for the map components [107].

Scalability to large-data. For datasets with large n, it is computationally intensive to store and continually process all of the samples from π to approximate f. For instance, the optimization of a map with $p = 10$ parameters given $n = 10^6$ samples and 64 quadrature points consumes about 10GB of memory, which is close to the limit of many laptop computers. In these large-data settings, machine learning algorithms only process small random subsets of the training data at once during the optimization process. For instance, stochastic gradient descent (SGD) uses small mini-batches of samples to estimate the objective function's gradient in each iteration of a gradient descent procedure. One can adopt SGD to identify the map coefficients $\boldsymbol{\alpha}$ given a prescribed set of multi-indices (e.g., in step 7 of Algorithm 1). We can also use SGD when selecting new multi-indices by ensuring the batch-size is large enough so that a nearly optimal multi-index is selected from within the reduced margin. This is formalized by ensuring the multi-index $\boldsymbol{\alpha}_t^*$ selected at iteration t using a subset of the data satisfies

$$|\nabla_{\boldsymbol{\alpha}_t^*} \widehat{\mathcal{L}}_k(f_t)| \geq \theta \max_{\boldsymbol{\alpha} \in \Lambda_t^{\mathrm{RM}}} |\nabla_{\boldsymbol{\alpha}} \widehat{\mathcal{L}}_k(f_t)| \tag{3.41}$$

for some relaxation parameter $\theta \in (0, 1]$. Let us remark that if $\theta = 1$, we recover Algorithm 1. If the selected multi-index satisfies (3.41) for all iterations t, then the greedy procedure is known as a week greedy algorithm. When minimizing convex functionals, [214] showed that condition (3.41) is sufficient to guarantee convergence of a week greedy algorithm to the global minimizer. Future work will demonstrate this numerically and explore other techniques to increase the algorithm's scalability

in large-data settings.

Choice of reference and objective function. In practice, the choice of reference density η affects both the approximation and statistical estimation of the map components S_k. If $\mathbf{X} \sim \pi$ is a heavy-tailed random variable (e.g., with a Laplace distribution), the pullback of a map that grows linearly as $\|x\| \to \infty$ through a Gaussian reference density will not be able to represent the target density π. Furthermore, using a Gaussian reference will yield a map optimization problem in (3.5) that is very sensitive to samples in the tails of π. Instead, choosing a reference $\mathbf{Z} \sim \eta$ that follows a Laplace distribution with independent marginal components (i.e., $\eta(\boldsymbol{z}) = \prod_{i=1}^{d} \eta_i(z_i)$ where $\eta_i(z_i) = \frac{1}{2}\exp(-|z_i|)$, will result in the objective function $\mathcal{J}_k(s) = \mathbb{E}_\pi[|s(\mathbf{X})| - \log|\partial_k s(\mathbf{X})|]$ for map component S_k. As compared to (3.4), this objective measures absolute values instead of squared values of S_k at each sample $\mathbf{X}$. For samples $\mathbf{X}$ in the tails of π that are typically mapped to the tails of η, this modification makes the objective function less dominated by single samples with large values for $S_k(\mathbf{X})$. This is analogous to using the absolute or the Huber loss instead of the squared loss in robust regression to reduce outlier sensitivity [73]. Another class of alternative loss function for the KL divergence are the tail-adaptive f-divergences proposed in [224]. Lastly, a related research direction is to develop transport map parameterizations that match the tail properties of an arbitrary target distribution.

Chapter 4

Learning non-Gaussian graphical models

4.1 Introduction

An undirected probabilistic graphical model represents the conditional independence, or Markov, properties satisfied by a collection of random variables $X = (X_1, \ldots, X_d)$ with law $\boldsymbol{\nu}_\pi$. In particular, it is a graph $\mathcal{G} = (V, E)$ where the set of vertices $V = \{1, \ldots, d\}$ contains the indices of the random variables and where the set of edges E encodes the statistical dependence structure of $\boldsymbol{X}$ in the following way: for any disjoint subsets A, B, and C of V, $\boldsymbol{X}_A$ and $\boldsymbol{X}_B$ are conditionally independent given $\boldsymbol{X}_C$ if C *separates* A and B in the graph $\mathcal{G}$. That is, if after removing nodes C from $\mathcal{G}$, there is no path between sets A and B in the resulting graph, then the conditional independence property

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \,|\, \boldsymbol{X}_C \tag{4.1}$$

holds. In this case, we say that $\mathcal{G}$ is an independence map (I-map), or Markov network, for $\boldsymbol{\nu}_\pi$. The goal of this chapter is to learn the Markov network of $\boldsymbol{\nu}_\pi$ given only a finite set of samples drawn from $\boldsymbol{\nu}_\pi$.

Learning the Markov properties of a distribution, given a set of data drawn from it, is useful for various reasons. Undirected graphs can reveal paths, hierarchical

and cyclic interactions between the variables [64, 194]. These graph structures have been used to interpret datasets from many application areas such as handwriting recognition [70, 68], natural language processing [21, 148], protein folding [215], and many more [25, 188]. The graphical model also defines efficient factorizations of high-dimensional distributions by representing the density as a product of potential functions that each depend on low-dimensional subsets of variables [223]. In some fields, such as image processing and spatial statistics, the Markov structure of the problem may be immediately recognizable [112], and exploiting the structure of the undirected graphical model greatly simplifies algorithms for inference and prediction.

Previous work in learning probabilistic graphical models has mainly focused on certain parametric families of distributions [59]. For Gaussian random variables, the conditional independence properties are encoded by the sparsity of the inverse covariance, or *precision*, matrix. That is, the (i, j) entry of the precision matrix is zero if and only if variables X_i and X_j are conditionally independent given the rest. Thus, learning the graph reduces to identifying the support of the non-zero entries in the precision matrix. An active area of research considers how to learn the graph with (very) sparse data relative to the dimension d of the variables. One of the best-known methods to tackle this problem is the graphical lasso (GLASSO), introduced by Banerjee, Ghaoui, and d'Aspremont [13] and Yuan and Lin [237], which solves an ℓ_1-penalized maximum likelihood estimation problem for the precision matrix. One popular method for finding the GLASSO estimator is the coordinate descent algorithm of Friedman, Hastie, and Tibshirani [72]. In the case of discrete random variables, approaches for finding sparse graphs include ℓ_1-penalized logistic regression [223], and the estimation of a generalized covariance matrix whose inverse (via its support) encodes conditional independence properties [132].

In the case of continuous and non-Gaussian data, the connection between the inverse covariance matrix and conditional independence is lost. Outside of the Gaussian setting, a regularized score matching method was proposed to learn sparse graphs for distributions within the exponential family [128]. Recently, a large class of multivariate graphical models was considered by combining node-wise conditional distributions

in the exponential family [233, 208]. Another area of research has proposed semi-parametric methods based on Gaussian copulas to model non-Gaussian data [131]. In this case, the observations are assumed to come from marginal nonlinear transformations of a multivariate Gaussian random vector with known Markov properties. The marginal transformations yield potentially non-Gaussian marginal distributions while at the same time preserving the I-map of the original multivariate Gaussian distribution. Per Sklar's theorem, any multivariate distribution can be written in terms of a copula and the univariate marginals [154]. Recently, these copulas were extended to fit models based on elliptical distributions [130]. However, the specific families of copulas that can be easily or explicitly described remain rather limited [11]. Moreover, data generated via Gaussian copula transformations may fail to fully test a structure learning algorithm's ability to handle non-Gaussianity.

So far, there has been relatively little work for general non-Gaussian distributions outside of the exponential family. The current work is concerned with providing a mathematical and algorithmic framework to describe and identify the Markov properties of a continuous and non-Gaussian distribution. As an expansion of our NeurIPS paper, [150], the main contributions of this chapter are as follows.

First, we establish a framework which allows for the description and computation of conditional independence properties in the non-Gaussian setting. This is represented by a new *conditional independence score* matrix $\Omega \in \mathbb{R}^{d \times d}$: each entry Ω_{ij} is a score for the independence of Z_i and Z_j conditioned on the remaining variables. The entry Ω_{ij} is defined by the expected magnitude of certain mixed derivatives—in other words, integrated *Hessian* information—from the joint density π. We show that, under certain assumptions, the score provides an upper bound for the *conditional mutual information*, a widely used measure of conditional dependence. To compute Ω given only samples from π, an estimate of the joint density is needed. This is achieved via a transport map—a transformation that deterministically couples one probability measure to another.

Second, we expand the use and analysis of an algorithm called SING (Sparsity Identification in Non-Gaussian distributions). A key element of the algorithm is a

thresholding scheme, and thus we propose a class of threshold estimators that are *consistent* for graph recovery. The success of the algorithm also depends on a strong and explicit connection between the sparsity of the graph and the sparsity of triangular transport maps associated with π. This connection is exploited to iteratively reveal sparsity in the graph. We show numerically that this iterative algorithm provides an improved estimator for the conditional independence score, compared to a non-iterative approach that does not account for the structure in the map.

Third, we explore the relationship between the distribution in question and how to learn its corresponding graph. Because there exist an infinite number of probability distributions with the same minimal I-map, intuitively it should be easier to identify the graph than to estimate the entire joint distribution. We refer to this notion as the *information gap*. Through several empirical and analytical studies, we explore the extent to which this notion is in fact true and discuss its consequences for graph identification. Moreover, we see how this notion can be directly exploited by the algorithm. That is, in some cases, we still recover the correct graph even with a biased approximation to the density resulting from a constrained parameterization for the transport map. On the other hand, it is also possible that constraining the form of the map yields an incorrect graph. We show examples of both cases in Section 4.5.

The remainder of this chapter is organized as follows. Section 4.2 introduces the conditional independence score and its connections to conditional mutual information. Section 4.3 describes a sample-based estimator for the score based on a transport map approximation to the density. Here we introduce our first algorithm for learning the Markov structure, and propose a class of threshold estimators that are consistent for this structure. To take advantage of the connection between the sparsity of the graph and the transport map, Section 4.4 presents the iterative algorithm SING. Section 4.5 demonstrates the performance of the algorithm on a variety of numerical examples and explores the notion of the information gap. Section 4.6 provides a discussion and outlook on future research directions. Appendix C contains the proofs of our main results on the conditional independence score, the consistency of our estimator, and

the conditional independence properties of various distributions.

4.2 Measures of conditional independence

An alternative to the global Markov properties (4.1) for characterizing conditional independence are the pairwise Markov properties. A distribution for $\boldsymbol{X} = (X_1, \ldots, X_d)$ satisfies a pairwise Markov property between two variables when the associated nodes are not connected in the graph $\mathcal{G} = (V, E)$. That is, given two variables X_i and X_j for $i \neq j$, the lack of an edge between nodes i and j means the two variables are conditionally independent given the remaining variables:

$$(i, j) \notin E \iff X_i \perp\!\!\!\perp X_j \,|\, \boldsymbol{X}_{-ij}.$$

Here $\boldsymbol{X}_{-ij}$ denotes the random vector obtained by removing the ith and jth components from $\boldsymbol{X}$. The pairwise Markov property is, in general, weaker than the global, but when the density is strictly positive, i.e., $\pi(\boldsymbol{x}) > 0 \; \forall \boldsymbol{x} \in \mathbb{R}^d$, the global and the pairwise Markov properties are equivalent [see 119]. In this work, we restrict our attention to this setting.

By definition, $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{-ij}$ means that the joint conditional density $\pi(x_i, x_j | \boldsymbol{x}_{-ij}) = \pi(x_i, x_j, \boldsymbol{x}_{-ij}) / \int \pi(x_i', x_j', \boldsymbol{x}_{-ij}) \mathrm{d}x_i' \mathrm{d}x_j'$ factorizes as the product of the marginal conditionals

$$\pi(x_i, x_j | \boldsymbol{x}_{-ij}) = \pi(x_i | \boldsymbol{x}_{-ij}) \pi(x_j | \boldsymbol{x}_{-ij}), \tag{4.2}$$

where $\pi(x_i | \boldsymbol{x}_{-ij}) = \int \pi(x_i, x_j | \boldsymbol{x}_{-ij}) \mathrm{d}x_j$ and $\pi(x_j | \boldsymbol{x}_{-ij}) = \int \pi(x_i, x_j | \boldsymbol{x}_{-ij}) \mathrm{d}x_i$. Thus, $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{-ij}$ allows the joint density to factorize as $\pi(\boldsymbol{x}) = \pi(x_i | \boldsymbol{x}_{-ij}) \pi(x_j | \boldsymbol{x}_{-ij}) \pi(\boldsymbol{x}_{-ij})$. If we further assume that π is continuously differentiable, this yields

$$\partial_i \partial_j \log \pi(\boldsymbol{x}) = 0, \tag{4.3}$$

for any $\boldsymbol{x} \in \mathbb{R}^d$. Conversely, if a strictly positive continuously differentiable $\pi(\boldsymbol{x})$ satisfies (4.3) for any $\boldsymbol{x} \in \mathbb{R}^d$, then $\pi(x_i, x_j | \boldsymbol{x}_{-ij})$ necessarily factors as in (4.2) so

that $X_i \perp\!\!\!\perp X_j | \boldsymbol{X}_{-ij}$. The characterization (4.3) of conditional independence has been already observed in Lemma 4.1 of Spantini, Bigoni, and Marzouk [204]. Based on (4.3), we propose to measure the conditional independence of X_i and X_j by the score $\Omega_{ij} \geq 0$ defined as

$$\Omega_{ij} := \int |\partial_i \partial_j \log \pi(\boldsymbol{x})|^2 \, \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{4.4}$$

A similar measure also appears in [150] under the name of "generalized precision," the difference being the square inside the integral.[1] For a strictly positive and continuously differentiable density π, the condition $\Omega_{ij} = 0$ yields (4.3) so that X_i and X_j are conditionally independent. That is, the sparsity pattern of Ω gives the Markov structure of π. The entries of Ω also provide a natural score for conditional independence: a value of Ω_{ij} near zero means that X_i and X_j are nearly conditionally independent, whereas a large value of Ω_{ij} means that X_i and X_j are strongly conditionally dependent. In Section 4.3.4, we will use this interpretation to estimate the Markov structure of π by *thresholding* the entries of an estimator for Ω.

We can relate the magnitude of score Ω_{ij} to another popular measure of conditional independence, the conditional mutual information (CMI). The CMI $I(X_i; X_j | \boldsymbol{X}_{-ij})$ is defined as the expected (with respect to $\boldsymbol{X}_{-ij}$) Kullback–Leibler divergence from the product of the marginal conditionals $\pi(x_i|\boldsymbol{x}_{-ij})$ and $\pi(x_j|\boldsymbol{x}_{-ij})$ to the joint conditional $\pi(x_i, x_j|\boldsymbol{x}_{-ij})$; that is,

$$I(X_i; X_j | \boldsymbol{X}_{-ij}) = \int \left(\int \pi(x_i, x_j|\boldsymbol{x}_{-ij}) \log \frac{\pi(x_i, x_j|\boldsymbol{x}_{-ij})}{\pi(x_i|\boldsymbol{x}_{-ij})\pi(x_j|\boldsymbol{x}_{-ij})} \mathrm{d}x_i \mathrm{d}x_j \right) \pi(\boldsymbol{x}_{-ij}) \mathrm{d}\boldsymbol{x}_{-ij} \tag{4.5}$$

$$= \int \log \left(\frac{\pi(x_i, x_j|\boldsymbol{x}_{-ij})}{\pi(x_i|\boldsymbol{x}_{-ij})\pi(x_j|\boldsymbol{x}_{-ij})} \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{4.6}$$

The CMI is widely adopted in part due to its information theoretic interpretations. The following theorem shows that $I(X_i; X_j | \boldsymbol{X}_{-ij})$ can in fact be *bounded above* by Ω_{ij}. This result relies on logarithmic Sobolev inequalities.

[1] In this work, we no longer use the term "generalized precision," but keep the same notation Ω to denote the new conditional independence score.

Definition 1. *A probability density function π on $\mathbb{R}^d$ satisfies the logarithmic Sobolev inequality if there exists a constant $C < \infty$ such that*

$$\int h \log \frac{h}{\int h \mathrm{d}\pi} \mathrm{d}\pi \leq \frac{C}{2} \int \|\nabla \log h\|_2^2 \, h \, \mathrm{d}\pi, \tag{4.7}$$

holds for any continuously differentiable function $h : \mathbb{R}^d \to \mathbb{R}_{>0}$. Here $\|\cdot\|_2$ denotes the canonical Euclidean norm on $\mathbb{R}^d$. The smallest constant $C = C(\pi)$ such that (4.7) holds is called the logarithmic Sobolev constant of π.

Theorem 4.2.1. *Let π be a strictly positive continuously differentiable probability density function on $\mathbb{R}^d$. Assume that all the conditional densities of the form $\pi(\cdot|\boldsymbol{x}_{-i})\colon x_i \mapsto \pi(x_i|\boldsymbol{x}_{-i})$ and $\pi(\cdot,\cdot|\boldsymbol{x}_{-ij})\colon (x_i, x_j) \mapsto \pi(x_i, x_j|\boldsymbol{x}_{-ij})$ satisfy the logarithmic Sobolev inequality with constants uniformly bounded by some constant $C_0 < \infty$, meaning*

$$C(\pi(\cdot|\boldsymbol{x}_{-i})) \leq C_0 \quad \textit{and} \quad C(\pi(\cdot,\cdot|\boldsymbol{x}_{-ij})) \leq C_0, \tag{4.8}$$

for all $\boldsymbol{x}_{-i} \in \mathbb{R}^{d-1}$ and $\boldsymbol{x}_{-ij} \in \mathbb{R}^{d-2}$ and for all $1 \leq i \neq j \leq d$. Then

$$I(X_i; X_j|\boldsymbol{X}_{-ij}) \leq C_0^2 \, \Omega_{ij}, \tag{4.9}$$

holds for any $i \neq j$.

The proof of Theorem 4.2.1 is provided in Appendix C.

Let us now comment on assumption (4.8) of Theorem 4.2.1. In general, there is no simple way to compute exactly the logarithmic Sobolev constant of a density. The Holley–Stroock perturbation criterion [87] and the Bakry–Émery criterion [12] are commonly used to bound the logarithmic Sobolev constant. Following Zahm, Cui, Law, Spantini, and Marzouk [239], one can combine these two criteria to show that if there exists a log-concave probability density π_0 and two scalars $\alpha > 0$ and $\beta < \infty$ such that $\alpha\pi_0(\boldsymbol{x}) \leq \pi(\boldsymbol{x}) \leq \beta\pi_0(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$, then π satisfies (4.8) with constant

$$C_0 = \frac{\beta}{\alpha\lambda}, \tag{4.10}$$

where $\lambda > 0$ is a lower bound for the smallest eigenvalue of $-\nabla^2 \log \pi_0(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$. We refer the reader to Zahm, Cui, Law, Spantini, and Marzouk [239, Section 2.2] for more details and examples of probability densities that satisfy the above conditions.

Remark (Gaussian case)**.** *Suppose that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \Sigma)$ is a Gaussian vector with mean $\boldsymbol{m} \in \mathbb{R}^d$ and non-singular covariance $\Sigma \in \mathbb{R}^{d\times d}$. Because $\pi(\boldsymbol{x}) \propto \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{m}))$ we have*

$$\Omega_{ij} = (\Sigma^{-1})^2_{ij}, \tag{4.11}$$

and so the sparsity pattern of the precision matrix Σ^{-1} gives the Markov structure of $\boldsymbol{X}$. This is a well known property of Gaussian vectors; see for instance Proposition 3.3.6. in [60]. This also explains the name "generalized precision" in [150].

Next we show how sharp the inequality (4.9) is for a $d = 2$ dimensional Gaussian vector $\boldsymbol{X} = (X_1, X_2)$ with covariance matrix $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ for some correlation $-1 \le \rho \le 1$. One can compute the (conditional) mutual information analytically $I(X_1; X_2|\boldsymbol{X}_{-12}) = I(X_1; X_2) = -\frac{1}{2}\log(1-\rho^2)$ and the score $\Omega_{12} = (\frac{\rho}{1-\rho^2})^2$. It remains to compute C_0. Using the formula (4.10) with $\beta = \alpha = 1$ (since π is log-concave we can chose $\pi_0 = \pi$), we obtain $C_0 = \lambda_{\min}(-\nabla^2 \log \pi(\boldsymbol{z}))^{-1} = \lambda_{\max}(\Sigma)$. Because $\lambda_{\max}(\Sigma) \le 2$, we then have

$$I(X_1; X_2) = -\frac{1}{2}\log(1-\rho^2) \overset{(4.9)}{\le} C_0^2 \Omega_{12} = \lambda_{\max}(\Sigma)^2 \left(\frac{\rho}{1-\rho^2}\right)^2 \le 4\left(\frac{\rho}{1-\rho^2}\right)^2.$$

We complete this section by comparing the *computational cost* of evaluating the CMI with that of evaluating the conditional independence score Ω_{ij}, for one variable pair (X_i, X_j). Given the joint density $\pi(\boldsymbol{x})$, computing the CMI requires the ability to evaluate the *normalized* conditional density $\pi(x_i, x_j|\boldsymbol{x}_{-ij})$ and the two normalized marginal conditionals, $\pi(x_i|\boldsymbol{x}_{-ij})$ and $\pi(x_j|\boldsymbol{x}_{-ij})$, for any value of $x_i, x_j \in \mathbb{R}$, $\boldsymbol{x}_{-ij} \in \mathbb{R}^{d-2}$. Breaking this down further, evaluating $\pi(x_i, x_j|\boldsymbol{x}_{-ij})$ requires integrating the joint density with respect to (x_i, x_j), while the marginal densities involve further integration with respect to the x_i and x_j variables individually. Then one must

take an outer expectation over $\boldsymbol{x}$. Evaluating the CMI therefore requires computing *nested* integrals, which is known to be challenging without additional structure [173]. In practice, information theoretic quantities that involve nested integration, such as CMI, are approximated using nested Monte Carlo (NMC) estimators. Given a budget of n samples, the best-case root-mean-squared error of NMC converges at a slow asymptotic rate of $\mathcal{O}(n^{-1/3})$. Moreover, as these Monte Carlo estimators involve the repeated computation of normalizing constants, controlling variance requires the construction of suitable biasing distributions for importance sampling [74], often in a way that depends on the conditioning variables [67].

On the other hand, computing the score entry Ω_{ij} requires evaluating a single mixed derivative of the log density $\log \pi(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$. In computing derivatives of the log density, we only require access to the unnormalized $\pi(\boldsymbol{x})$. Moreover, the outer expectation of the score can be approximated with standard non-nested Monte Carlo estimators, converging at the usual $\mathcal{O}(n^{-1/2})$ rate. Samples can be drawn directly from the target π, and no importance sampling of any kind is required.

In either case, evaluating the CMI or the conditional independence score requires a explicit functional form of the joint density π. In a setting where we only have access to samples from π, many structure learning algorithms rely on density estimation techniques as a step to learn the Markov structure. In general, density estimation in high dimensions, especially for non-Gaussian data, can become computationally expensive. Thus, a major question becomes: What is an efficient or advantageous way to represent the density given the goal of learning the Markov structure of a distribution? To answer this, we rely on a particular method of density estimation, using transport maps.

4.3 Estimators of conditional independence score

In this section, we propose a sample-based estimator for the sparsity of the conditional independence score Ω. Given samples from the target density π, we first estimate a transport map S (see Chapters 2 and 3) that pulls back a standard normal density

η to the target density π, i.e. $\pi = S^\sharp \eta$. In 4.3.1 we present the parameterization we use for the map in this chapter, and in 4.3.2 we show some asymptotic statistical properties of the estimated map parameters. In 4.3.3 and 4.3.4, we show to build a thresholding sample-based estimator for the score Ω and prove its asymptotic consistency for recovering the Markov structure of π in 4.3.6.

4.3.1 Representation for density via monotone transport maps

To approximate the KR rearrangement between a pair of densities on $\mathbb{R}^d$, we consider a triangular map with monotone increasing components. As shown in Chapter 3, monotonicity is enforced component-wise by ensuring the derivative of the kth component S_k with respect to the kth variable is a strictly positive function, i.e., $\partial_k S_k(x_1, \ldots, x_k) > 0$ for all $(x_1, \ldots, x_k) \in \mathbb{R}^k$. In this chapter, we consider the following general parameterization that guarantees monotonicity of each component S_k, which is given by:

$$S_k(x_1, \ldots, x_k) = c_k(x_1, \ldots, x_{k-1}) + \int_0^{x_k} g\left(h_k\left(x_1, \ldots, x_{k-1}, t\right)\right) dt, \tag{4.12}$$

for some positive function $g \colon \mathbb{R} \to \mathbb{R}_+$ and functions $c_k \colon \mathbb{R}^{k-1} \to \mathbb{R}$ and $h_k \colon \mathbb{R}^k \to \mathbb{R}$ [174].

Remark. *Letting $c_k(\boldsymbol{x}_{<k}) = f_k(\boldsymbol{x}_{<k}, 0)$ and $h_k(\boldsymbol{x}_{\leq k}) = \partial_k f_k(\boldsymbol{x}_{\leq k})$ for some smooth function $f_k \colon \mathbb{R}^k \to \mathbb{R}$ and choosing g to be a bijective function yields the representation for S_k proposed in Chapter 3.*

Two common choices for g in (4.12) are $g(x) = \exp(x)$ and $g(x) = x^2$, which result in the "integrated exponential" and the "integrated squared" parameterizations, respectively. In this chapter, we use the integrated squared parameterization because of its computational advantage of closed-form integration under certain choices for h_k [204, 28]. Following Baptista, Zahm, and Marzouk [16], we parameterize the

functions c_k and h_k using linear expansions

$$c_k(\boldsymbol{x}) = \sum_j c_{k,j}\psi_j(\boldsymbol{x}), \qquad h_k(\boldsymbol{x}) = \sum_j h_{k,j}\phi_j(\boldsymbol{x}), \tag{4.13}$$

in terms of tensor product Hermite functions (ψ_j, ϕ_j) and unknown coefficients $\boldsymbol{\alpha} = (c_{k,j}, h_{k,j})$. We note that recent methods for autoregressive density estimation alternatively represent the functions c_k and h_k using neural networks [162, 99].

In practice, we truncate the expansions in (4.13) by prescribing a maximum total degree β for the multivariate Hermite functions in c_k and h_k. We denote the space of lower-triangular maps with total degree β as $\mathcal{S}_\Delta^\beta$. As expected, a higher degree provides a richer basis for density estimation, but requires more computational effort to optimize and more samples to accurately estimate the coefficients. Here we follow the convention in [204] where a map of degree β uses basis functions up to degree β for c_k and $\beta - 1$ for h_k, since the latter is then integrated once. To include affine maps within the space $\mathcal{S}_\Delta^\beta$, we also include constant and linear functions with respect to each variable in the expansions (4.13). Computations in Section 4.5 using the transport map parameterizations above are performed using the publicly-available software TransportMaps.[2]

4.3.2 Optimization of the transport map

We complete the discussion of transport maps by describing the optimization procedure for finding the map. After defining a maximum polynomial degree for the basis functions within each map component, the resulting map $S_{\boldsymbol{\alpha}} \in \mathcal{S}_\Delta^\beta$ is parameterized by a finite number of coefficients $\boldsymbol{\alpha} \in \mathbb{R}^p$. In this subsection we include the subscript $\boldsymbol{\alpha}$ on S to emphasize the map's parametric dependence.

A computational approach to find the KR rearrangement that was explored in Marzouk, Moselhy, Parno, and Spantini [141] is to minimize the Kullback–Leibler divergence $D_{\mathrm{KL}}(\pi||S^\sharp\eta) = \mathbb{E}_\pi[\log(\pi/S^\sharp\eta)]$ from the pullback density $S^\sharp\eta$ to π over the space of monotone increasing triangular maps $\mathcal{S}_\Delta$. After parameterizing the maps

[2] http://transportmaps.mit.edu

with coefficients $\boldsymbol{\alpha}$, this is equivalent to solving

$$\begin{aligned}\boldsymbol{\alpha}^* &= \arg\min_{\boldsymbol{\alpha}} \; D_{\mathrm{KL}}\left(\pi || S_{\boldsymbol{\alpha}}^\sharp \eta\right) \\ &= \arg\max_{\boldsymbol{\alpha}} \; \mathbb{E}_\pi\left[\log S_{\boldsymbol{\alpha}}^\sharp \eta(\boldsymbol{X})\right] \end{aligned} \tag{4.14}$$

As shown in Marzouk, Moselhy, Parno, and Spantini [141] and Parno and Marzouk [163], for standard Gaussian η and lower triangular S, this optimization problem is separable across the map components $S_1, \ldots, S_d$. In addition, when using the parameterizations in Section 4.3.1 for each map component, the optimization problem is unconstrained and differentiable with respect to the coefficients $\boldsymbol{\alpha}$. Therefore, in practice we can use an iterative method such as BFGS to find the optimal solution [155]. While the problem is not in general convex in $\boldsymbol{\alpha}$, suitable choices of g in (4.12) can ensure that the problem has a unique global minimizer [16].

Given i.i.d. data $\{\boldsymbol{X}^l\}_{l=1}^n$ from π, we can approximate the expectation in (4.14) and maximize the likelihood associated with this data set. That is,

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= \arg\max_{\boldsymbol{\alpha}} \; \frac{1}{n}\sum_{l=1}^n \log S_{\boldsymbol{\alpha}}^\sharp \eta(\boldsymbol{X}^l), \\ &= \arg\max_{\boldsymbol{\alpha}} \; \frac{1}{n}\sum_{l=1}^n \sum_{k=1}^d \left[-\frac{1}{2}(S_{\boldsymbol{\alpha}})_k(\boldsymbol{X}_{\le k}^l)^2 + \log \partial_k (S_{\boldsymbol{\alpha}})_k(\boldsymbol{X}_{\le k}^l)\right]. \end{aligned} \tag{4.15}$$

where the last equality follows from the form of the standard Gaussian reference density η. The optimal solution $\widehat{\boldsymbol{\alpha}}$ in (4.15) is the maximum likelihood estimate (MLE) of $\boldsymbol{\alpha}^*$. Here we assume that the solutions of (4.14) and (4.15) are unique.

Under suitable regularity conditions on the log-likelihood in (4.15), $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}^*$. Furthermore, it is a random variable that converges in distribution as $n \to \infty$ to a normal random vector given by

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \Gamma(\boldsymbol{\alpha}^*)^{-1}\right), \tag{4.16}$$

where $\Gamma(\boldsymbol{\alpha}) \in \mathbb{R}^{p\times p}$ is the non-singular Fisher information matrix [36] for the transport

map representation of the density π with coefficients $\boldsymbol{\alpha}$. Entry (i,j) of the Fisher information matrix is given by

$$\Gamma(\boldsymbol{\alpha})_{ij} := -\mathbb{E}_\pi \left[\partial_{\alpha_i}\partial_{\alpha_j} \log S_{\boldsymbol{\alpha}}^\sharp \eta(\boldsymbol{X})\right]. \tag{4.17}$$

We conclude this section with a description of the closed-form solution to (4.15) when the coefficients parameterize transport maps $S_{\boldsymbol{\alpha}}(\boldsymbol{x})$ that are affine in $\boldsymbol{x}$.

Proposition 7 (Affine map optimization)**.** *Suppose π is an arbitrary continuous density on $\mathbb{R}^d$, and let $\{\boldsymbol{X}^l\}_{l=1}^n$ be samples drawn from π with $n \geq d$. If the map components are restricted to be affine functions of the input variables (i.e., polynomial degree $\beta = 1$), maximizing the log-likelihood function that follows from the pullback density in (4.15) yields a Gaussian approximation to π given by $S_{\widehat{\alpha}}^\sharp \eta = \mathcal{N}(\widehat{\mathbf{m}}, \widehat{\Sigma})$ with empirical mean $\widehat{\mathbf{m}} = \frac{1}{n}\sum_{l=1}^n \boldsymbol{X}^l$ and empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{l=1}^n (\boldsymbol{X}^l - \widehat{\mathbf{m}})(\boldsymbol{X}^l - \widehat{\mathbf{m}})^T$.*

4.3.3 Computation of Ω

After optimizing the coefficients, the pullback density $S_{\widehat{\boldsymbol{\alpha}}}^\sharp \eta$ defines an approximation to the target density π. The conditional independence score Ω_{ij} in (4.4) is then estimated as

$$\widehat{\Omega}_{ij} = \mathbb{E}_\pi |\partial_i \partial_j \log S_{\widehat{\boldsymbol{\alpha}}}^\sharp \eta(\boldsymbol{X})|^2. \tag{4.18}$$

We can approximate the expectation above using n i.i.d. samples $\{\boldsymbol{X}^l\}_{l=1}^n$ from π, yielding a sample-based estimator of $\widehat{\Omega}$:

$$\widetilde{\Omega}_{ij} = \frac{1}{n}\sum_{l=1}^n |\partial_i \partial_j \log S_{\widehat{\boldsymbol{\alpha}}}^\sharp \eta(\boldsymbol{X}^l)|^2. \tag{4.19}$$

In this work we estimate $\widehat{\Omega}$ using the same n samples from π as those used in (4.15) to estimate the map coefficients $\widehat{\boldsymbol{\alpha}}$. Reusing the samples produces a biased approximation of the non-zero values of $\widehat{\Omega}$ at finite n. We observe in our numerical experiments, however, that this bias has no significant impact on the estimation of the sparsity

pattern of Ω.

4.3.4 Threshold estimator of Ω

As a result of the sample-based approximations of S and $\widehat{\Omega}$, the sparsity pattern of $\widetilde{\Omega}$ will not exactly match that of Ω. For instance, a zero entry in Ω may result in a small but numerically non-zero entry in $\widetilde{\Omega}$. To account for this mismatch, we introduce a threshold estimator $\overline{\Omega}$ defined as

$$\overline{\Omega}_{ij} = \begin{cases} \widetilde{\Omega}_{ij}, & \widetilde{\Omega}_{ij} \geq \tau_{ij} \\ 0, & \text{otherwise} \end{cases}, \tag{4.20}$$

for some $\tau_{ij} > 0$. Threshold estimators are commonly used for sparse covariance matrix estimation (see Cai and Liu [30]) and τ_{ij} is usually chosen proportional to the standard deviation of $\widetilde{\Omega}_{ij}$. The rationale behind this choice is to threshold the entries of $\widetilde{\Omega}$ whose standard deviation makes them indistinguishable from zero.

To compute the standard deviation or variance of $\widetilde{\Omega}_{ij}$, empirical estimation is not feasible since we only have a unique realization of $\widetilde{\Omega}_{ij}$. Instead we approximate its variance with

$$\mathbb{V}(\widetilde{\Omega}_{ij}) \approx \frac{1}{n} \left(\nabla_{\boldsymbol{\alpha}} \widetilde{\Omega}_{ij}\right)^T \Gamma(\boldsymbol{\alpha})^{-1} \left(\nabla_{\boldsymbol{\alpha}} \widetilde{\Omega}_{ij}\right) \Big|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}} := \frac{\widetilde{v}_{ij}^2}{n}, \tag{4.21}$$

where $\Gamma(\boldsymbol{\alpha})$ is the Fisher information in (4.17) and $\nabla_{\boldsymbol{\alpha}} \widetilde{\Omega}_{ij}|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}}$ denotes the gradient of $\boldsymbol{\alpha} \mapsto \widetilde{\Omega}_{ij}(\boldsymbol{\alpha}) = \frac{1}{n} \sum_{i=1}^n |\partial_i \partial_j \log S_{\boldsymbol{\alpha}}^\sharp \eta(\boldsymbol{X}^l)|^2$ evaluated at $\widehat{\boldsymbol{\alpha}}$. We assume here that $\boldsymbol{\alpha} \mapsto \widetilde{\Omega}_{ij}(\boldsymbol{\alpha})$ is a continuously differentiable function of the parameters $\boldsymbol{\alpha}$ for each entry (i, j). The variance approximation in (4.21) is inspired by the delta method [156], which exploits the fact that, given a sequence of random variables $\theta_n \in \mathbb{R}^p$ satisfying $\sqrt{n}\,(\theta_n - \theta) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \Lambda)$ and a continuously differentiable function $g \colon \mathbb{R}^p \to \mathbb{R}$ such that $\nabla g(\theta) \neq 0$, we have $\sqrt{n}\,(g(\theta_n) - g(\theta)) \xrightarrow{D} \mathcal{N}\left(0, \nabla g(\theta)^T \Lambda \nabla g(\theta)\right)$. Here we consider θ_n to be the map coefficients $\widehat{\boldsymbol{\alpha}}$ and g to be the sample-based score estimator $\widetilde{\Omega}_{ij}$. The main difference between (4.21) and the variance predicted by the delta

method is that we evaluate the gradients and the Fisher information at the estimated map coefficients $\widehat{\boldsymbol{\alpha}}$ because the limiting coefficients $\boldsymbol{\alpha}^*$ are unknown in practice.

For the threshold in (4.20), we then use

$$\tau_{ij} = f(n)\frac{\widetilde{v}_{ij}}{\sqrt{n}}, \tag{4.22}$$

where $f(n)$ is a function that increases slowly with n. In our numerical experiments, we will choose $f(n) \propto \sqrt{\log n}$. In Section 4.3.6, we will justify this choice by identifying a class of functions f that make the resulting threshold estimator $\overline{\Omega}$ a *consistent* estimator of the sparsity pattern of Ω as $n \to \infty$.

4.3.5 Non-iterative SING algorithm

The tools above suffice to build an algorithm to learn the Markov structure of a continuous, non-Gaussian distribution. Algorithm 3 learns the graph structure from the support of the threshold estimator $\overline{\Omega}$ for the conditional independence score. We refer to this sequence of steps as the non-iterative Sparsity Identification in Non-Gaussian distributions (N-SING) algorithm. In Section 4.4 we propose an iterative version of this algorithm that uses the sparsity of $\overline{\Omega}$ to define improved estimators for the map and the resulting score matrix Ω.

Algorithm 3: Non-iterative Sparsity Identification in Non-Gaussian distributions (N-SING)

Input : i.i.d. samples $\{\boldsymbol{X}^l\}_{l=1}^n \sim \pi$, maximum polynomial degree β, threshold scaling f

Output: Edge set $\widehat{E}_n$ of minimal I-map for $\boldsymbol{\nu}_\pi$

1 Compute transport map: $S_{\widehat{\boldsymbol{\alpha}}} = \arg\max_{S_{\boldsymbol{\alpha}} \in \mathcal{S}_\Delta^\beta} \sum_{l=1}^n \log S_{\boldsymbol{\alpha}}^\sharp \eta(\boldsymbol{X}^l)$

2 Estimate Ω: $\widetilde{\Omega}_{ij} = \frac{1}{n}\sum_{l=1}^n |\partial_i \partial_j \log S_{\widehat{\boldsymbol{\alpha}}}^\sharp \eta(\boldsymbol{X}^l)|^2$

3 Threshold $\widetilde{\Omega}$, with $\tau_{ij} = f(n)\widetilde{v}_{ij}/\sqrt{n}$, to yield $\overline{\Omega}$

4 Compute $\widehat{E}_n$: $(i,j) \in \widehat{E}_n$ if $\overline{\Omega}_{ij} \neq 0$

4.3.6 Analysis of consistency

Here we establish conditions under which the proposed threshold estimator is consistent for recovering the edge set E of the minimal I-map for $\boldsymbol{\nu}_\pi$. For simplicity, we consider a variant of N-SING that uses an exact expectation $\widehat{\Omega}_{ij} = \mathbb{E}_\pi |\partial_i \partial_j \log S^\sharp_{\widehat{\boldsymbol{\alpha}}} \eta(\boldsymbol{X})|^2$ in step 2 of Algorithm 3. Using the finite-sample approximation $\widetilde{\Omega}_{ij}$ would complicate but not fundamentally change the following analysis. We also assume that the map parameterization is sufficiently rich to recover the target density exactly, i.e., that there exists a map $S_{\boldsymbol{\alpha}^*} \in \mathcal{S}^\beta_\Delta$ with coefficients $\boldsymbol{\alpha}^* \in \mathbb{R}^p$, for some polynomial degree β, such that $\pi = (S_{\boldsymbol{\alpha}^*})^\sharp \eta$, where $\boldsymbol{\alpha}^*$ is obtained from the solution of (4.14). Intuitively, since the maximum likelihood estimate $\widehat{\boldsymbol{\alpha}}$ converges to $\boldsymbol{\alpha}^*$ as $n \to \infty$ (recall (4.16)), each entry $\widehat{\Omega}_{ij}$ of the estimated score matrix for a sufficiently smooth map S should then converge to the true score Ω_{ij}:

$$\widehat{\Omega}_{ij} = \mathbb{E}_\pi |\partial_i \partial_j \log S^\sharp_{\widehat{\boldsymbol{\alpha}}} \eta(\boldsymbol{X})|^2 \quad \underset{n\to\infty}{\longrightarrow} \quad \mathbb{E}_\pi |\partial_i \partial_j \log S^\sharp_{\boldsymbol{\alpha}^*} \eta(\boldsymbol{X})|^2 = \Omega_{ij}.$$

We will formalize this notion within the analysis below. To recover the support of Ω from $\widehat{\Omega}$, we consider the threshold estimator

$$\overline{\Omega}_{ij} = \widehat{\Omega}_{ij} \, \mathbb{1}(\widehat{\Omega}_{ij} > \tau_{ij}), \qquad \text{with } \tau_{ij} = f(n)\widehat{v}_{ij}/\sqrt{n}, \tag{4.23}$$

where $\widehat{v}^2_{ij} \coloneqq (\nabla_{\boldsymbol{\alpha}} \widehat{\Omega}_{ij})^T \Gamma(\boldsymbol{\alpha})^{-1} (\nabla_{\boldsymbol{\alpha}} \widehat{\Omega}_{ij})|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}}$.

A proper choice of f is critical to guaranteeing that the support of the threshold estimator $\overline{\Omega}$ converges to the support of Ω with increasing n. For instance, when $\Omega_{ij} = 0$, both $\widehat{\Omega}_{ij}$ and $\widehat{v}_{ij}/\sqrt{n}$ go to zero at the same rate as $n \to \infty$. As a result, the event $\{\widehat{\Omega}_{ij} > \widehat{v}_{ij}/\sqrt{n}\}$ asymptotically occurs with a constant non-zero probability, resulting in false positive edges. The role of $f(n)$ in this case is to adjust the rate of convergence of the threshold to ensure that $\widehat{\Omega}_{ij} < \tau_{ij}$ asymptotically, i.e., that there are no false positives. A similar argument holds for false negatives. The following proposition gives sufficient conditions on f to guarantee the recovery of the edge set in the minimal I-map for $\boldsymbol{\nu}_\pi$. The proof is provided in Appendix C.

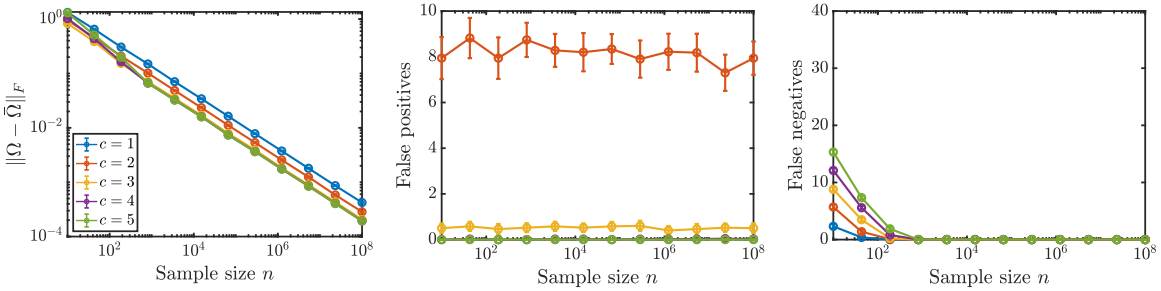
Proposition 8. *For the threshold estimator (4.23), let f be a function such that $f(n) \to \infty$ and $f(n)/\sqrt{n} \to 0$ as $n \to \infty$. Then the edge set $\widehat{E}_n$ returned by the associated* N-SING *algorithm is a consistent estimator of E, i.e., $\mathbb{P}(\widehat{E}_n = E) \to 1$ as $n \to \infty$.*

The main idea behind this result is to show that the probability of false positives (i.e., type 1 errors) or false negatives (i.e., type 2 errors) converges to zero with the prescribed threshold. As demonstrated above, the occurrence of these events is determined by the magnitude of the estimated score $\widehat{\Omega}_{ij}$ relative to the threshold τ_{ij}, or, equivalently, by the magnitude of the ratio $\sqrt{n}\widehat{\Omega}_{ij}/\widehat{v}_{ij}$ in comparison to $f(n)$. For each pair (i, j), we consider the function $g(\boldsymbol{\alpha}) = \mathbb{E}_\pi|\partial_i\partial_j \log S^\sharp_{\boldsymbol{\alpha}}\eta(\boldsymbol{x})|^2$ and analyze the asymptotic statistics of the ratio $\sqrt{n}\widehat{\Omega}_{ij}/\widehat{v}_{ij} = \sqrt{n}g(\widehat{\boldsymbol{\alpha}})/(\nabla_{\boldsymbol{\alpha}}g(\widehat{\boldsymbol{\alpha}})^T\Gamma(\widehat{\boldsymbol{\alpha}})^{-1}\nabla_{\boldsymbol{\alpha}}g(\widehat{\boldsymbol{\alpha}}))^{1/2}$ as a function of the estimated map coefficients $\widehat{\boldsymbol{\alpha}}$. When the variables X_i and X_j are conditionally dependent, we have $g(\boldsymbol{\alpha}^*) = \Omega_{ij} \neq 0$, and we assume that the gradient $\nabla_{\boldsymbol{\alpha}}g(\boldsymbol{\alpha}^*) \neq 0$. In this case, the limiting distribution of the ratio is Gaussian by an application of the delta method. On the other hand, when X_i and X_j are conditionally independent, we have not only $g(\boldsymbol{\alpha}^*) = \Omega_{ij} = 0$ but also $\nabla_{\boldsymbol{\alpha}}g(\boldsymbol{\alpha}^*) = 0$, because both of these terms depend on $\partial_i\partial_j \log \pi(\boldsymbol{x})$, which is zero for all $\boldsymbol{x} \in \mathbb{R}^d$. Thus both $\widehat{\Omega}_{ij}$ and $\widehat{v}_{ij}$ approach zero, and their ratio becomes singular as $n \to \infty$. Now, the delta method is no longer valid; instead, we consider the asymptotic distribution of singular Wald statistics, as analyzed in [61, 166]. Here, to characterize the limiting distribution, we assume that $\nabla^2_{\boldsymbol{\alpha}}g(\boldsymbol{\alpha}^*) \neq 0$, but higher-order derivatives could also be considered if this condition does not hold. Under the asymptotic distributions for these two scenarios, we show that the probabilities of false positives and false negatives converge to zero given the conditions above on the function f in (4.23).

While Proposition 8 guarantees the threshold estimator is consistent for *any* f satisfying the criteria above, the selected f may affect the algorithm's finite-sample performance. A function f that grows more quickly with n will produce higher thresholds and reduce the probability of false positive edges, while a more slowly growing function will produce lower thresholds and reduce the probability of false negative edges. Future work will investigate the impact of these choices on finite-

sample bounds, i.e., the number of samples required by the algorithm to recover the graph with high probability.

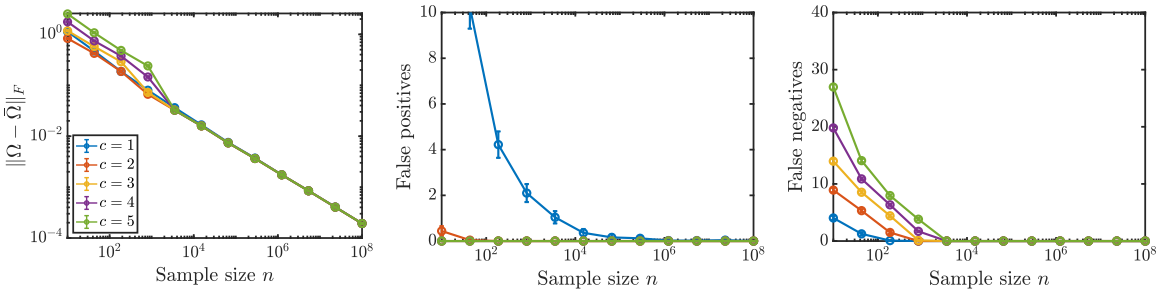
To validate the consistency of the proposed thresholds, we consider a simple model for recovering the sparsity of a matrix. Let Ω be a positive matrix with sparse nonzero entries that is corrupted with additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \rho)$ where $\rho \propto 1/\sqrt{n}$ (as we have from the delta method) so that we observe $\widehat{\Omega} = \Omega + \epsilon$. To recover the sparsity of Ω, we define the threshold estimator $\overline{\Omega} = \widehat{\Omega}\mathbb{1}(\widehat{\Omega} > \tau)$ and compare the sparsity of $\overline{\Omega}$ and Ω. Figures 4-1-4-3 plot the error in Frobenius norm $\|\Omega - \overline{\Omega}\|_F$, the Type 1 (false positive) errors, and the Type 2 (false negative) errors as a function of n for three different threshold scaling functions $\tau = c\rho$, $\tau = c\sqrt{\log(n)}\rho$, and $\tau = c\sqrt{n}\rho$, respectively, where c is a positive constant. While some of the thresholds produce estimators that converge to Ω entrywise (i.e., in the Frobenius norm), the threshold that doesn't satisfy $f(n) \to \infty$ has a constant number of false positives as we increase the sample size, while the threshold that is constant with respect to n has a constant number of false negatives for large n. On the other hand, both Type 1 and Type 2 errors decrease to zero for the threshold $\tau = c\sqrt{\log(n)}\rho$ that satisfies the conditions in Proposition 8. Furthermore, this threshold appears to be less sensitive to the choice of constant c, than the other thresholds.



(a) (b) (c)

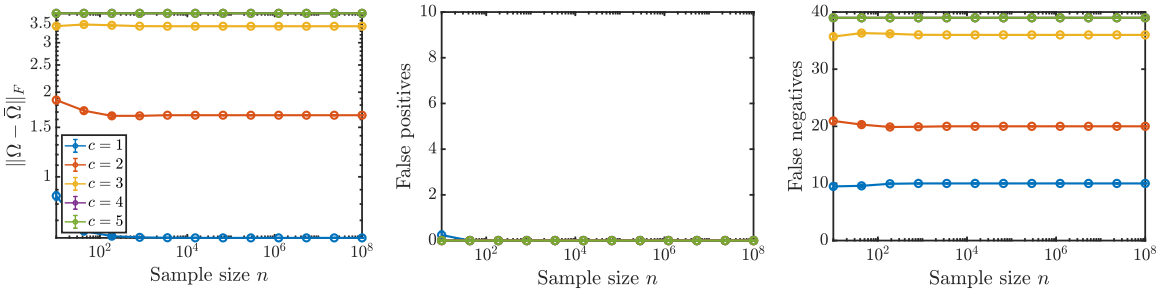
Figure 4-1: (a) Convergence of $\widehat{\Omega}$ in Frobenius norm for threshold $\tau = c/\sqrt{n}$; (b) Convergence of Type 1 (false negative errors) (c) Convergence of Type 2 (false negative errors)

As a final validation, we investigate the performance of different thresholding functions for estimating the score matrix Ω in the context of the N-SING algorithm. To do so, we consider a $d = 10$-dimensional Gaussian target density $\pi(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \Sigma)$ with



(a) (b) (c)

Figure 4-2: (a) Convergence of $\widehat{\Omega}$ in Frobenius norm for threshold $\tau = c\log(n)/\sqrt{n}$; (b) Convergence of Type 1 (false negative errors) (c) Convergence of Type 2 (false negative errors)



(a) (b) (c)

Figure 4-3: (a) Convergence of $\widehat{\Omega}$ in Frobenius norm for threshold $\tau = c\sqrt{n}/\sqrt{n}$; (b) Convergence of Type 1 (false negative errors) (c) Convergence of Type 2 (false negative errors)

inverse covariance $\Sigma^{-1} = \mathbf{L}^T\mathbf{L}$ where $\mathbf{L} \in \mathbb{R}^{d\times d}$ is a sparse lower-triangular matrix. Then, by Proposition 7, $S(\boldsymbol{x}) = \mathbf{L}\boldsymbol{x}$ is the transport map that pulls back the standard Gaussian density η to π. Given n i.i.d. samples from π, a maximum likelihood estimator $\widehat{\mathbf{L}}$ of $\mathbf{L}$ is found using (4.15). The MLE is asymptotically distributed as $\widehat{\mathbf{L}} \sim \mathcal{N}(\mathbf{L}, \Gamma/n)$ by 4.16 where Γ is the Fisher information matrix for the estimated parameters $\mathbf{L}$. In this experiment, we set L to an identity matrix for the first $d-1$ rows and randomly sample the dth row using independent standard Gaussian entries for half of the off-diagonal elements and a uniform random variable between $[0, 1]$ for the diagonal element. As a result, the first $d-1$ components of $\mathbf{X}$ are standard Gaussian and the unknown map parameters $\boldsymbol{\alpha}$ correspond to the elements in the last row of L.

We now test the threshold estimator for Ω assuming that we are in an asymptotic

regime where the MLE for the unknown map parameters follows a Gaussian distribution. Given a sample of $\widehat{\boldsymbol{\alpha}}$ from this asymptotic distribution for each n, we estimate the entries in the score matrix $\widehat{\Omega}_{dj} = (\widehat{\mathbf{L}}_{dj}\widehat{\mathbf{L}}_{dd})^2$ for $j = 1, \ldots, d$. Let us remark that for a linear map (i.e., a multivariate Gaussian pullback density), the Hessian of the log-density is constant with respect to $\boldsymbol{x}$ and computing the last row of the score matrix does not require a Monte Carlo estimator over the support of π, i.e., $\widehat{\Omega} = \widetilde{\Omega}$. For each estimator $\widehat{\Omega}$, we compute the variances of the score matrix entries using (4.21) and construct the threshold estimator $\overline{\Omega}$ in (4.20). Figure 4-4 plots the probability of false positive and negative errors for the parameters $\boldsymbol{\alpha}$ with different threshold scaling functions $f(n)$. The probabilities are estimated empirically for each sample size n given 1000 samples of $\widehat{\mathbf{L}}$. We observe that the functions $f(n) = \sqrt{\log n}$ and $f(n) = n^{1/4}$, which satisfy the conditions in Proposition 8, result in correct recovery of the edges as $n \to \infty$, while the remaining choices result in inconsistent threshold estimators.

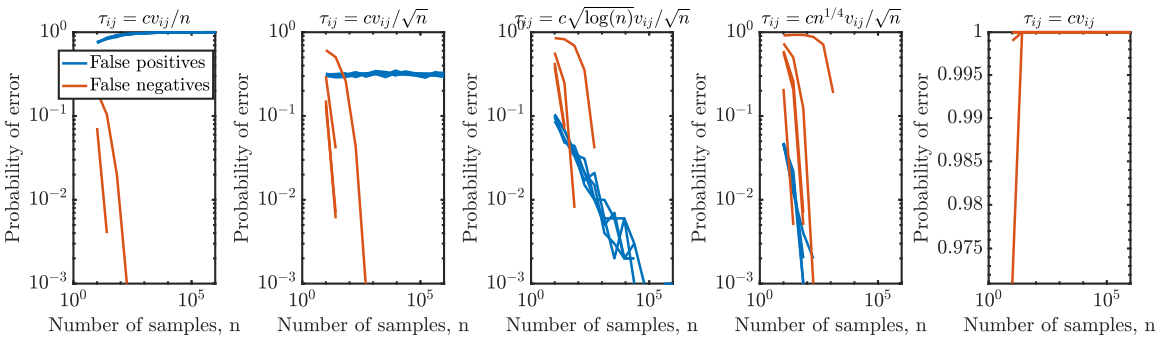


Figure 4-4: False positive and negative errors with different sample size scalings for the threshold of the Ω estimator in the SING algorithm with Gaussian samples

4.4 Improved estimator for Markov structure

In Section 4.3 we showed how to estimate the conditional independence score using a transport map representation of the target density. This transport map is a lower-triangular function and thus is a sparse map where each component does not depend on all of its input variables. However, when the target measure $\boldsymbol{\nu}_\pi$ satisfies conditional independence properties, the transport map we seek to approximate can inherit

additional sparse structure [see 204, Section 5]. In this Section, we take advantage of this connection between the sparsity of the graph and the sparsity of the transport map to present an iterative algorithm for learning the graph.

4.4.1 Sparsity of the transport map

For a lower triangular function S, the sparsity pattern of the map, $\mathcal{I}_S$, is defined in Spantini, Bigoni, and Marzouk [204] as:

$$\mathcal{I}_S \coloneqq \{(j,k) : j < k, \partial_j S_k = 0\}. \tag{4.24}$$

That is, the sparsity pattern is the set of all integer pairs (j,k) with $j < k$, such that the kth component of the map does not depend on the jth input variable.[3] The complement of this set, i.e.,

$$\mathcal{I}_S^c \coloneqq \{(j,k) : j < k, \partial_j S_k \neq 0\}, \tag{4.25}$$

determines the *active variables* of the map. That is, if $(j,k) \in \mathcal{I}_S^c$, then the kth component of the map must depend on the jth input variable. We denote the set of lower triangular maps that respect the sparsity pattern given by $\mathcal{I}_S$ as $\mathcal{S}_{\mathcal{I}_S} \subset \mathcal{S}_\Delta$.

Given a target density π, Spantini, Bigoni, and Marzouk [204] showed that the Markov structure of $\boldsymbol{\nu}_\pi$ yields a *tight lower bound* on the sparsity pattern $\mathcal{I}_S$ of the KR rearrangement that pulls back η to π. Knowledge of this sparsity can be used when solving the variational problem in (4.14) by restricting the feasible domain to transport maps with a reduced set of active variables. To determine this sparsity pattern, we perform a series of graph operations on the minimal I-map $\mathcal{G}$ of the target measure $\boldsymbol{\nu}_\pi$. These operations define the active variables for each map component based on a sequence of intermediate graphs $(\mathcal{G}^k)$. The graph $\mathcal{G}^k$ is identical to the graph obtained in the variable elimination algorithm before marginalizing node k according to the elimination ordering $(d, d-1, \ldots, 1)$. However, we emphasize that this

[3] The lower triangular function also satisfies $\partial_j S_k = 0$ for all $j > k$ by construction.

sparsity pattern is identified only by inspecting the graph, without actually performing variable elimination or additional computation (e.g., marginalization) on the joint density. We restate the relevant part of this result below.[4]

Theorem 4.4.1 (Spantini, Bigoni, and Marzouk [204], Theorem 3 (Part 1))**.** *Let* $\mathbf{Z} \sim \boldsymbol{\nu}_\eta$, $\mathbf{X} \sim \boldsymbol{\nu}_\pi$ *with Lebesgue absolutely continuous measures* $\boldsymbol{\nu}_\eta$, $\boldsymbol{\nu}_\pi$, *and let* $\boldsymbol{\nu}_\eta$ *be a product measure on* $\mathbb{R}^d$. *Moreover, assume that* $\boldsymbol{\nu}_\pi$ *is globally Markov with respect to* $\mathcal{G}$, *and define, recursively, the sequence of graphs* $(\mathcal{G}^k)_{k=1}^d$ *as: (1)* $\mathcal{G}^d := \mathcal{G}$ *and (2) for all* $1 \leq k < d$, $\mathcal{G}^{k-1}$ *is obtained from* $\mathcal{G}^k$ *by removing node* k *and by turning its neighborhood* $Nb(k, \mathcal{G}^k)$ *into a clique. Then the following holds:*

1. *If* $\mathcal{I}_S$ *is the sparsity pattern of the transport map* S *pushing forward* $\mathbf{X}$ *to* $\mathbf{Z}$, *then*

$$\widehat{\mathcal{I}}_S \subset \mathcal{I}_S, \tag{4.26}$$

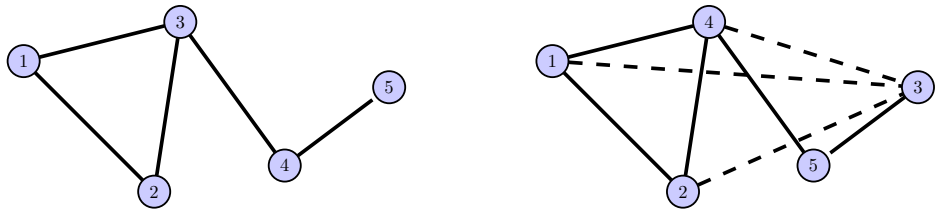
   where $\widehat{\mathcal{I}}_S$ *is the set of integer pairs* (j, k) *such that* $j \notin Nb(k, \mathcal{G}^k)$.

4.4.2 Ordering variables in the map

In the process above, the sparsity pattern of the map decreases (relative to that of the original I-map $\mathcal{G}$) when adding edges to the intermediate graphs $\mathcal{G}^k$ to create cliques. These edges produce *fill-in*. Fill-in will occur unless $\mathcal{G}$ is chordal *and* the variable ordering corresponds to the perfect elimination ordering [184]. Whether or not $\mathcal{G}$ is chordal, the amount of fill-in is dependent on the ordering of the input variables. For example, consider the graph and variable ordering in Figure 4-5a. The corresponding lower bound for the sparsity pattern of the map is

$$\widehat{\mathcal{I}}_S = \{(1,4), (2,4), (1,5), (2,5), (3,5)\}, \tag{4.27}$$

[4] Parts 2 and 3 of Theorem 3 in Spantini, Bigoni, and Marzouk [204] provide sparsity bounds for the transport map S^{-1}, which are related to the marginal independence of $\boldsymbol{\nu}_\pi$ and do not concern the current work.



(a) (b)

Figure 4-5: (a) A sparse graph with an optimal node ordering; (b) Suboptimal ordering induces extra edges.

and the dependence of each map component on the input variables is

$$S(\boldsymbol{x}) = \begin{bmatrix} S_1(x_1) \\ S_2(x_1, x_2) \\ S_3(x_1, x_2, x_3) \\ S_4(\quad\quad x_3, x_4) \\ S_5(\quad\quad\quad x_4, x_5) \end{bmatrix}. \tag{4.28}$$

With the variable ordering shown in Figure 4-5a, no edges are added during the process to identify $\widehat{\mathcal{I}}_S$ and the resulting transport map is more sparse than a dense lower-triangular map. In contrast, using the suboptimal ordering shown in Figure 4-5b, edges must be added to the induced graph, shown in dashed lines. The associated sparsity pattern is now $\widehat{\mathcal{I}}_S = \{(1,5),(2,5)\}$, and hence the transport map is predicted to be less sparse.

With a larger set $\widehat{\mathcal{I}}_S$, we can simplify the parameterization of the map and reduce the size of the coefficient vector $\boldsymbol{\alpha}$; this in turn reduces the variance of the estimator $\widehat{\boldsymbol{\alpha}}$ (4.15), for any given sample size n [see demonstrations in 150, Section 3.2]. Thus, it is worthwhile to find a variable ordering that maximizes the sparsity of the map. A variable ordering is equivalent to a permutation $\varphi\colon [d] \to [d]$ of the nodes in the graph. [204] presented an optimization problem for identifying the permutation that induces the least fill-in. This optimization problem is in general NP-complete [236]. Nevertheless, several heuristics have been proposed to construct permutations based only on the graph $\mathcal{G}$, including weighted min-fill and min-degree orderings [112].

Another that we have found to perform well in practice is the reverse of the ordering used by sparse Cholesky factorization algorithms [187]. We will use this ordering for the numerical examples in Section 4.5.

4.4.3 The iterative SING algorithm

In Section 4.4.1 we saw how sparsity in the graph implies sparsity in the transport map. To take advantage of this sparsity, we need to know the Markov structure of the target density π. This Markov structure is not available *a priori* given only samples from π. However, the N-SING algorithm from Section 4.3 can provide an initial estimate of this Markov structure, based on a dense lower triangular map S (i.e., with $\mathcal{I}_S = \emptyset$). This Markov structure can be used to identify a new variable ordering, and a corresponding bound $\widehat{\mathcal{I}}_S$ on the sparsity pattern of the transport map via Theorem 4.4.1. We can then enforce this sparsity bound to compute a new estimate of the transport map, and in turn obtain a new threshold estimate of the Markov structure of π. We repeat this procedure until the sparsity of the estimated graph no longer changes. This defines an iterative application of N-SING that we call the SING algorithm.[5] The formal steps of SING are presented in Algorithm 4.

The first steps of SING are identical to the N-SING algorithm. The rationale for the remaining steps, and for subsequent iterations, is essentially to exploit sparsity for variance reduction. Sparsity in the graph, coupled with a good variable ordering, leads to sparsity in the map, i.e., a smaller number of active variables. As noted above, such a map can be described by fewer coefficients $\boldsymbol{\alpha}$; a maximum likelihood estimate of these coefficients—based on the same finite sample from π as in previous iterations—in turn has a smaller variance, as observed in [150]. Thus, the iterations of the SING algorithm provide improved estimators of the target density and of the conditional independence score for the goal of learning the graph.

Remark. *For a multivariate Gaussian target density $\pi = \mathcal{N}(\mathbf{0}, \Sigma)$, the* SING *algorithm with affine maps (i.e., polynomial degree $\beta = 1$) alternates between estimating a*

[5] The SING algorithm first appeared with slight modifications in Morrison, Baptista, and Marzouk [150].

sparse Cholesky factor of the inverse covariance matrix (see Proposition 7 and Example 1) and defining a threshold estimator for Σ^{-1} to learn the Markov structure of π. The sparsity of the Cholesky factor is dependent on the sparsity of Σ^{-1} and the ordering of the input variables. Recently, several methods have also been proposed to learn sparse Cholesky factors of a sparse inverse covariance matrix for the goal of density estimation in the Gaussian setting. These methods are based on ℓ_1-penalized maximum likelihood estimation [96], banded sparsity patterns for the Cholesky factor [20, 126], and combinations of multiple variable orderings [106].

To conclude, let us comment on the stopping criterion used in Algorithm 4. During the procedure, the edges in the estimated graph can change freely. For instance, we do not impose any constraint on the edge set $\widehat{E}^t$ at iteration t (e.g., $\widehat{E}^t \subseteq \widehat{E}^{t-1}$) or on the sparsity pattern of the map. Nevertheless, we have observed in most of our numerical experiments that the cardinality of estimated edges $|\widehat{E}^t|$ decreases until the algorithm finds a good estimate for E. Thus, checking that $|\widehat{E}^t|$ is non-decreasing works well as a practical stopping criterion.

Algorithm 4: Sparsity Identification in Non-Gaussian distributions (SING)

Input : i.i.d. samples $\{\boldsymbol{X}^l\}_{l=1}^n \sim \pi$, maximum polynomial degree β, threshold scaling f

Output: Edge set $\widehat{E}_n$ of minimal I-map for $\boldsymbol{\nu}_\pi$

Define : $\mathcal{S}^\beta_{\widehat{\mathcal{I}}^1} = \mathcal{S}^\beta_\Delta$, $t = 1$

while $|\widehat{E}^t|$ *is decreasing* **do**

- Compute transport map: $S_{\widehat{\boldsymbol{\alpha}}} = \arg\max_{S_{\boldsymbol{\alpha}} \in \mathcal{S}^\beta_{\widehat{\mathcal{I}}^t}} \sum_{l=1}^n \log S^\sharp_{\boldsymbol{\alpha}} \eta(\boldsymbol{X}^l)$
- Estimate Ω: $(\widetilde{\Omega}^t)_{ij} = \frac{1}{n}\sum_{l=1}^n |\partial_i \partial_j \log S^\sharp_{\widehat{\boldsymbol{\alpha}}} \eta(\boldsymbol{X}^l)|^2$
- Threshold $\widetilde{\Omega}^t$ with $\tau_{ij} = f(n)(\widetilde{v}^t)_{ij}/\sqrt{n}$, to yield $\overline{\Omega}^t$
- Compute $|\widehat{E}^t|$ where $(i,j) \in \widehat{E}^t$ if $(\overline{\Omega}^t)_{ij} \neq 0$
- Find permutation of the variables φ^{t+1} (e.g., using reverse Cholesky ordering)
- Identify sparsity pattern of subsequent map $\widehat{\mathcal{I}}^{t+1}$
- $t \leftarrow t+1$

Set $\widehat{E}_n = \widehat{E}^t$

4.5 Numerical examples

In this section we apply the SING algorithm to learn the Markov structure of several non-Gaussian datasets. Section 4.5.1 presents results for the butterfly distribution and demonstrates the value of an iterative algorithm for recovering the Markov structure. Section 4.5.2 applies SING to data from nonparanormal distributions that were considered in Liu, Lafferty, and Wasserman [131]. Surprisingly, affine transport map approximations to the target density (i.e., with polynomial degree $\beta = 1$) work well for these nonparanormal examples, even with highly non-Gaussian marginals. In Section 4.5.3 we use an analytical example to further investigate why a linear map might still work for non-Gaussian data. We then examine a generalization of the nonparanormal setting in Section 4.5.4, where the target distribution is given by a diagonal transformation of a non-Gaussian base distribution. In these examples, we use the approximation class of the transport map that represents the base density also to approximate the target density, and still recover the correct graph. Finally, in Section 4.5.5 we consider a higher-dimensional physics-based dataset arising from the Lorenz-96 dynamical system.

In all of our numerical experiments, we parameterize the transport maps using the integrated squared representation introduced in Section 4.3.1, thereby enforcing the monotonicity of each map component by construction. Within the SING algorithm, we employ the reverse Cholesky ordering as a heuristic to order the variables in the map. Following the analysis in Section 4.3.6, we set the threshold to $f(n) = c\sqrt{\log n}$ where $c \in \mathbb{R}$ is a constant. While c can be chosen via cross-validation to improve empirical performance, any value is consistent for learning the graph, and we set $c = 1$ in our numerical investigations. Before running SING, we standardize each variable in the dataset by subtracting the empirical mean and dividing by the empirical standard deviation. This normalization step ensures all of the variables are on a similar scale and improves empirical performance.

To quantitatively evaluate the results of the SING algorithm for recovering the Markov structure using samples from π, we measure the errors in the estimated edge

sets $\widehat{E}$. For each graph, we measure the number of false positives: edges that are in $\widehat{E}$ and not in E (Type 1 errors) and the number of false negatives: edges in E that are not in $\widehat{E}$ (Type 2 errors). In the figures below, we report the mean type 1 and 2 errors across 25 runs of the algorithm with independent batches of samples, as well as the 95% confidence interval for the mean.

4.5.1 Butterfly distribution

The first example consists of r i.i.d. pairs of random variables (P_i, Q_i), where:

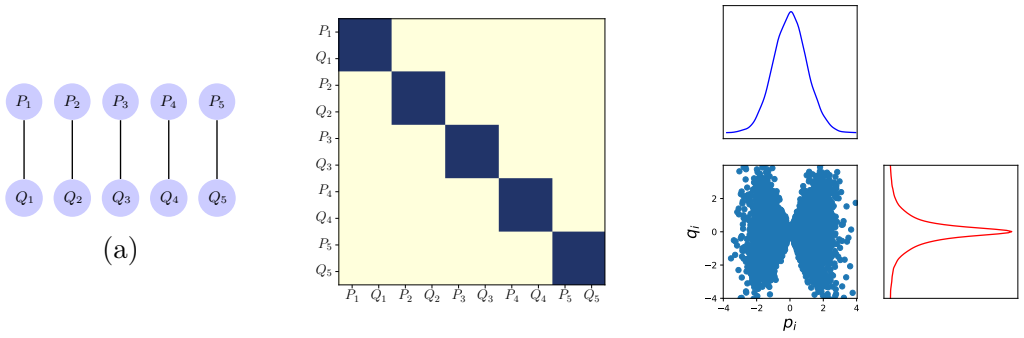
$$P_i \sim \mathcal{N}(0,1) \tag{4.29}$$

$$Q_i = W_i P_i, \quad \text{with } W_i \sim \mathcal{N}(0,1), \quad W_i \perp\!\!\!\perp P_i, \quad i = 1, \ldots, r. \tag{4.30}$$

One such pair of random variables, or a variation of the above, is a commonly used example to illustrate how two random variables can be uncorrelated but not independent.

Figures 4-6a–4-6b show the minimal I-map and corresponding adjacency matrix of the graph for $r = 5$ pairs, with the variables ordered as $P_1, Q_1, \ldots, P_5, Q_5$. Figure 4-6c shows the one- and two-dimensional marginal densities for one pair (P_i, Q_i). Each one-dimensional marginal is symmetric and unimodal, but the two-dimensional marginal (shown as samples) displays strongly non-Gaussian behavior.

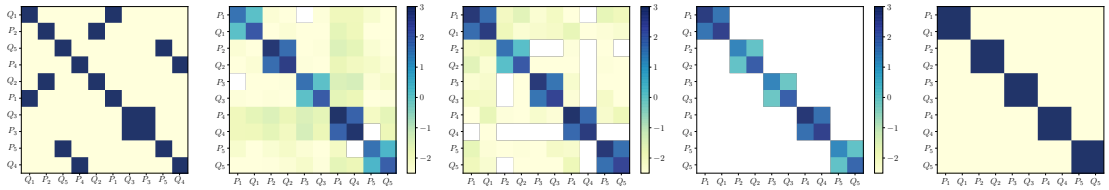
Figure 4-7 shows the progression of the identified graph (based on the sparsity of the estimated conditional independence score) over the iterations of SING, with $n = 2000$ samples and a polynomial degree $\beta = 3$. The variables in the data set are initially permuted, to verify that SING identifies a good ordering. After the first iteration of the algorithm (the output of the N-SING algorithm), the estimator for the conditional independence score has the block diagonal pattern in Figure 4-7b, but the off-diagonals of $\overline{\Omega}$ are not yet zero, resulting in many extra edges. In the next iterations, the algorithm leverages the sparsity of the graph estimated thus far to reveal sparsity in the transport map and improve the estimator for Ω in Figures 4-7c and 4-7d, thereby removing all erroneous edges. After the fifth iteration, the sparsity



(a) (b) (c)

Figure 4-6: (a) The undirected graphical model; (b) Adjacency matrix of true graph (dark blue corresponds to an edge, off-white to no edge); (c) One- and two-dimensional marginal densities for one pair (P_i, Q_i).

of the graph (and thus the size of the edge set) has not changed and the algorithm returns the correct graph in Figure 4-7e.



(a) (b) (c) (d) (e)

Figure 4-7: (a) Adjacency matrix of original graph under random permutation; (b) Entrywise logarithm of the thresholded score matrix $\overline{\Omega}$ (white indicates a zero entry for $\overline{\Omega}$) after iteration 1; (c) Iteration 3; (d) Iteration 5; (e) Adjacency matrix of the graph after iteration 5. SING returns the correct graph with $\beta = 3$.

There are two important conclusions from this example. First, assuming normality of the data returns the incorrect graph, as displayed in Figure 4-8. If the data were normal, then a linear map ($\beta = 1$) would suffice. However, SING with $\beta = 1$ yields not only an incorrect graph, but in fact fails to detect any edges at all. This result does not improve with increasing n, and incorrectly implies that all ten variables are marginally independent. Furthermore, this result indicates that methods based on precision matrix estimation that assume the data to be Gaussian (e.g., GLASSO) will also fail to recover the correct graph in this example, as demonstrated in [150].

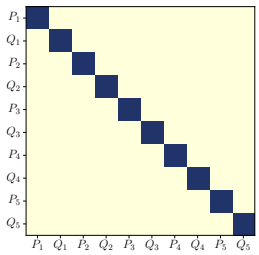


Figure 4-8: When the data are assumed normal ($\beta = 1$), all ten variables are incorrectly found to be independent.

Second, the density estimation problem here is quite challenging: the exact transport map that characterizes π in this case would in fact require an infinite expansion of polynomials. That is, no transport map with finite polynomial degree β can perfectly represent π. However, the correct graph can be identified with $\beta = 3$. Thus, in this case, allowing for some nonlinearity in the map is sufficient. A linear map does not work at all, but β also need not be excessively high.[6]

4.5.2 Nonparanormal data: Gaussian CDF and power transformations

Now we test SING on data from nonparanormal distributions with arbitrary Markov structures. Let $D_k : \mathbb{R} \to \mathbb{R}$ be a monotone and differentiable transformation for $k = 1, \ldots, d$. Following [131], we say that $\mathbf{X} = (X_1, X_2, \ldots, X_d)$ has a nonparanormal distribution with measure $\boldsymbol{\nu}_\pi$ and density π when $\mathbf{Z} = D(\mathbf{X}) \coloneqq (D_1(X_1), D_2(X_2), \ldots, D_d(X_d))$ follows a multivariate Gaussian distribution with measure $\boldsymbol{\nu}_\rho$ and density $\rho = \mathcal{N}(\mathbf{0}, \Sigma)$. We refer to this Gaussian as the "base" distribution in the following subsections. The transformation D acts component-wise on each variable, and is referred to as a *diagonal transport map* that pushes forward π to ρ (or equivalently D pulls back ρ to π). One unique property of diagonal maps is that the pullback measure $D^\sharp \boldsymbol{\nu}_\rho$ through a diagonal map has the same Markov structure as $\boldsymbol{\nu}_\rho$. This is formalized in the following proposition, which we prove in Appendix C.

[6] A tutorial on this example is provided online by Transport Maps Team [216].

Proposition 9. *Let $\boldsymbol{\nu}_\rho$ be a measure with strictly positive density ρ that is Markov with respect to $\mathcal{G}$, and let D be a differentiable diagonal transport map. Then, the pullback measure $D^\sharp\boldsymbol{\nu}_\rho$ is also Markov with respect to $\mathcal{G}$.*

If the pullback density $\pi = D^\sharp\rho$ satisfies the hypotheses in Theorem 4.2.1, we can also observe this property of diagonal transformations in the conditional independence score of π. The following proposition shows that the entries of the score matrix are related to the mixed partial derivatives of the log-density of ρ and the derivatives of the diagonal map components.

Proposition 10. *Let ρ be a strictly positive continuously differentiable density and let D and its inverse be differentiable diagonal transport maps. Then, the conditional independence score of the pullback density $D^\sharp\rho$ is:*

$$\Omega_{ij} = \int \left|\partial_i\partial_j \log \rho(\boldsymbol{z})\, \partial_i D_i^{-1}(z_i)\partial_j D_j^{-1}(z_j)\right|^2 \rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z}. \tag{4.31}$$

This result is proved in Appendix C.

Remark (Gaussian case)**.** *Suppose that ρ is a multivariate Gaussian density with non-singular covariance $\Sigma \in \mathbb{R}^{d\times d}$. Then the conditional independence score of $D^\sharp\rho$* [7] *has the form:*

$$\Omega_{ij} = (\Sigma^{-1})_{ij}^2 \int \left|\partial_i D_i^{-1}(z_i)\partial_j D_j^{-1}(z_j)\right|^2 \rho(\boldsymbol{z})\mathrm{d}\boldsymbol{z}.$$

Thus, if the transformation D is strictly increasing so that $D_i(z_i) \neq 0$ for all i, then the support of Ω is identical to the support of the inverse covariance matrix, i.e., $\Omega_{ij} = 0$ if and only if $(\Sigma^{-1})_{ij} = 0$. Furthermore, each nonzero entry of the conditional independence score Ω_{ij} is proportional to $(\Sigma^{-1})_{ij}^2$.

From the result in Proposition 9 for diagonal maps, the minimal I-maps of $\boldsymbol{\nu}_\pi$ and of $\boldsymbol{\nu}_\rho$ are equivalent. Thus the Markov structure of the target distribution is prescribed by the sparsity of the precision matrix Σ^{-1} of ρ, but the data can have very

[7] The pullback of a multivariate Gaussian density ρ through a diagonal map D is an example of a Gaussian copula. It is well known that Gaussian copulas preserve the Markov properties of ρ, while introducing non-Gaussianity via nonlinear diagonal transformations.

non-Gaussian features (see Figure 4-9b for one example). Within the field of structure learning from non-Gaussian data, this is known as a nonparanormal distribution and is an example of a Gaussian copula. Nonparanormal distributions are common test cases for algorithms that handle non-Gaussianity.

To generate data from a nonparanormal distribution with an arbitrary graph structure, we follow the steps in Liu, Lafferty, and Wasserman [131]. First, a random sparse graph $\mathcal{G} = (V, E)$ is generated. For each node $i \in (1, \ldots, d)$, we associate a pair of random variables $(Y_i^{(1)}, Y_i^{(2)}) \in [0, 1]^2$ to i where

$$Y_1^{(l)}, Y_2^{(l)}, \ldots Y_d^{(l)} \sim \mathcal{U}[0, 1] \tag{4.32}$$

for $l = 1, 2$. Then, each pair of nodes (i, j) is included in the edge set E with probability

$$\mathbb{P}((i, j) \in E) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\|y_i - y_j\|_2^2}{2s}\right), \tag{4.33}$$

where s is a parameter that controls the sparsity of the graph, $y_k \equiv (y_k^{(1)}, y_k^{(2)})$ is a sample of $(Y_k^{(1)}, Y_k^{(2)})$, and $\|\cdot\|_2$ represents the Euclidean norm. In our numerical experiments we set $s = 3$ and limit the maximum degree of the graph, i.e., the number of the edges incident to each node, to be four. A realization of a graph generated according to this procedure is shown in Figure 4-9a. After defining the graph, the entries of the inverse covariance Σ^{-1} are given by:

$$\Sigma_{ij}^{-1} = \begin{cases} 1 & i = j \\ 0.245 & (i, j) \in E \\ 0 & \text{otherwise.} \end{cases} \tag{4.34}$$

We note that with maximum degree four, the value 0.245 ensures that the inverse covariance matrix is positive definite by the Gershgorin circle theorem.

To sample from $\boldsymbol{\nu}_\pi$, we generate i.i.d. samples $\boldsymbol{Z}^l$ from ρ and apply the inverse diagonal transformation to generate i.i.d. samples $\boldsymbol{X}^l = D^{-1}(\boldsymbol{Z}^l)$ from π. In this work we consider two possibilities for the function D as in [131]: a scaled Gaussian

CDF and a power transformation. We now detail how we construct the two functions D_k.

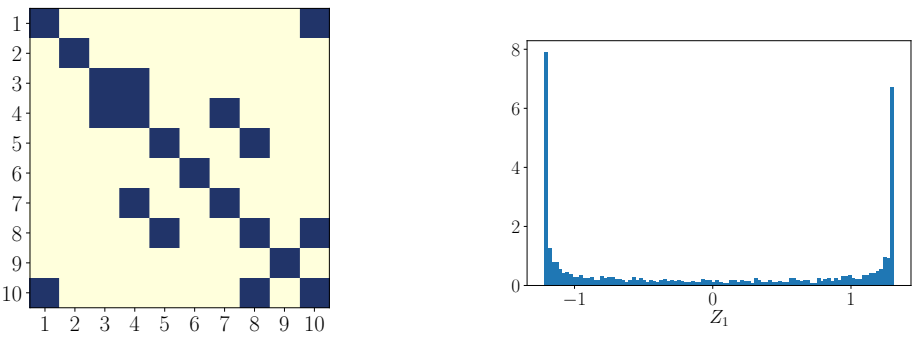
Gaussian CDF transformation. Let $f_0 \colon \mathbb{R} \to [0,1]$ be the univariate Gaussian CDF with mean μ_{f_0} and standard deviation σ_{f_0}, i.e.,

$$f_0(t) = \Phi\left(\frac{t - \mu_{f_0}}{\sigma_{f_0}}\right), \tag{4.35}$$

where Φ is the univariate standard Gaussian CDF. The inverse CDF transformation $F_k = D_k^{-1}$ applied to the kth variable is defined as:

$$F_k(z_k) = \sigma_k \left(\frac{f_0(z_k) - \int f_0(t)\Phi\left(\frac{t-\mu_k}{\sigma_k}\right)dt}{\sqrt{\int \left(f_0(y) - \int f_0(t)\Phi\left(\frac{t-\mu_k}{\sigma_k}\right)dt\right)^2 \Phi\left(\frac{y-\mu_k}{\sigma_k}\right)dy}} \right) + \mu_k. \tag{4.36}$$

In our experiments we apply the same transformation to each marginal by setting $\mu_{f_0} = 0.05$, $\sigma_{f_0} = 0.4$, $\mu_k = 0$ and $\sigma_k = \sqrt{\Sigma_{kk}}$. A representative marginal PDF of π is shown (as a histogram) in Figure 4-9b. Each marginal displays very non-Gaussian behavior as a result of the nonlinearity in the function F_k.



(a) (b)

Figure 4-9: (b) The true minimal I-map of $\boldsymbol{\nu}_\pi$. (b) A histogram of one of the marginals from the nonparanormal dataset. In this case, the data has bimodal, non-Gaussian behavior.

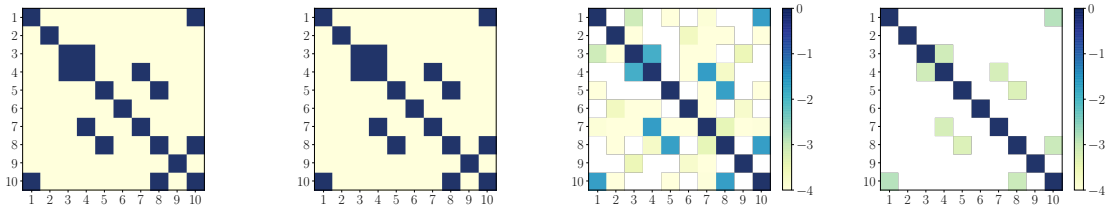
Power transformation. Let $f_0(t) = \text{sign}(t)|t|^a$, for $a > 0$. The inverse power transformation $F_k = D_k^{-1}$ applied to the kth variable is defined as

$$F_k(z_k) = \sigma_k \left(\frac{f_0(z_k - \mu_k)}{\sqrt{\int (f_0(t - \mu_k))^2 \Phi\left(\frac{t-\mu_k}{\sigma_k}\right) dt}} \right) + \mu_k. \tag{4.37}$$

We set μ_k and σ_k as above in the CDF transformation, and set $a = 3$.

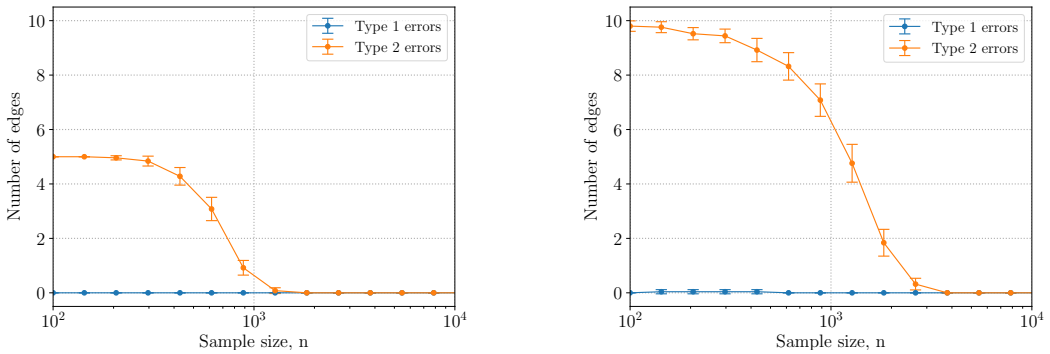
Surprisingly, the conditional independence properties of these nonparanormal distributions are recovered with a linear map. As seen in Figure 4-10a for the Gaussian CDF transformation, we recover the correct graph with $n = 3000$ samples and a polynomial degree $\beta = 1$. Figure 4-10b shows that the correct graph is also returned with $\beta = 2$. While a transport map with higher polynomial degree could be used, a biased approximation to π based on a $\beta = 1$ map in this case is sufficient for learning the graph. We note that the dominant entries of the inverse of the empirical covariance matrix of the data, shown in Figure 4-10c, for this example also reveal the true graph—just as the sparsity of the precision matrix would in the Gaussian case. The next subsection will investigate this connection further. But the inverse of the empirical covariance contains many noisy and non-zero entries as compared to the final score matrix $\overline{\Omega}$, as shown in Figure 4-10d; the latter benefits from the thresholding process in SING and the resulting sparsification of the transport map used to estimate the target density.

Figure 4-11 shows the errors made by the SING algorithm versus the sample size n for both the Gaussian CDF and the power transformation, with $\beta = 1$. For both transformations, the number of type 1 errors (i.e., erroneous edges that are not present in the true graph) is always zero for all n, and the number of type 2 errors (i.e., undetected edges that are present in the true graph) decreases to zero as n increases.



(a) (b) (c) (d)

Figure 4-10: (a) Recovered graph with $\beta = 1$. The correct graph is returned. (b) The recovered graph with $\beta = 2$ is also correct, but Hermite functions of polynomial degree two in the transport map are unnecessary. (c) Entrywise logarithm of the inverse of the empirical covariance matrix of the data. (d) Entrywise logarithm of $\overline{\Omega}$ for $\beta = 1$.



(a) (b)

Figure 4-11: Type 1 and 2 errors for recovering the graph of a nonparanormal distribution versus sample size n for the (a) Gaussian CDF, and (b) power transformation using $\beta = 1$.

4.5.3 Nonparanormal data: cubic transformation

In this section we investigate the effect of biased approximations to the target density when learning the graph. We consider a simplified form of the power transformation applied to a 3-dimensional Gaussian vector $\mathbf{Z} \in \mathbb{R}^3$. Let $\mathbf{Z} \sim \rho = \mathcal{N}(\mathbf{0}, \Sigma_\rho)$ where the precision and the covariance matrix are given by

$$\Sigma_\rho^{-1} = \begin{bmatrix} 1 & 0.2 & 0 \\ 0.2 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{bmatrix}, \quad \Sigma_\rho = 0.92 \begin{bmatrix} 0.96 & -0.2 & 0.04 \\ -0.2 & 1 & -0.2 \\ 0.04 & -0.2 & 0.96 \end{bmatrix} \tag{4.38}$$

From the sparsity of the precision matrix, the random variables satisfy $Z_1 \perp\!\!\!\perp Z_3 | Z_2$. The corresponding graph is a chain connecting nodes Z_1 to Z_2 and Z_2 to Z_3.

We now consider a component-wise monotonic transformation (i.e., a diagonal transport map) given by $F_k(z_k) = z_k^3$ for $k = 1, 2, 3$. The transformed random variable $\mathbf{X} = F(\mathbf{Z}) = (F_1(Z_1), F_2(Z_2), F_3(Z_3))$ is non-Gaussian with density $\pi = F_\sharp \rho$. By the result of Proposition 10 for diagonal transformations, π satisfies the conditional independence property $X_1 \perp\!\!\!\perp X_3 | X_2$, and the Markov structure of π is equivalent to the Markov structure of ρ.

To characterize the target density, suppose we approximate π by the pullback of a standard Gaussian density through a triangular transport map with polynomial degree $\beta = 1$. Using Proposition 7 with $n \to \infty$, the approximate density is a multivariate Gaussian distribution $\mathcal{N}(\mu_\pi, \Sigma_\pi)$ with mean $\mu_\pi = \mathbb{E}_\pi[\mathbf{X}] = \mathbb{E}_\rho[F(\mathbf{Z})] = \mathbf{0}$ and covariance $\Sigma_\pi = \mathbb{E}_\pi[\mathbf{X}\mathbf{X}^T] = \mathbb{E}_\rho[F(\mathbf{Z})F(\mathbf{Z})^T]$. Each entry of the covariance matrix can be expressed analytically in terms of $\sigma_{ij} = (\Sigma_\rho)_{ij}$ as

$$(\Sigma_\pi)_{ij} = 9\sigma_{ii}\sigma_{jj}\sigma_{ij} + 6\sigma_{ij}^3, \tag{4.39}$$

since each entry is in fact a higher-order moment of the multivariate Gaussian distribution $\boldsymbol{\nu}_\rho$. As a result, we can also analytically compute the inverse of Σ_π, which is given by

$$\Sigma_\pi^{-1} = \begin{bmatrix} 5.96 \times 10^{-2} & 6.99 \times 10^{-3} & -5.56 \times 10^{-4} \\ 6.99 \times 10^{-3} & 5.36 \times 10^{-2} & 6.99 \times 10^{-3} \\ -5.56 \times 10^{-4} & 6.99 \times 10^{-3} & 5.96 \times 10^{-2} \end{bmatrix}, \tag{4.40}$$

up to three significant digits. The (1,3) entry of Σ_π^{-1} is not zero, so in principle a Gaussian approximation to π will *not* recover the correct graph. The (1,3) entry is still quite small, however—and, importantly, the relative magnitudes of entries in Σ_π^{-1} are similar to those in Σ_ρ^{-1}. Thus, when using a numerical approximation, the small $(1, 3)$ entry can easily be thresholded and set to zero. And after this thresholding, the correct graph is in fact returned.

Figures 4-12a and 4-12b investigate this phenomenon, by showing the errors made

by SING for sample sizes $n \in [10^2, 10^6]$ using $\beta = 1$ and $\beta = 2$, respectively. The correct graph is returned for a broad range of sample sizes. This is a case where the finite-sample estimators of S and Ω and an under-parameterized map (i.e., a biased approximation to π) interact in a surprisingly beneficial way to correctly learn the graph. We believe the same phenomenon explains the success of the linear $\beta = 1$ transport maps in the previous subsection. When the sample size becomes large enough to resolve the smallest entries in the precision matrix (4.40) with sufficiently high confidence, however, we observe in Figure 4-12 that the SING algorithm with $\beta = 1$ also includes $(1, 3)$ in the estimated edge set of the graph. A similar trend is observed for $\beta = 2$. Note that the class of $\beta = 2$ transport maps is still insufficient to exactly capture π in this example (the exact transport map would be the composition of a linear map with a diagonal map that applies component-wise cubic root transformations). For sufficiently large n, the bias for $\beta = 2$ yields the type 1 errors seen at the right of Figure 4-12b. We also point the reader to a recent article [151] that explains why a Gaussian approximation (i.e., $\beta = 1$ transport maps) can have sparse inverse covariance matrices for these non-paranormal distributions.

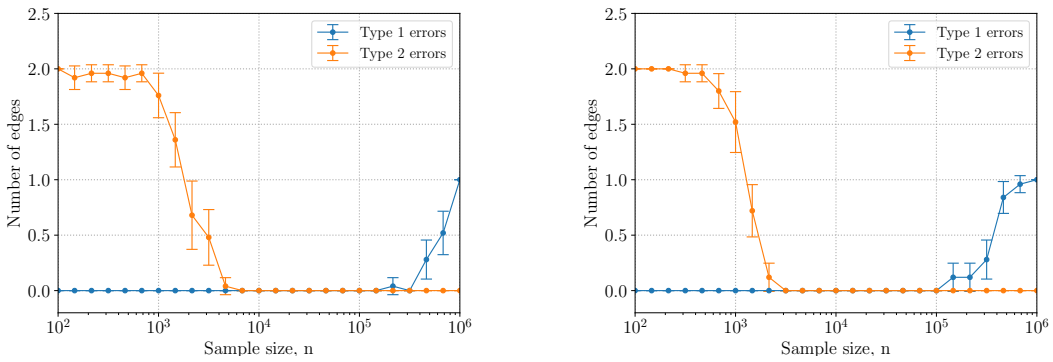


Figure 4-12: Type 1 and 2 errors for recovering the graph of the non-paranormal distribution with the cubic transformation versus sample size n using (a) $\beta = 1$, and (b) $\beta = 2$.

4.5.4 Diagonal transformations of a non-Gaussian base distribution

In this example, we consider the class of distributions $\boldsymbol{\nu}_\pi$ defined as the pullback of a *non-Gaussian* base distribution $\boldsymbol{\nu}_\rho$ through a nonlinear diagonal transport map. When the base density ρ is represented as the pullback of a standard Gaussian density through a triangular transport map that uses polynomials of maximum degree β, we refer to $\boldsymbol{\nu}_\rho$ as a distribution in base class β. The nonparanormal distributions in subsections 4.5.2–4.5.3 are a subset of this class, where ρ is a multivariate Gaussian density (i.e., ρ is in base class $\beta = 1$). For $\beta > 1$, the class considered here is a *generalization* of the nonparanormal family of distributions.

As an example, we consider a distribution $\boldsymbol{\nu}_\rho$ in base class $\beta = 2$. The density ρ is given by the pullback of a standard Gaussian density η through a sparse transport map S of the form

$$S_1(x_1) = ax_1, \qquad S_k(x_1, x_k) = (x_1^2 + b) + ax_k \quad \text{for} \quad k = 2, \ldots, d, \tag{4.41}$$

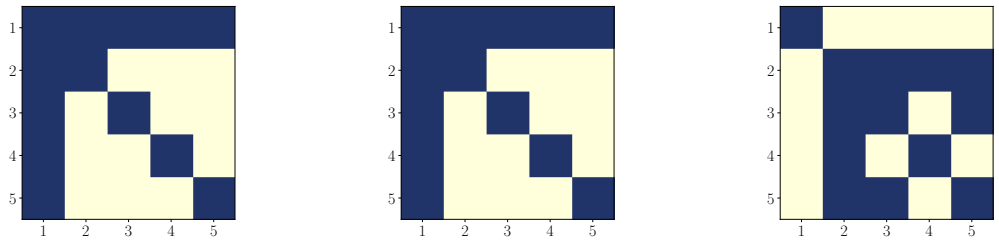
where we set the parameters $a = 1$ and $b = 1$ to adjust the moments of the distribution $S^\sharp \boldsymbol{\nu}_\eta$. The transformed random variable $\mathbf{Z} = S^{-1}(\boldsymbol{Y})$ for $\boldsymbol{Y} \sim \eta$ has the density $S^\sharp \eta$ and its Markov structure is displayed in Figure 4-13 for a $d = 5$ dimensional problem. The star graph associated with the conditional independence structure of ρ is a commonly used graph benchmark for structure learning algorithms [100].

Figure 4-13: Markov structure and sparsity pattern $\mathcal{I}_S$ of the transport map S for the $d = 5$ dimensional pullback density $S^\sharp \eta$.

After defining the base density ρ, we apply the nonlinear inverse CDF transformation in (4.36) to each component of $\mathbf{Z}$ in order to define the random vari-

able $\mathbf{X} = D^{-1}(\mathbf{Z})$. Our target density is that of $\mathbf{X}$: the pullback of a standard Gaussian density η through the composition of S and the diagonal map D, which we denote as $\pi = (S \circ D)^\sharp \eta$. To sample from π, we generate i.i.d. samples $\boldsymbol{Y}^l$ from the standard Gaussian reference density η and apply the composition of the inverse maps $D^{-1} \circ S^{-1} \colon \mathbb{R}^d \to \mathbb{R}^d$ to each sample, thus generating i.i.d. samples $\boldsymbol{X}^l = D^{-1} \circ S^{-1}(\boldsymbol{Y}^l)$.

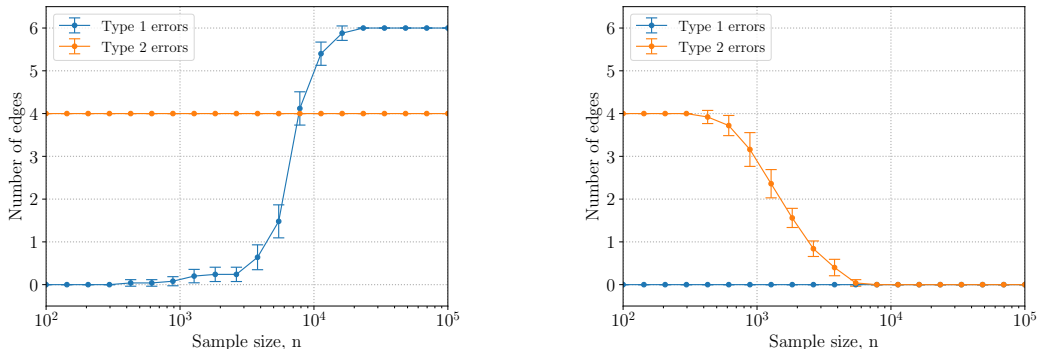
To learn the graph structure of π, we run the SING algorithm with $\beta = 2$ using $n = 10^4$ samples. The true graph and the recovered graph are displayed in Figures 4-14a and 4-14b, respectively. The graph structure is learned correctly. On the other hand, Figure 4-14c shows that using $\beta = 1$ does not recover the true graph of π.



(a) (b) (c)

Figure 4-14: (a) The true graph structure. (b) The recovered graph with $\beta = 2$ is correct. (c) Recovered graph with $\beta = 1$ is incorrect.

To demonstrate the sample sizes needed to learn the true graph in this example, we run the SING algorithm for sample sizes $n \in [10^2, 10^5]$. Figures 4-15a and 4-15b display the average type 1 and type 2 errors in the estimated graphs for $\beta = 1$ and $\beta = 2$, respectively, at each sample size n. For $n \geq 10^4$ samples roughly, the number of type 1 and type 2 errors using $\beta = 2$ remains close to zero and represents successful graph recovery. In contrast, Figure 4-15a shows that using $\beta = 1$, there is no sample size within the tested range where one can recover the exact graph. The conclusion here is analogous to the nonparanormal case of subsection 4.5.2: the SING algorithm can return the correct Markov structure when the β parameter matches the polynomial degree necessary to represent the base distribution and not the target.



(a) (b)

Figure 4-15: Type 1 and 2 errors for recovering the graph of a nonlinear marginal transformation of a *non-Gaussian* base distribution, versus sample size n, using (a) $\beta = 1$, and (b) $\beta = 2$.

4.5.5 Lorenz-96 dynamical system

In our final example, we apply SING to a physics-based dataset generated by a chaotic dynamical system. In particular, we consider the Lorenz-96 model that is commonly used to represent features of the atmosphere (e.g., temperature, vorticity) on a mid-latitude circle of the Earth [134]. The state at d discrete locations thus represents the discretization of a spatially *periodic* domain, described by the state vector $\boldsymbol{X}(t) = (X_1(t), \ldots, X_d(t)) \in \mathbb{R}^d$ at time t. The evolution of this state in time is defined by the set of coupled nonlinear ODEs

$$\frac{dX_j}{dt} = (X_{j+1} + X_{j-2})\, X_{j-1} - X_j + F, \quad j = 1, \ldots, d, \tag{4.42}$$

where $X_{-1} \equiv X_{d-1}$, $X_0 \equiv X_d$, $X_1 \equiv X_{d+1}$. In our experiments we use $d = 15$ and $F = 8$, which leads to chaotic dynamics [178].

To characterize the conditional independence properties of the invariant distribution of the state, we collect data from long-time trajectories of the system. To generate a trajectory, we sample a random initial condition $\boldsymbol{X}(0) \sim \mathcal{N}(\boldsymbol{0}, I_d)$ and use a 4th order Runge-Kutta method with a time step of $\Delta t = 0.01$ for $t \in [0, 1600]$ to approximate the state vector $\boldsymbol{X}(k\Delta t)$ of the ODE in (4.42) at $k \in \mathbb{N}_0$. To reduce correlation between the samples, we sub-sample the trajectory by collecting samples only

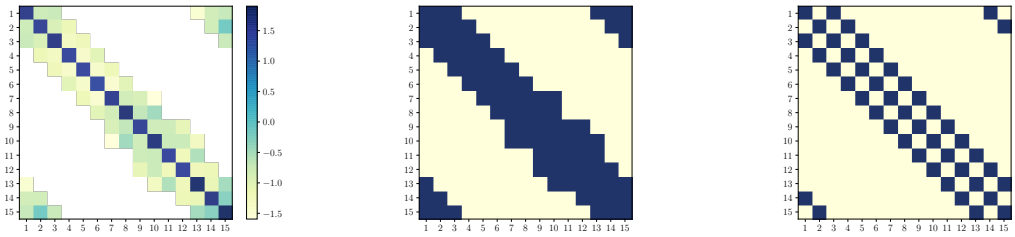
for $k = 40m$ and $m \in \mathbb{N}_0$. Furthermore, to reduce the effect of the initial condition, we discard the first 1000 samples of the trajectory.

To learn the Markov structure of the invariant distribution of the Lorenz-96 system, we run the SING algorithm with $\beta = 2$ using $n = 3000$ samples from this dataset. In this problem, the dynamics at each time step introduce local interactions between each state variable X_j and its neighboring variables on the discretized periodic domain, as seen in the structure of the ODE in (4.42). As a result, the repeated application of the dynamics results in full dependence amongst the variables. However, the invariant distribution of the system can be well approximated by a Markov random field where each variable conditioned on its closest few neighbors in the physical grid is independent of the others. To account for the weak conditional independence between distant states in the grid, we use a threshold for each entry of the conditional independence score given by $\tau_{ij} = \tau_0 + f(n)\widetilde{v}_{ij}/\sqrt{n}$, where $\tau_0 \geq 0$ is a constant offset for all (i, j). This can be seen as a generalization of the threshold proposed in subsection 4.3.4 and applied in the previous numerical examples, where we simply used $\tau_0 = 0$. Here we set $\tau_0 = 0.1$.

The thresholded conditional independence score $\overline{\Omega}$ and the corresponding adjacency matrix of the graph found with SING are shown in Figures 4-16a and 4-16b, respectively. To emphasize the decay of the entries in $\overline{\Omega}$ away from the diagonal, we plot the entrywise logarithm of $\overline{\Omega}$. This figure demonstrates the banded dependence of each variable on neighboring variables separated by at most 3 nodes in either direction, as well as the periodic structure of the graph. On the other hand, the SING algorithm with $\beta = 1$ (i.e., a Gaussian approximation to the target density π) entirely misses the conditional dependence of each variable on its immediate neighbors, as seen in Figure 4-16c.

4.6 Discussion and extensions

This chapter develops a framework for learning the Markov structure of continuous non-Gaussian probability distributions from data. The framework is built on two



(a) (b) (c)

Figure 4-16: (a) Logarithm of nonzero entries in $\overline{\Omega}$. (b) Adjacency matrix of the graph learned for the 15-dimensional Lorenz-96 dataset using $\beta = 2$. (c) The adjacency matrix obtained with $\beta = 1$ incorrectly suggests that each variable X_j is conditionally independent of its immediate neighbors, X_{j-1} and X_{j+1}.

key elements. First, we introduce a computationally tractable score for conditional independence, based on averaged Hessian information from the target log-density. Though this score is immediately useful in its own right, we also show that it bounds the conditional mutual information under appropriate assumptions. Second, we use a transport-based density estimation method, based on parametric approximations of the triangular Knothe–Rosenblatt rearrangement, that explicitly and iteratively exploits sparsity. In particular, the algorithm uses sparsity bounds for triangular maps, which follow directly from the Markov structure; variance-based thresholding for the score estimator; and variable ordering schemes designed to preserve sparsity. Putting these elements together yields the SING algorithm for structure learning. Our analysis shows consistent graph recovery for a single iteration of the algorithm in an asymptotic regime, while numerical results demonstrate the benefits of an iterative algorithm with finite samples.

Indeed, the parameterization of the triangular transport is a useful degree of freedom of our overall framework. Polynomials (or the closely related Hermite functions) are convenient as they include the Gaussian setting as a special case, but future work can certainly explore other options, e.g., the nonlinear separable representations in Spantini, Baptista, and Marzouk [203], or single layers of autoregressive flows. Further analysis of the information gap described above might make it possible to develop transport map estimators with the *minimal* representation that is needed to learn the

Markov structure of the target distribution. Relatedly, it would be valuable to develop information theoretic (*representation-independent*) lower bounds on the number of samples needed to identify the graph in the non-Gaussian setting. To our knowledge, these bounds only exist for Gaussian and discrete Markov random fields [225, 190].

We outline a few additional avenues for future work as follows:

Different data sources. In practical settings, it is of interest to learn the graph from heterogeneous data that may not be independent and identically distributed. Examples include data collected from several related populations that have common structure, and data collected over time. There has been some work to address these problems in the Gaussian setting: Guo, Levina, Michailidis, and Zhu [83] and Danaher, Wang, and Witten [55] propose to combine optimization problems for separate precision matrices with shared ℓ_1-penalties to identify common sparsity, while Zhou, Lafferty, and Wasserman [243] exploits smooth changes in the precision matrix to estimate the evolution of the graph over time.

Latent structure. We also envision extending the SING algorithm to learn the structure of graphical models with latent variables and partial observations. For Gaussian distributions, Chandrasekaran, Parrilo, and Willsky [37] proposes a penalized maximum likelihood approach to identify a precision matrix with sparse and low-rank structure. A similar approach may explore other properties of the conditional independence score matrix Ω, besides sparsity, to reveal hidden variables and multiscale structure in the target density.

High dimensional models. To learn the graph of a d-dimensional distribution, the SING algorithm computes transport maps with d components and stores $\mathcal{O}(d^2)$ entries for the estimated conditional independence score in memory. For high-dimensional datasets, it may not be feasible to jointly estimate the associated graph. Instead, neighborhood selection methods identify local dependencies by independently estimating the neighborhood of each node in $\mathcal{G}$, i.e., $\mathrm{Nb}(k, \mathcal{G})$ for $k = 1, \ldots, d$ such that

$\pi(x_k|\boldsymbol{x}_{-k}) = \pi(x_k|\boldsymbol{x}_{\mathrm{Nb}(k,\mathcal{G})})$. Methods for finding the neighborhood include greedy selection strategies [29] and penalized maximum likelihood estimators [143]. Future work will consider a neighborhood selection version of the SING algorithm that avoids estimating the global transport map for the joint density; this would reduce the run time and memory required to learn the graph.

Chapter 5

Likelihood-free Bayesian inference via couplings

5.1 Introduction

We now turn to the topic of Bayesian inference to learn about unknown model parameters from observed data. Let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ denote model parameters that follow a prior distribution with density $\pi_{\mathbf{X}}$ and let $\mathbf{Y} \in \mathcal{Y} \subset \mathbb{R}^m$ denote observations that follow a likelihood model $\pi_{\mathbf{Y}|\mathbf{X}}$. Given a realization of the data $\boldsymbol{y}^*$, the posterior distribution characterizes the uncertainty in $\mathbf{X}$ given $\boldsymbol{y}^*$. Bayes' theorem provides an expression for the posterior density as

$$\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*) = \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}^*|\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})}{\pi_{\mathbf{Y}}(\boldsymbol{y}^*)}, \tag{5.1}$$

where $\pi_{\mathbf{Y}}(\boldsymbol{y}) = \int \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ is a normalizing constant, that is also referred to as the model evidence or the marginal likelihood. For ease of notation, we will often denote the posterior density as $\pi_{\mathbf{X}|\boldsymbol{y}^*}(\boldsymbol{x}) \coloneqq \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*)$. The goal of Bayesian computation is to provide an explicit expression or to draw samples from the posterior density in (5.1). These samples can be used to estimate the expectation of functions $f(\boldsymbol{x})$ under $\pi_{\mathbf{X}|\mathbf{Y}}$ and to perform downstream tasks such as model selection, model validation, and experimental design.

Many Bayesian inference problems involve a prior $\pi_{\mathbf{X}}$ and/or a likelihood model $\pi_{\mathbf{Y}|\mathbf{X}}$ that are analytically unavailable or computationally prohibitive to evaluate. For instance, the likelihood may involve evaluating a high-dimensional integral over latent variables in a state space model or may only be defined through its data-generating process. In these cases, it not feasible to use classical likelihood-based inference methods such as Markov chain Monte Carlo (MCMC) or importance sampling to generate samples from the posterior density $\pi_{\mathbf{X}|\mathbf{Y}}$ as these rely on likelihood evaluations. As a remedy, approximate Bayesian computation (ABC) provides a suite of techniques to perform inference in generative models using only joint samples of the parameters and observations $(\boldsymbol{X}^i, \boldsymbol{Y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$. While ABC has asymptotic guarantees for correct posterior sampling in the large sample limit, common algorithms are based on rejection sampling and become computationally expensive for models with high-dimensional parameters and data.

In this chapter, we present alternative methods to ABC that are based on deterministic couplings (or transport maps) between random variables. These techniques are used to characterize or sample from the posterior distribution in Bayesian inference problems. In Section 3.2, we provide background on ABC and introduce how transport maps can be used for posterior approximation. Section 5.3 presents an application to characterize the parameter uncertainty of stochastic PDEs used in materials science. Then, we present two novel approaches specifically for posterior sampling. First, Section 5.4 derives a transformation that directly maps prior to posterior samples by *composing* triangular transport maps. We then present an algorithm in Section 5.5 to exploit additional structure in these composed maps. Second, Section 5.6 shows how to estimate and evaluate block-triangular maps which are less sensitive to variable ordering than triangular maps.

5.2 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is one class of Bayesian inference techniques that characterize the posterior without direct likelihood evaluations. While

ABC is sometimes used interchangeably with likelihood-free (or simulated-based) inference, the term was first coined in [17] to perform Bayesian inference for an application in population genetics by imperfectly matching synthetic and observed data samples.

The core idea in classical ABC techniques is to use rejection sampling to find candidate parameters that closely match the observed data $\boldsymbol{y}^*$. To do so, ABC draw pairs of joint samples $(\boldsymbol{X}^i, \boldsymbol{Y}^i)$ where the parameters $\boldsymbol{X}^i \sim \pi_{\mathbf{X}}$ are sampled from the prior and synthetic observations $\boldsymbol{Y}^i \sim \pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\boldsymbol{X}^i)$ are sampled from the likelihood model. We note that more sophisticated ABC algorithms select other proposal distributions for drawing samples $\boldsymbol{X}^i$ (e.g., ABC-MCMC uses approximate posterior samples for $\boldsymbol{X}^i$ that are less likely to be rejected [140]). For each pair of samples, ABC then accepts a parameter $\boldsymbol{X}^i$ if the data $\boldsymbol{Y}^i$ is close to the observed data $\boldsymbol{y}^*$ with respect to some discrepancy measure d (e.g., the Euclidean metric) and tolerance level $\epsilon \geq 0$, i.e., ABC accepts $\boldsymbol{X}^i$ if $d(\boldsymbol{Y}^i, \boldsymbol{y}^*) < \epsilon$.

For $\epsilon = 0$, the ABC algorithm only accepts samples that *exactly* reproduce the data, which can be shown to correspond to sampling perfectly from the posterior distribution. For $\epsilon > 0$, the accepted samples will be drawn from the approximate posterior density[1]

$$\widehat{\pi}_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*) \propto \int_{\mathcal{Y}} \mathbb{1}_{\{d(\boldsymbol{y},\boldsymbol{y}^*)<\epsilon\}} \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}) \mathrm{d}\boldsymbol{y} \pi_{\mathbf{X}}(\boldsymbol{x}). \tag{5.2}$$

In the limit of $\epsilon \to 0$, the approximate posterior in (5.2) matches the posterior density, i.e., $\widehat{\pi}_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*) \to \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*)$. In practice, however, it is unlikely to generate samples that match the data exactly and so it is often necessary to use a strictly positive tolerance ϵ. While smaller values of ϵ result in a lower-bias approximation in (5.2), larger ϵ values accept more parameter samples. As $\epsilon \to 0$, ABC algorithms often require an impractical number of simulations, that are not realistic for models with expensive data-generating procedures. Furthermore, the likelihood of matching

[1]Replacing the indicator by a smooth kernel function produces a kernel density estimator for the posterior density. The estimator weights different synthetic observations based on how well they satisfy the constraint.

the observed data for any ϵ decreases with higher-dimensional data. Hence, fewer samples are accepted and the quality of the non-parametric posterior approximation in (5.2) suffers from the curse-of-dimensionality. We refer the reader to [196] for a comprehensive overview of ABC methods and their properties.

A common approach to reduce the sensitivity of ABC to the tolerance ϵ is to work with summary statistics $\mathbf{Q} = \mathcal{Q}(\mathbf{Y})$ where $\mathcal{Q}\colon \mathbb{R}^m \to \mathbb{R}^r$ is a map from data to low-dimensional statistics, i.e., for $r \ll m$. By comparing summary statistics, ABC produces samples from the conditional distribution $\pi_{\mathbf{X}|\mathbf{Q}}(\cdot|\boldsymbol{q}^*)$ where $\boldsymbol{q}^* = \mathcal{Q}(\boldsymbol{y}^*)$ are the observed statistics. In practice, this will introduce an approximation in the ABC algorithm, unless these statistics are sufficient, meaning that $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y}) = \pi_{\mathbf{X}|\mathbf{Q}}(\cdot|\mathcal{Q}(\boldsymbol{y}))$ for all $\boldsymbol{y}$. While sufficient statistics are often unattainable, a central topic in ABC literature is to construct *informative* statistics. See [66] for an approach that estimates the conditional expectation of the unknown parameters given the observations to define these statistics.

An alternative to ABC that is more sample-efficient with high-dimensional data uses a parametric approximation to the likelihood function. For instance, the Bayesian synthetic likelihood (BSL) method uses n_y observation samples $\boldsymbol{Y}^i \sim \pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\boldsymbol{x})$ to build the Gaussian approximation $\widehat{\pi}_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \widehat{\mathbf{m}}(\boldsymbol{x}), \widehat{\mathbf{C}}(\boldsymbol{x}))$ for each parameter $\boldsymbol{x}$ as $\widehat{\mathbf{m}}(\boldsymbol{x}) \coloneqq \frac{1}{n_y}\sum_{i=1}^{n_y} \boldsymbol{Y}^i$ and $\widehat{\mathbf{C}}(\boldsymbol{x}) \coloneqq \frac{1}{n_y-1}\sum_{i=1}^{l}(\boldsymbol{Y}^i - \widehat{\mathbf{m}}(\boldsymbol{x}))(\boldsymbol{Y}^i - \widehat{\mathbf{m}}(\boldsymbol{x}))^T$. These approximate likelihoods can then be evaluated at $\boldsymbol{y}^*$ within traditional sampling algorithms, such as MCMC, to approximately sample from the posterior density [170]. While Gaussian approximations are more robust in high-dimensions than the non-parametric likelihood in ABC, the BSL is biased in settings with non-Gaussian likelihoods. Moreover, both ABC and BSL must be repeated for each new realization of the data $\boldsymbol{y}^*$.

We now present a transport-based approach for characterizing conditional distributions. We begin by considering the triangular transport map S that pushes forward the joint distribution $\pi_{\mathbf{Y},\mathbf{X}}$ to a reference distribution with independent components of the same dimensions, i.e., $\eta_{\mathbf{Z}_1,\mathbf{Z}_2} = \eta_{\mathbf{Z}_1}\eta_{\mathbf{Z}_2}$ (e.g., a multivariate standard normal). Given that S is lower triangular, we can choose the variable ordering and partition

the map into the following two blocks

$$S(\boldsymbol{y}, \boldsymbol{x}) = \begin{bmatrix} S^{\mathcal{Y}}(\boldsymbol{y}) \\ S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}) \end{bmatrix}, \tag{5.3}$$

where $S^{\mathcal{Y}} \colon \mathbb{R}^m \to \mathbb{R}^m$ pushes forward the data marginal $\pi_{\mathbf{Y}}$ to the reference marginal $\eta_{\mathbf{Z}_1}$ and $\boldsymbol{\xi} \mapsto S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{\xi})$ is a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ that pushes forward $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ to the other marginal $\eta_{\mathbf{Z}_2}$ for each $\boldsymbol{y} \in \mathbb{R}^m$. Hence, the inverse of the map $\boldsymbol{\xi} \mapsto S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{\xi})$ pulls back $\eta_{\mathbf{Z}_2}$ to the conditional $\pi_{\mathbf{X}|\boldsymbol{y}}$. We denote this inverse map by $S^{\mathcal{X}}(\boldsymbol{y}, \cdot)^{-1}$. Let us remark that the map S in (5.3) can be triangular as in Chapter 3, but also generalized to have a block-triangular structure. Section 5.6 presents an approach to find invertible block-triangular maps for conditional sampling.

Given n i.i.d. samples from the joint distribution $\{(\boldsymbol{X}^i, \boldsymbol{Y}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Y}}$ we can learn the components in block $S^{\mathcal{X}}$ of a triangular map by solving the optimization problems

$$\arg\min_{s} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{2} s(\boldsymbol{Y}^i, \boldsymbol{X}^i_{\leq k})^2 - \log |\partial_{m+k} s(\boldsymbol{Y}^i, \boldsymbol{X}^i_{\leq k})| \right] \tag{5.4}$$

under the monotonicity constraint $\partial_{m+k} s(\boldsymbol{y}, \boldsymbol{x}_{\leq k}) > 0$ for all $k = 1, \ldots, d$. We refer the reader to Chapter 3 for more details on how to represent triangular maps and solve these optimization problems.

Solving the constrained problem in (5.4) for each component produces an estimator $\widehat{S}^{\mathcal{X}}$ that approximates the conditional density $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ for any $\boldsymbol{y}$. Hence, finding a map parameterized by $\boldsymbol{y}$ amortizes the cost of inference for multiple realizations of the data. In particular, evaluating the map at $\boldsymbol{y}^*$ yields the approximation of the posterior density $\widehat{\pi}_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*) \coloneqq \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x})$. Furthermore, the estimated map can be used to sample from the approximate posterior density by generating samples $\boldsymbol{Z}_2^i \sim \eta_{\mathbf{Z}_2}$ and inverting the monotone map $S^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{X}^i) = \boldsymbol{Z}_2^i$ for each $\boldsymbol{X}^i$.

Figure 5-1 displays the non-Gaussian conditional densities captured by the map for different values of a one-dimensional observation y^* from a quadratic likelihood model. The single map $S^{\mathcal{X}}(x, y)$ parameterized by y represents both unimodal and bimodal conditional densities $\pi_{X|Y}$. We remark that all of these conditionals are captured

given only samples from the joint density $\pi_{X,Y}$, which can be more easily obtained than samples from the conditional density $\pi_{X|Y}$. Each joint sample $(X^i, Y^i) \sim \pi_{X,Y}$ can be seen as a sample from the conditional density $\pi_{X|Y^i}$. We leverage the smooth dependence of the conditionals densities $\pi_{X|Y}$ on y when using the map to sample from the conditional density corresponding to an previously unseen realization of Y.

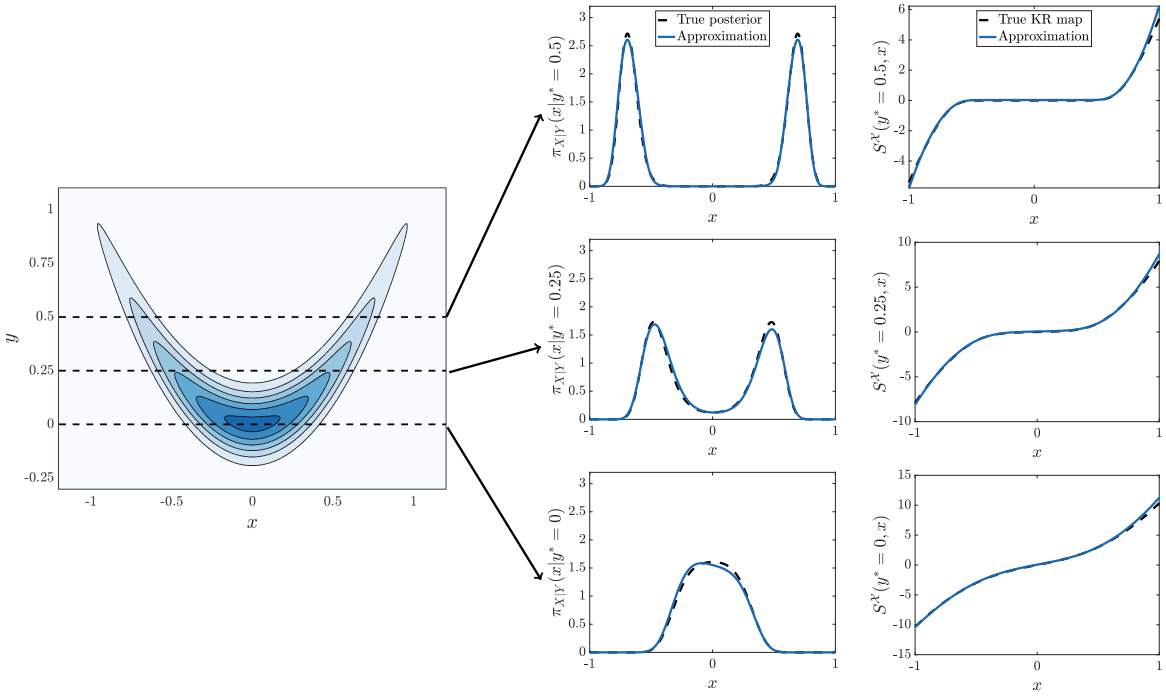


Figure 5-1: Conditional density estimation for the distribution of $X|Y$ with the quadratic likelihood model: $Y^2 = X^2 + \epsilon$ for $X \sim \mathcal{N}(0, 0.25)$ and $\epsilon \sim \mathcal{N}(0, 0.01)$. Contours of the joint density $\pi_{X,Y}$ for the parameter X and the observation Y are on the left. The posterior densities $\pi_{X|Y}(x|y^*)$ for three values of the observation $y^* \in \{0, 0.25, 0.5\}$ and the monotone maps that pull back a standard normal density to each posterior are plotted on the right.

5.3 Application to Stochastic PDEs

In this section, we apply the measure transport technique from Section 5.2 to quantify parameter uncertainties in a computational model for diblock copolymer (Di-BCP) self-assembly. Di-BCPs are polymers consisting of chains of two separate monomers (often labeled as A and B) that are linked to form single copolymers chains. Collections of these Di-BCPS are known as polymer melts. While at higher temperatures

the melts are spatially homogeneously, at lower temperatures the polymers enter ordered states and form periodic nanostructures (i.e., morphologies) such as lines (lamallae), spheres and cylinders, as seen in Figure 5-2. This process is known as microphase separation. The resulting structures yield polymer melts with different physical characteristics. Furthermore, this process could be guided to design complex nanostructures with specific properties for nano-manufacturing applications [102]. In particular, the self-assembly of Di-BCPs is a promising method for fabricating low-cost materials with sub-10 nanometer designs, which is currently very costly with traditional lithography [167].

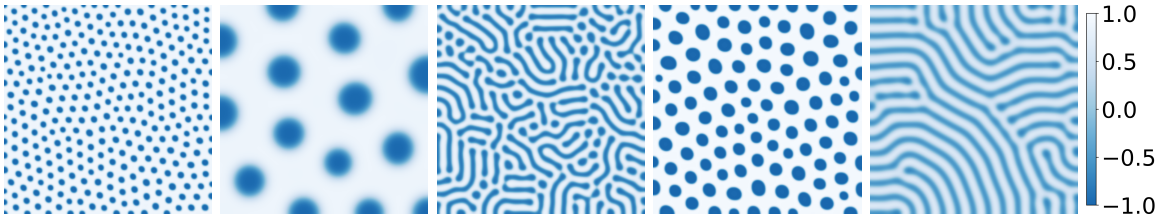


Figure 5-2: Examples of Di-BCP melt microphase seperation simulated using the OK model for various choices of the model parameters (m, ϵ, σ). The two phases are indicated numerically as -1 and $+1$.

Recently, it has been shown that computational models are a valuable tool for predicting melt morphologies and have the potential to be used during the self-assembly process [75]. These models vary in fidelity based on the finest length scale at which they can resolve the arrangements of polymer structures. To rely on these models for design (in particular, the models of coarser-grained fidelity), however, it is important they are calibrated according to observational data in order to make reliable predictions in experimental settings. In this section, we perform Bayesian calibration of model parameters based on observations of the polymer melt's equilibrium states from microscopy or X-ray scattering techniques. This Bayesian calibration quantifies the uncertainty in parameters by accounting for noise in the observations and aleatoric uncertainty in the range of possible equilibrium states.

The remainder of this section is organized as follows. Subsection 5.3.1 introduces the Di-BCP computational model, and subsection 5.3.2 describes the likelihood model for the collected observations. Subsection 5.3.3 presents the posterior approxima-

tions based on measure transport, and subsection 5.3.4 uses these transport maps to measure the information contained in various types of observations about the model parameters.

5.3.1 Di-BCP forward model

In this study we adopt the Ohta-Kawaski (OK) model for the microphase separation of Di-BCP melts. The OK model is a density functional theory that defines the equilibrium Di-BCP states u as minimizers of a free energy functional [157], as an alternative to solving the nonlocal Cahn-Hilliard PDE. Let $u_A, u_B \in [0,1]$ denote the densities for two monomer phases in a domain $\mathcal{D} \subset \mathbb{R}^2$, labelled as A and B, which satisfy $u_A + u_B = 1$. Then, we let $u = u_A - u_B \in [-1,1]$ be the order parameter, which represents the difference of these two phases. The free energy for the order parameter in the OK model $\mathcal{E}^{\mathrm{OK}}(u)$ is given by

$$\mathcal{E}^{\mathrm{OK}}(u) = \int_{\mathcal{D}} \frac{\kappa}{4}(1-u(\boldsymbol{s})^2)^2 + \frac{\epsilon^2}{2}|\nabla u(\boldsymbol{s})|^2 + \frac{\sigma}{2}(u(\boldsymbol{s})-m)(-\Delta^{-1})(u(\boldsymbol{s})-m)\mathrm{d}\boldsymbol{s}, \quad (5.5)$$

where $(\kappa, \epsilon, \sigma) \in \mathbb{R}^3_+$ are scalar model parameters and m denotes the mass average of the order parameter, i.e., $m = \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} u(\boldsymbol{s})\mathrm{d}\boldsymbol{s}$. The free energy functional in (5.5) consists of three terms: a double-well energy that favors pure monomer values taking values of ± 1, an interfacial energy that minimizes the formation of interfaces and large gradients in u, and a non-local energy that characterizes long-range interactions and produces periodic structures in the melt. In this study, we follow the common assumption of taking $\kappa = 1$ by appealing to higher-fidelity SCFT models [40]. The two parameters ϵ and σ are related to the length scale of the interfaces between both phases and the degree of polymerization, respectively. Our goal for this study is to calibrate the model parameters $\mathbf{X} = (m, \epsilon, \sigma)$.

Given a setting for the parameters $\boldsymbol{x}$, the Di-BCP equilibrium state u^* is given by the local minimizer of the OK free energy functional in (5.5) starting from a random initial state z. That is,

$$u^* \in \underset{u \in V_m}{\arg\min}\, \mathcal{E}^{\mathrm{OK}}(u(z)), \quad (5.6)$$

where V_m denotes the state space of functions in $H^1_m = \{u \in H^1(\mathcal{D}), \int_{\mathcal{D}} u(\boldsymbol{s})\mathrm{d}\boldsymbol{s} = m\}$ which satisfy homogeneous Neumann boundary conditions on the boundary of $\mathcal{D}$. Figure 5-2 displays examples of these equilibrium states. In practice, we use the modified Newton method proposed in [31] for minimizing the OK energy. We refer to the mapping from $\boldsymbol{x}$ and z to $u^* = \mathcal{F}(\boldsymbol{x}, z)$ via this optimization procedure as the forward operator.

As a result of the quartic double-well energy term, the OK energy in (5.5) is a non-convex functional. Hence, the minimizer (5.6) is non-unique and depends heavily on the starting point z. To identify multiple equilibrium solutions, it is common to sample z from a uniform white noise random field [172]. In our study, we generate these fields as transformations of an isotropic Gaussian random field with zero mean and covariance operator $(\tau\mathrm{Id}-\Delta)^{-1}$. Here we set $\tau = 200$ so that points in the domain are spatially correlated on a smaller length scale than the features of all equilibrium states. Given a random field with variance σ^2, we apply the pointwise transformation $z \mapsto m + (1-m)\mathrm{erf}(z/(\sqrt{2}\sigma))$ so that the starting point is within the bounds $(-1, 1)$ and has pointwise mean m. We denote the density for the random fields z as $\pi_{\mathbf{Z}}$. The push-forward of $\pi_{\mathbf{Z}}$ through the forward operator $\mathcal{F}(\boldsymbol{x}, \cdot)$ generates a distribution of equilibrium states for each $\boldsymbol{x}$. Moreover, this random input results in a *stochastic* forward operator (i.e., analogous to the Cahn-Hilliard PDE with stochastic inputs) that maps parameters to random realizations of the state.

5.3.2 Likelihood model for observations

The data $\mathbf{Y}$ we use to infer the model parameters $\mathbf{X}$ consist of images of Di-BCPs equilibrium states captured using scanning electron microscopy (SEM) techniques. The SEM sensor extracts a blurred and noise-corrupted version of the equilibrium state on a grid. We model the imaging sensor as the blurring convolution kernel $\mathcal{I}\colon V_m \to V_m$ applied to the polymer melt specimem. In our study, we use the

Gaussian kernel

$$\mathcal{I}(u)(\boldsymbol{s}_2) = \int_{\mathcal{D}} \frac{1}{\sqrt{2}\sigma_{\text{blur}}} \exp\left(-\frac{1}{2\sigma_{\text{blur}}^2}\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|^2\right) u(\boldsymbol{s}_1)\mathrm{d}\boldsymbol{s}_1, \tag{5.7}$$

where σ_{blur} is a blurring strength parameter. We evaluate the effect of three choices for $\sigma_{\text{blur}} \in \{0, 5 \times 10^{-3}, 10^{-2}\}$ on the recovery of the unknown parameters $\mathbf{X}$. In addition to blurring, we assume each sensor is independently corrupted by additive Gaussian noise. This results in an observation model of the form

$$\mathbf{Y}_k = \mathcal{I}(u)(\boldsymbol{s}_k) + \mathbf{N}_k, \tag{5.8}$$

where the observations are extracted on a $m = 256 \times 256$ grid of equidistant sensor locations $\boldsymbol{s}_k$ with i.i.d. Gaussian noise $\mathbf{N}_k \sim \mathcal{N}(0, \sigma_N^2)$ of variance σ_N^2. In our numerical experiments, we compare the effect of the sensor noise variance $\sigma_N^2 = \mathbb{E}[u]/\text{SNR}$ on the recovery for the signal-to-noise ratios $\text{SNR} \in \{\infty, 5.0, 3.0, 2.5\}$.

The additive sensor noise and randomness in the stochastic forward operator produces a conditional distribution $\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})$ for each observation $\boldsymbol{y}$ given the model parameters $\boldsymbol{x}$. Given that multiple values of the random initial point z can generate equilibrium melts u with the same observation, the likelihood for the data $\boldsymbol{y}^*$ must account for all of these values of z. Hence, we have an *integrated likelihood* that marginalizes the inputs z via the conditional expectation

$$\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}^*|\boldsymbol{x}) = \int_{V^z} \pi_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\boldsymbol{y}^*|\boldsymbol{x}, z)\pi_{\mathbf{Z}}(z)\mathrm{d}z, \tag{5.9}$$

where $\pi_{\mathbf{Y}|\mathbf{X},\mathbf{Z}}(\boldsymbol{y}^*|\boldsymbol{x}, z) \propto \exp\left(-\frac{1}{2\sigma_N^2}\sum_{k=1}^m \|\boldsymbol{y}_k^* - \mathcal{I}(\mathcal{F}(\boldsymbol{x}, z))(\boldsymbol{s}_k)\|^2\right)$ is a Gaussian conditional distribution defined by the additive sensor noise. The integration of the infinite-dimensional random input z makes the integrated likelihood in (5.9) intractable to evaluate in practice. This makes inference with the Di-BCP likelihood model a candidate for likelihood-free inference methods.

While in principle we can use the observations in (5.8) to infer the parameters $\mathbf{X}$, in this study we consider low-dimensional summary statistics $\mathbf{Q} = \mathcal{Q}(\mathbf{Y}) \in \mathbb{R}^r$ of the

images $\mathbf{Y}$ for inference. In particular, we hypothesize that some information in the image (e.g., the specific orientation of the material grains) is not relevant to predict the parameters and hence compressing the data should not greatly affect the posterior density. Furthermore, working with statistics has the following two advantages: (1) it can improve the tractability of LFI methods that depend on the dimension of data, and (2) it can provide insight into what information contained in the data is relevant for inference. The two classes of summary statistics we consider in our study are functionals from the OK free energy in (5.5) that are summarized in Table 5.1, and Fourier-based observables.

Table 5.1: Summary statistics for inference of model parameters that are derived from the OK energy functional. In practice, we apply these functionals to images $\boldsymbol{y}$, the noisy and blurred realizations of u.

Summary statistic	**Functional**
Empirical spatial average	$\mathcal{M}(u) = \int_{\mathcal{D}} u \, \mathrm{d}\boldsymbol{s}$
Double well	$\mathcal{Q}_1(u) = \int_{\mathcal{D}} (1-u^2)^2/4 \, \mathrm{d}\boldsymbol{s}$
Interfacial	$\mathcal{Q}_2(u) = \int_{\mathcal{D}} \lvert\nabla u\rvert^2 \, \mathrm{d}\boldsymbol{s}$
Non-local	$\mathcal{Q}_3(u) = \int_{\mathcal{D}} (u-\hat{m}(u))(-\Delta)^{-1}(u-\hat{m}(u)) \, \mathrm{d}\boldsymbol{s}$
Total-variation	$\mathcal{Q}_4(u) = \dfrac{2\int_{\mathcal{D}} (\nabla u \cdot \nabla u)^{1/2} \, \mathrm{d}\boldsymbol{s}}{1-\lvert\mathcal{M}(u)\rvert}$

To summarize periodic patterns in the equilibrium melts we consider Fourier-based observables based on the azimuthal-averaged power spectrum of the images. We hypothesize that this averaging preserves most information about the parameters given that the specific orientation of melt patterns is not unique. To obtain the averaged power spectrum from an image $\boldsymbol{y}^*$ of dimensions $N \times M$, we first apply the discrete Fourier transform to compute the coefficient matrix $\widehat{\boldsymbol{y}} \in \mathbb{C}^{N\times M}$ with the (n, m) entry

$$\widehat{\boldsymbol{y}}(n,m) = \sum_{k=0}^{N-1} \sum_{l=0}^{M-1} \boldsymbol{y}^*(n,m) \exp\left(-i\frac{2\pi nk}{N}\right) \exp\left(-i\frac{2\pi ml}{M}\right), \tag{5.10}$$

where i denotes the imaginary unit satisfying $i^2 = -1$. The power spectrum $\mathbf{P}_s \in \mathbb{R}^{N\times M}$ is given by the modulus of the complex coefficient matrix, i.e., $\mathbf{P}_s = |\widehat{\boldsymbol{y}}|$. The azimuthal-averaged power spectrum $\mathbf{P}_a$ is computed by changing $\mathbf{P}_s$ to radial

coordinates (r, θ) and averaging over $N_\theta(r_k)$ orientations θ for each radial frequency $r_k \in (0, r_{\max}]$. That is,

$$\mathbf{P}_a(r_k) = \frac{1}{N_\theta(r_k)} \sum_{i=1}^{N_\theta(r_k)} \mathbf{P}_s(r_k, \theta_i). \tag{5.11}$$

In our numerical experiments, we use images of dimensions $N, M = 101$ and consider radial frequencies up to $r_{\max} = 40$.

5.3.3 Posterior computation

In this study, we consider a prior distribution that is uniform over a set of admissible parameters $\mathcal{A} \subseteq \mathbb{R}^3_+$. That is, $\pi_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{|\mathcal{A}|}\mathbb{1}_{\{\boldsymbol{x}\in\mathcal{A}\}}$. The boundary of the admissible set $\partial\mathcal{A}$ is defined by describing constraints on each parameter that exclude trivial equilibrium states and regions of non-identifiable parameters. First, we consider mass averages $m \in [0, \sqrt{1/3}]$ that do not produce equilibrium states without phase separation, i.e., disordered states where the two monomers are perfectly mixed. Second, the parameters (ϵ, σ) are selected as a function of m so that the phase transitions are not infinitesimally small and the length scales of the morphological structures are smaller than the domain size. Figure 5-3 plots the resulting admissible set along with $10,000$ i.i.d. samples drawn from $\pi_{\mathbf{X}}$.

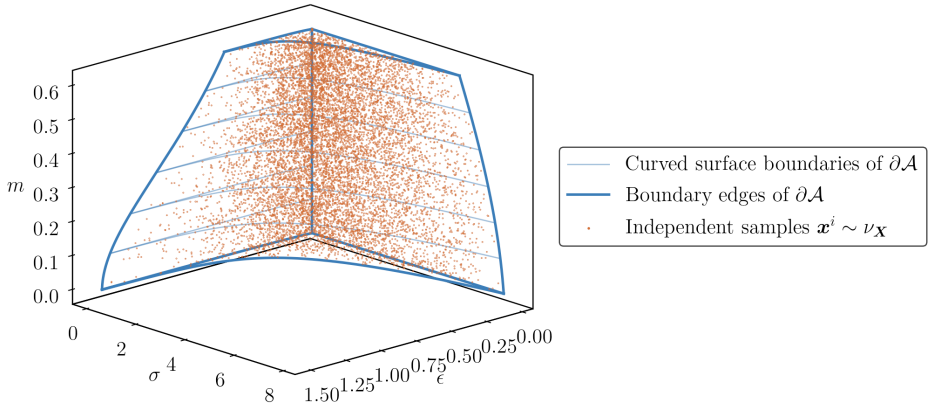


Figure 5-3: Admissible set and samples from the prior density for the model parameters $\pi_{\mathbf{X}}$

Most importantly, the conditional constraints on each parameter yield a hierar-

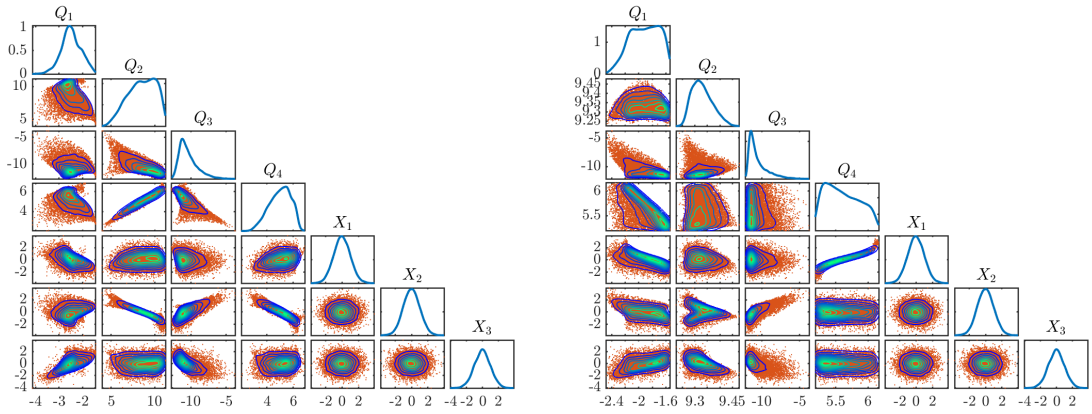
chical prior that factorizes in the form

$$\pi_{\mathbf{X}}(m, \epsilon, \sigma) = \pi_M(m)\pi_{E|M}(\epsilon|m)\pi_{\Sigma|M,E}(\sigma|m, \epsilon). \tag{5.12}$$

Each uniform conditional density in (5.12) can be expressed as the pushforward of an invertible transformation through a univariate Gaussian reference density. This enables us to re-parameterize the model parameters using standard Gaussian random variables without any loss of information. Furthermore, this re-parametrized prior is fully supported on $\mathbb{R}^3$. In comparison, the original parameters result in a prior density, and hence a posterior density, that are constrained to a compact set contained in $\mathcal{A}$, which violates the assumption in Chapters 2 and 3 for using triangular maps. For the remainder of this section, we denote the transformed parameters as $\mathbf{X} \sim \pi_{\mathbf{X}}$ and present the results in this transformed space.

To apply the measure transport technique in Section 5.2, we collect $n = 50,000$ independent prior samples $\boldsymbol{X}^i \sim \pi_{\mathbf{X}}$ and data $\boldsymbol{Y}^i \sim \pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\boldsymbol{X}^i)$ generated from the forward and likelihood models in Sections 5.3.1 and 5.3.2, respectively. For each data sample $\boldsymbol{Y}^i$, we compute the energy observables $\mathbf{Q}^i = (Q_1^i, Q_2^i, Q_3^i, Q_4^i)$ and Fourier-based observables. Figure 5-4 presents joint samples from $\pi_{\mathbf{Q},\mathbf{X}}$ for the likelihood model with two different noise and blur settings. The contours of a kernel density estimate for the one and two-dimensional marginals indicates the non-Gaussianity of the joint densities.

In our experiments we investigate the effect of collecting different observations and applying different levels of noise and blurring to the Di-BCP equilibrium images. Given a parameter $\boldsymbol{x}^* \sim \pi_{\mathbf{X}}$ and observation $\boldsymbol{y}^*$ sampled from the likelihood model, we estimate the transport map $\widehat{S}^{\mathcal{X}}$ using the `ATM` algorithm in Chapter 3 given $n = 30,000$ samples. We then use the resulting map to evaluate the approximate posterior densities $\pi_{\mathbf{X}|\boldsymbol{y}^*}(\boldsymbol{x}) = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x})$. Figure 5-5 displays a marginal of the approximate posterior density for the model parameters $\boldsymbol{x} = (\epsilon, \sigma)$ and $m = 0$ when conditioning on an increasing number of energy functionals (with only additive noise corruption). With more observations, the posterior density concentrates around the



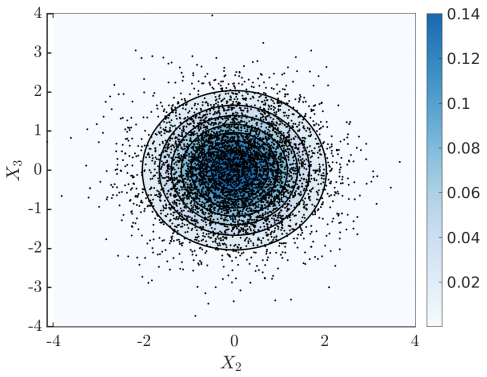
(a) (b)

Figure 5-4: Samples and kernel density estimates for the one and two-dimensional marginals of the joint density $\pi_{\mathbf{Q},\mathbf{X}}$ of observables and parameters given (left) noise-free and blurring-free images and (right) images with additive noise of SNR 2.5 and blurring strength $\sigma_{\text{blur}} = 10^{-2}$.

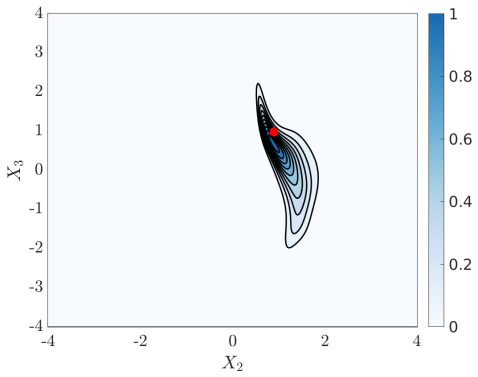
true value of the parameters $\boldsymbol{x}^*$ with lower uncertainty. Similarly, Figure 5-6 displays the $m = 0$ marginal of the approximate posterior after conditioning on two functionals (Q_2, Q_4) when adding either noise or blurring to the images in the likelihood model. We observe that these sources of corruption increase the posterior spread and decrease the posterior accuracy for recovering the true parameters $\boldsymbol{x}^*$ that generated the data.

5.3.4 EIG computation

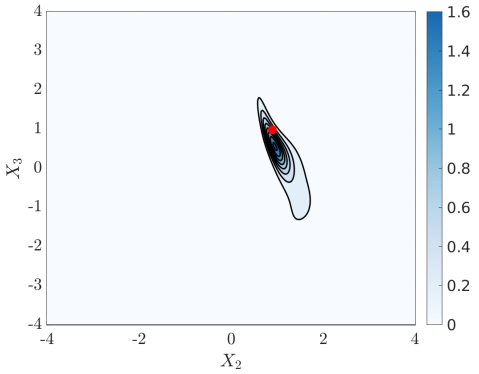
A core advantage of estimating the conditional densities for *all values* of $\boldsymbol{q}$ is that it enables the computation of expectations over the parameters and observable random variables. One of these quantities is the expected information gain (EIG), which represents the average informativeness of an observable $\mathbf{Q}$ for learning about a parameter $\mathbf{X}$. In the field of Bayesian optimal experimental design, the EIG is used to quantify information gain *a priori* to an experiment taking place and collecting a specific realization of the data [129]. Specifically, the EIG $I(\mathbf{X}; \mathbf{Q})$ is defined as the KL divergence from the prior to the posterior in expectation over the random variable



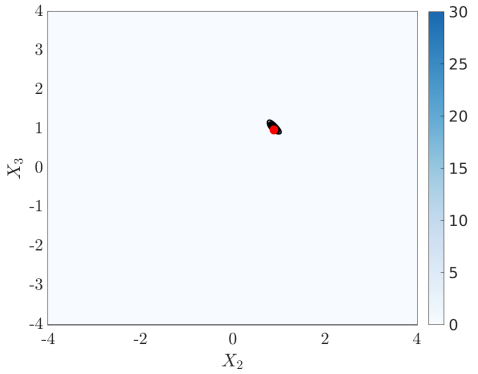
(a) Prior samples



(b) Posterior given $Y = (Q_4)$



(c) Posterior given $Y = (Q_2, Q_4)$



(d) Posterior given $Y = (Q_1, Q_2, Q_3, Q_4)$

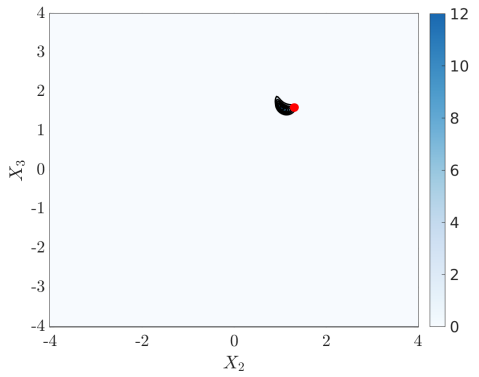
Figure 5-5: Approximate posterior densities for (ϵ, σ) given $m = 0$ when conditioning on an increasing number of energy observables with additive noise of SNR $= 2.5$ and no blurring, i.e., $\sigma_{\text{blur}} = 0$. The posterior concentrates around the true value of the parameters, indicated in red, when conditioning on more observables.

$\mathbf{Q}$. That is,

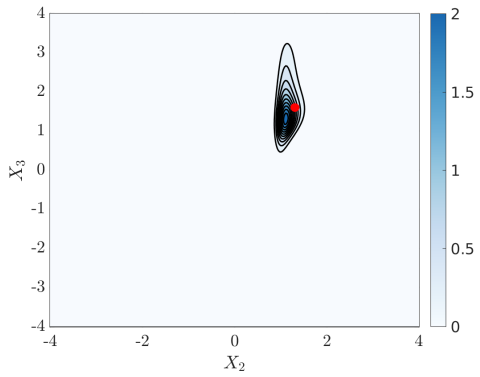
$$I(\mathbf{X}; \mathbf{Q}) = \int_{\mathbb{R}^r} D_{\text{KL}}(\pi_{\mathbf{X}|\boldsymbol{q}} || \pi_{\mathbf{X}}) \pi_{\mathbf{Q}}(\boldsymbol{q}) \mathrm{d}\boldsymbol{q}. \tag{5.13}$$

If the observable $\mathbf{Q}$ is uninformative of $\mathbf{X}$, then the posterior reverts to the prior for all $\boldsymbol{q}$ and the EIG is equal to zero. Otherwise, the EIG is a positive quantity.

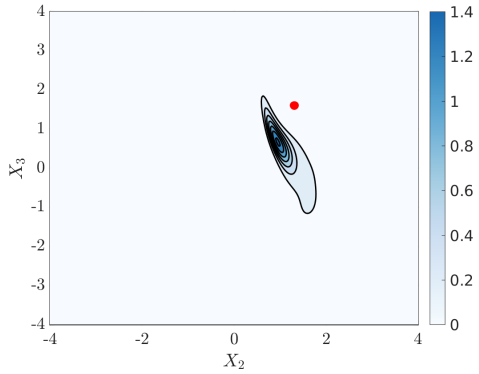
The EIG in (5.13) is equivalently expressed as the expected log-ratio of the posterior over the prior density, i.e., $I(\mathbf{X}; \mathbf{Q}) = \mathbb{E}_{\pi_{\mathbf{X},\mathbf{Q}}}[\log(\pi_{\mathbf{X}|\mathbf{Q}}(\mathbf{X}|\mathbf{Q})/\pi_{\mathbf{X}}(\mathbf{X}))]$. Given an approximate posterior density $\widehat{\pi}_{\mathbf{X}|\mathbf{Q}}$ (for instance, from the measure transport approach in Section 5.2), a simple calculation [69, Appendix A] shows that a lower



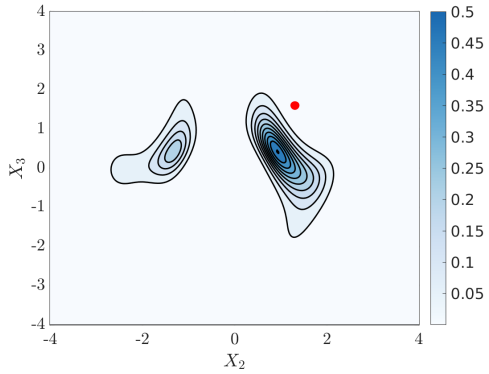
(a) No noise, no blurring



(b) No noise, with blurring



(c) With noise, no blurring



(d) With noise, with blurring

Figure 5-6: Approximate posterior densities $\widehat{\pi}_{\epsilon,\sigma|Q_2,Q_4}$ given $m = 0$ when conditioning on energy observables with additive noise of SNR $= 2.5$ and/or blurring of standard deviation $\sigma_{\text{blur}} = 10^{-2}$. The precision and accuracy of the posterior relative to the true parameter decrease when adding noise or blurring.

bound for the EIG in (5.13) is

$$I(\mathbf{X};\mathbf{Q}) \geq \widehat{I}(\mathbf{X};\mathbf{Q}) := \mathbb{E}_{\pi_{\mathbf{X},\mathbf{Q}}}\left[\log \frac{\widehat{\pi}_{\mathbf{X}|\mathbf{Q}}(\mathbf{X}|\mathbf{Q})}{\pi_{\mathbf{X}}(\mathbf{X})}\right]. \tag{5.14}$$

Moreover, the gap between the EIG and the lower bound is equal to the KL divergence from the approximation to the true posterior in expectation over the observables, i.e., $\mathbb{E}_{\pi_{\mathbf{Q}}}[D_{\text{KL}}(\pi_{\mathbf{X}|\mathbf{Q}}(\cdot|\mathbf{Q})||\widehat{\pi}_{\mathbf{X}|\mathbf{Q}}(\cdot|\mathbf{Q}))]$. Thus, the lower bound for the EIG approaches the true EIG as the approximate density $\widehat{\pi}_{\mathbf{X}|\boldsymbol{q}}$ better approximates the true posterior $\pi_{\mathbf{X}|\boldsymbol{q}}$ for each $\boldsymbol{q}$. Given n i.i.d. samples $\{(\boldsymbol{X}^i, \boldsymbol{Q}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Q}}$ from the joint density

of parameters and observables, an unbiased estimate for the lower bound in (5.14) is

$$\frac{1}{n}\sum_{i=1}^{n}\log\widehat{\pi}_{\mathbf{X}|\mathbf{Q}}(\boldsymbol{X}^i|\boldsymbol{Q}^i) - \mathbb{E}_{\pi_{\mathbf{X}}}[\log\pi_{\mathbf{X}}(\mathbf{X})]. \tag{5.15}$$

The second term in (5.15) is the entropy of the prior density for the parameters, which is often available in closed form. As an example, for a standard Gaussian prior density $\pi_{\mathbf{X}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \mathbf{0}, \mathbf{I}_d)$ with $\boldsymbol{x} \in \mathbb{R}^d$, the entropy is given by $\mathbb{E}_{\pi_{\mathbf{X}}}[-\log\pi_{\mathbf{X}}(\mathbf{X})] = \frac{1}{2}\log(2\pi e d)$. Figure 5-7 presents an example of using EIG to compare the informativeness of two observables for learning about a one-dimensional parameter $X \in \mathbb{R}$. In a

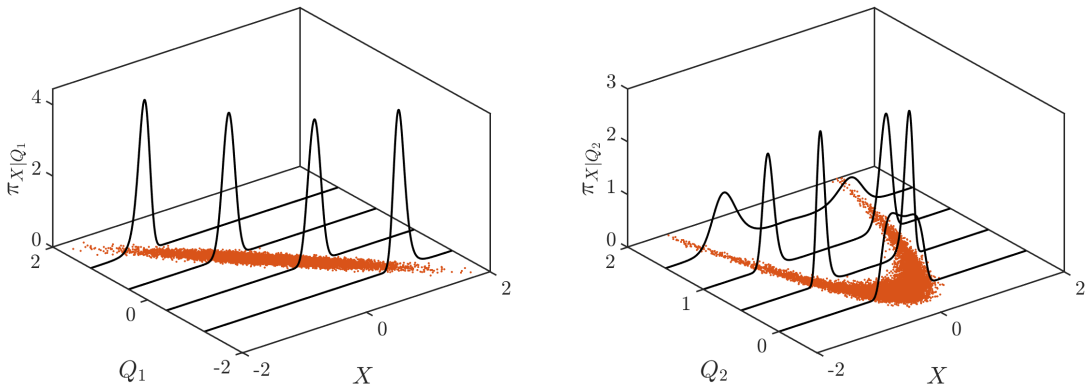


Figure 5-7: Posterior densities for the Gaussian parameter X in Figure 5-1 given two observables $Q_1 = -X + N$ (*left*) and $Q_2 = X^2 + N$ (*right*) for $N \sim \mathcal{N}(0, 0.01)$. The posterior densities for Q_1 are more concentrated as compared to the bi-modal posteriors for Q_2, indicating that Q_1 is more informative of X. Integrating the KL divergence between the prior density for X and the posteriors over the different values of Q_1, Q_2 results in expected information gains (EIG) of $\widehat{I}(X; Q_1) = 1.64 \pm 0.01$ and $\widehat{I}(X; Q_2) = 0.86 \pm 0.01$. EIG provides a quantitative approach to measure and rank the informativeness of different observables.

likelihood–free setting, the prior density $\pi_{\mathbf{X}}$ may also be analytically unavailable. In this case, we can also find an approximation $\widehat{\pi}_{\mathbf{X}}$ for $\pi_{\mathbf{X}}$ and define the EIG estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left[\log\widehat{\pi}_{\mathbf{X}|\mathbf{Q}}(\boldsymbol{X}^i|\boldsymbol{Q}^i) - \log\widehat{\pi}_{\mathbf{X}}(\boldsymbol{X}^i)\right], \tag{5.16}$$

where $\{(\boldsymbol{X}^i, \boldsymbol{Q}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Q}}$. Let us remark that the estimator in (5.16) does not yield a bound on EIG; see [69]. This is problematic when the goal is to maximize the

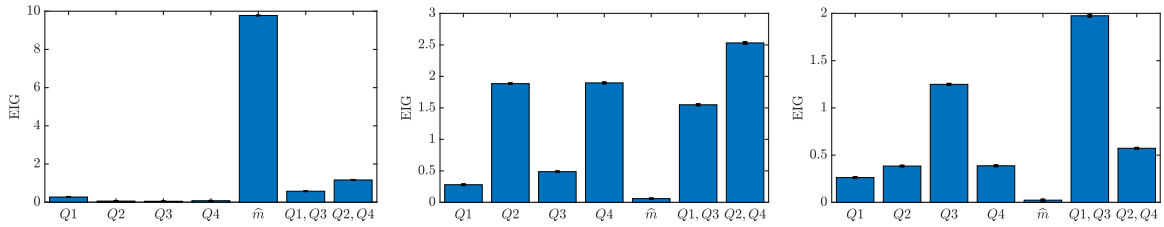
EIG with respect to different designs in the context of optimal experimental design.

An important property of mutual information, and hence expected information gain, is its invariance under marginal transformations of the random variables [46]. If we have two bijective functions f and g, then $I(\mathbf{X};\mathbf{Q}) = I(f(\mathbf{X});g(\mathbf{Q}))$. This property allows us to compute the EIG by finding approximations to the posterior of the transformed random variables so that the resulting EIG estimator has better statistical properties. In our numerical experiments, we transform the original model parameters $\mathbf{X}$ defined on a bounded subset of $\mathbb{R}^3_+$ to the unbounded domain $\mathbb{R}^3$. The transformed random variables have a standard Gaussian prior distribution allowing us to use the EIG estimator in (5.15). A similar transformation can be defined for the observations. For instance, the function $g(\boldsymbol{q}) = \log(\boldsymbol{q})$ will transform non-negative variables with full support on $\mathbb{R}_+$ to be fully supported on $\mathbb{R}$. We did not need these for our case study.

In our numerical experiments we use $n = 20,000$ i.i.d. samples from $\pi_{\mathbf{X},\mathbf{Q}}$ to estimate the expected information gain in (5.14), which are independent of the samples we use to learn the map. To account for outlier samples from numerical anomalies in the discretization and solver for (5.6), we compute the 0.005 and 0.995 quantiles of the log posterior evaluations $\log \widehat{\pi}_{\mathbf{X}|\mathbf{Q}}(\boldsymbol{X}^i|\boldsymbol{Q}^i)$ and discard evaluations that are outside of these quantiles for the EIG estimation.

Figure 5-8 plots the estimated EIG as a function of each marginal parameter for different observables. As naturally expected, the empirical spatial average $\widehat{m}$ is most informative of the mass average m. We also observe that the energy functionals based on gradient information, Q_2 and Q_4, are informative of the interface length ϵ, while the non-local functional Q_3 based on longer length scale information is informative of the overall polymer characteristics (e.g., the number of structural units in each chain), which is given by σ.

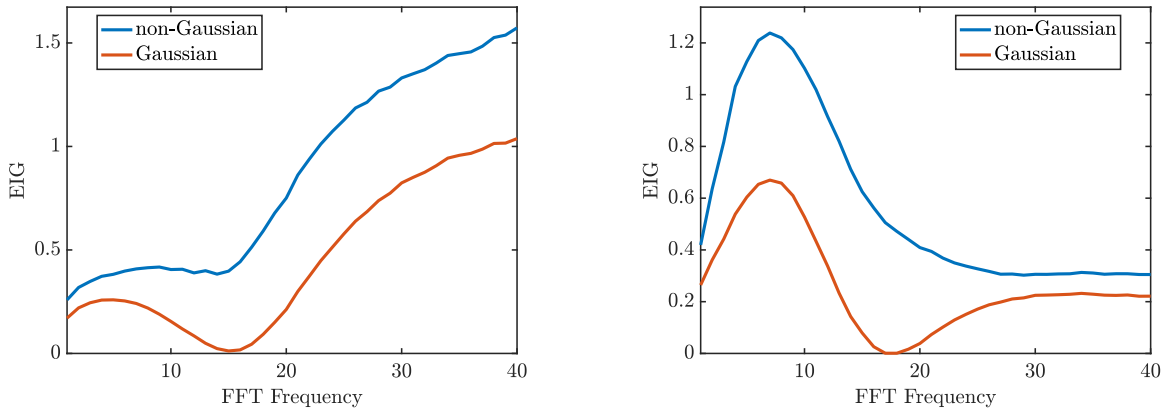
Next, we consider the information gain from the Fourier-based power spectrum of the equilibrium melts. Figure 5-9 plots the EIG for inferring ϵ and σ as a function of different radial frequencies in the power spectrum $\mathbf{P}_a$, without blurring and additive Gaussian noise in the data. The EIG estimates from the non-Gaussian measure



(a) $I(m; \mathbf{Q})$ (b) $I(\epsilon; \mathbf{Q})$ (c) $I(\sigma; \mathbf{Q})$

Figure 5-8: Expected information gain I between each marginal parameter and various energy functionals.

transport approximation of each conditional density $\pi_{\mathbf{X}|\mathbf{Q}}$ are compared to the EIG computed using Gaussian approximations of $\pi_{\mathbf{X}|\mathbf{Q}}$. The measure transport approach provides closer approximations to the true conditionals, and thus a tighter lower bound for EIG in (5.14). In comparison, the Gaussian approximation under-estimates the EIG as seen in Figure 5-9b.



(a) (b)

Figure 5-9: EIG for (a) ϵ and (b) σ when conditioning marginally on each radial frequency in the discrete Fourier transform of the equilibrium melts . The non-Gaussian posterior approximations using the measure transport approach with nonlinear maps are compared to underestimated EIG values using Gaussian approximations (i.e., linear transport maps).

To further compare the Gaussian and non-Gaussian approximations, we visualize the joint density of the non-local parameter σ and the power spectrum $\mathbf{P}_a$ for the radial frequency $r_k = 17$. Figure 5-10 plots i.i.d. samples (left), the approximation using the `ATM` procedure (middle) and a Gaussian approximation (right). We observe that the Gaussian approximation (or equivalently a linear transport map as

in Example 1) does not capture the multi-modal structure of the joint density and results in an approximation with near-zero correlation between the parameter and this observable. Hence, the resulting EIG is near zero at the frequency $r_k = 17$ in Figure 5-9b, in comparison to the EIG estimated using a nonlinear transport map. We conclude this section by emphasizing the importance of using rich and flexible parameterizations (offered by the `ATM` framework in Chapter 3) to characterize non-Gaussian distributions and accurately estimate EIG. This is especially important when extracting insights such as the most informative radial frequencies for learning about each marginal parameter.

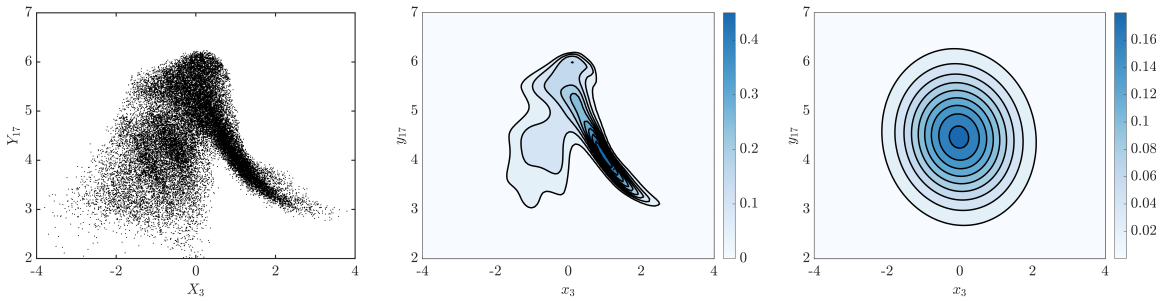


Figure 5-10: The two-dimensional marginal distribution for the parameter σ and the power spectrum for the radial frequency $r_k = 17$. The joint samples ($l$eft) are used to form a non-Gaussian ($m$iddle) and a Gaussian approximation ($r$ight) of the joint density.

5.4 Stochastic map inference algorithm

While the approach in Section 5.2 can be used to sample from or explicitly evaluate the posterior density, we now focus on combining transports to build simpler maps that can be used specifically for posterior sampling. To do so, we construct a prior-to-posterior transformation $T_{\boldsymbol{y}^*}\colon \mathbb{R}^{m+d} \to \mathbb{R}^d$ such that $\mathbf{X}_c = T_{\boldsymbol{y}^*}(\mathbf{Y}, \mathbf{X})$ follows the posterior law with distribution $\pi_{\mathbf{X}|\boldsymbol{y}^*}$ for $(\mathbf{X}, \mathbf{Y}) \sim \pi_{\mathbf{X},\mathbf{Y}}$.

Let S be a triangular transport map of the form in (5.3) that pushes forward the joint density $\pi_{\mathbf{Y},\mathbf{X}}$ to a reference density $\eta_{\mathbf{Z}_1,\mathbf{Z}_2}$ (e.g., a standard normal) of dimension $d+m$. Using the property that the map $\boldsymbol{\xi} \mapsto S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{\xi})$ pushes forward $\pi_{\mathbf{X}|\boldsymbol{y}}$ to $\eta_{\mathbf{Z}_2}$ for each $\boldsymbol{y}$, we have that $S^{\mathcal{X}}(\boldsymbol{Y}^i, \boldsymbol{X}^i) \sim \eta_{\mathbf{Z}_2}$ for each joint sample $(\boldsymbol{X}^i, \boldsymbol{Y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$. Thus,

we can define a composed transformation that maps samples from $\pi_{\mathbf{Y},\mathbf{X}}$ to $\pi_{\mathbf{X}|\boldsymbol{y}^*}$ as

$$T_{\boldsymbol{y}^*}(\boldsymbol{y},\boldsymbol{x}) \coloneqq S^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1} \circ S^{\mathcal{X}}(\boldsymbol{y},\boldsymbol{x}),$$

where $S^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1}$ denotes the inverse of the mapping $\boldsymbol{\xi} \mapsto S^{\mathcal{X}}(\boldsymbol{y}^*,\boldsymbol{\xi})$ that pulls back $\eta_{\mathbf{Z}_2}$ samples to $\pi_{\mathbf{X}|\boldsymbol{y}^*}$ samples. Figure 5-11 depicts this composition as an alternative to sampling using the inverse of $S^{\mathcal{X}}$ alone (i.e., the right hand side of the plot), which we refer to as a single map approach.

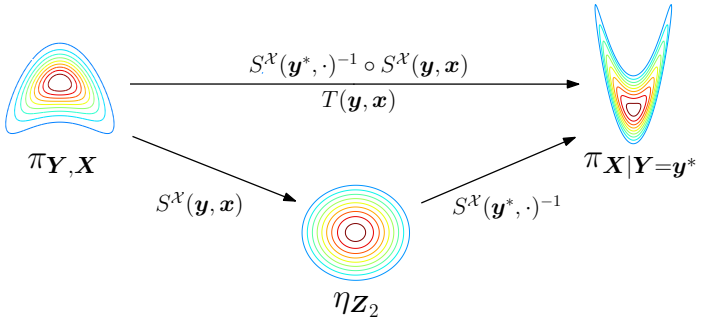


Figure 5-11: Composed maps from the joint distribution $\pi_{\mathbf{Y},\mathbf{X}}$ to the posterior $\pi_{\mathbf{X}|\boldsymbol{y}^*}$

In practice, we build an estimator $\widehat{T}_{\boldsymbol{y}^*}$ for $T_{\boldsymbol{y}^*}$ using n i.i.d. samples $\{(\boldsymbol{Y}^i,\boldsymbol{X}^i)\}_{i=1}^n$ from the joint density $\pi_{\mathbf{Y},\mathbf{X}}$. We estimate $S^{\mathcal{X}}$ by solving the optimization problem (5.4) in Section 5.2, and compose the resulting maps to produce the transformation

$$\widehat{T}_{\boldsymbol{y}^*}(\boldsymbol{y},\boldsymbol{x}) \coloneqq \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{y},\boldsymbol{x}). \tag{5.17}$$

We then generate approximate posterior samples by evaluating $\widehat{T}_{\boldsymbol{y}^*}$ at the same joint samples[2] that are used to estimate $S^{\mathcal{X}}$. The map in (5.17) can be seen as pushing prior samples $\mathbf{X} \sim \pi_{\mathbf{X}}$ to posterior samples through a transformation parametrized by the random variable $\mathbf{Y}$. Hence, we refer to this Bayesian inference method as the *stochastic map* (SM) algorithm and summarize its steps in Algorithm 5.

If the map $\widehat{S}^{\mathcal{X}}$ is estimated exactly such that $\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)$ pulls back $\pi_{\mathbf{X}|\boldsymbol{y}}$ to $\eta_{\mathbf{Z}_2}$ for

[2]Using the same n samples to estimate $\widehat{T}_{\boldsymbol{y}^*}$ and to generate posterior samples will introduce a bias when sampling from the distribution of $\widehat{T}_{\boldsymbol{y}^*}(\mathbf{X},\mathbf{Y})$. Alternatively, we can evaluate $\widehat{T}_{\boldsymbol{y}^*}$ at i.i.d. samples from $\pi_{\mathbf{Y},\mathbf{X}}$ that were not used to estimate $S^{\mathcal{X}}$. In settings with a finite set of joint samples, however, we observe that it is often more advantageous to construct low variance estimators for $S^{\mathcal{X}}$ using all available samples.

all $\boldsymbol{y}$, then using the composed map or the single map will perform identically for posterior sampling. These two approaches will differ, however, given a map $\widehat{S}^{\mathcal{X}}$ such that $\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x}) \neq \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})$. In this case, $T_{\boldsymbol{y}^*}$ is a often a simpler transformation and is less affected by estimation error than using $S^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1}$ alone. For instance, when the data is uninformative of $\mathbf{X}$, i.e., $\widehat{S}^{\mathcal{X}}$ is a constant function of $\boldsymbol{y}$, we have $\widehat{T}_{\boldsymbol{y}^*}(\boldsymbol{x},\boldsymbol{y}) = \mathrm{Id}(\boldsymbol{x})$. Thus, applying $\widehat{T}_{\boldsymbol{y}^*}$ to samples from $\pi_{\mathbf{X},\mathbf{Y}}$ will return exact prior samples while applying $\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1}$ to samples from $\eta_{\mathbf{Z}_2}$ will only return samples from an approximate prior distribution, that depends on how well we estimate $S^{\mathcal{X}}$.

Algorithm 5: Stochastic map (SM) for conditional sampling

Input : i.i.d. samples $\{(\boldsymbol{X}^i,\boldsymbol{Y}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Y}}$, observation $\boldsymbol{y}^*$

Output: Approximate posterior samples $\{\boldsymbol{X}_c^i\}_{i=1}^n$

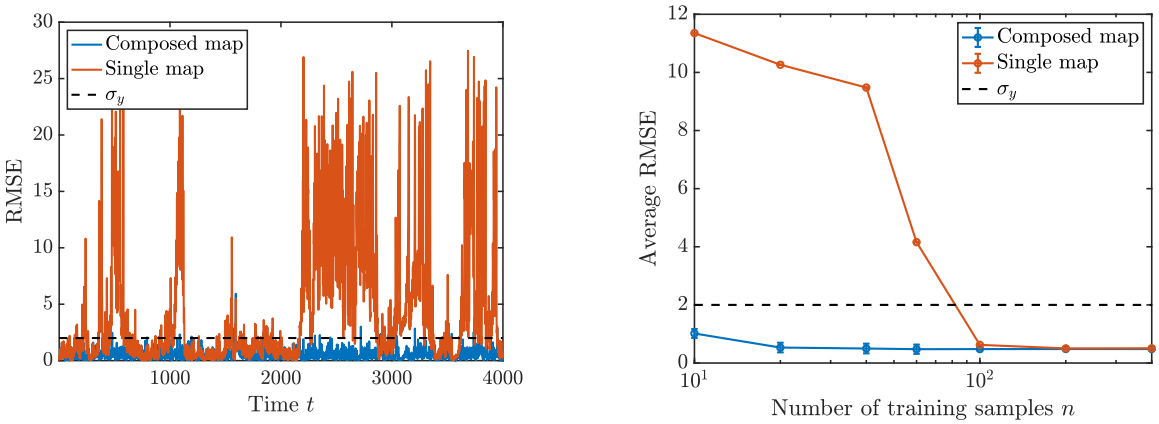
1 Compute block $\widehat{S}^{\mathcal{X}}$ given joint samples by solving (5.4)

2 Evaluate residual samples $\boldsymbol{Z}_2^i = \widehat{S}^{\mathcal{X}}(\boldsymbol{Y}^i,\boldsymbol{X}^i)$ for $i = 1,\ldots,n$

3 Invert map $\boldsymbol{Z}_2^i = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\boldsymbol{X}_c^i)$ to get $\boldsymbol{X}_c^i$ for $i = 1,\ldots,n$

Let us remark that this notion of transforming samples using the composed map rather than re-sampling is well-known in the data assimilation literature. As a simple experiment, we compare the performance of an ensemble Kalman filter (EnKF) for tracking the state of a fully-observed Lorenz-63 model over time using either the composition of linear maps $S^{\mathcal{X}}$ or a single map alone. We refer the reader to Chapter 6 for more details about the Lorenz-63 model and the EnKF. At each inference step, we use n joint samples to approximately sample from the posterior density and use the posterior mean as an estimate for the true underlying state that generated the observations. Figure 5-12 plots the root-mean-squared error (RMSE) for the state as a function of the ensemble size n. We observe that while both posterior sampling methods perform similarly for large ensemble sizes, their ability to track the state is very different for small n. In fact, using linear maps is sufficient to accurately track the state even with $n = 10$ samples if we use them to derive a (composed) prior-to-posterior transformation. In comparison, the single map approximates the posterior by a multivariate Gaussian (i.e., the pullback density of the linear map) at each inference step. As a result, we observe that the single map is unstable for

tracking the state with $n < 100$ samples.



(a) (b)

Figure 5-12: (a) The root-mean-squared error (RMSE) for estimating the state of a Lorenz-63 system over time with $n = 50$ samples for learning the map. (b) The average RMSE over time for increasing sample sizes n. The composed map results in stable performance for tracking the true state with small n.

The remainder of this section is organized as follows. Subsections 5.4.1 and 5.4.2 show the numerical and theoretical advantages of using the composed map over a single map. Lastly, subsection 5.4.3 relates the SM algorithm to a commonly used ABC technique for correcting approximate posterior samples known as regression adjustment.

5.4.1 Variance using composed map

In this section we compare the posterior variance that is estimated using the composed map $\widehat{T}_{\boldsymbol{y}^*}$ to the single map $\widehat{S}^{\mathcal{X}}$. Let the joint distribution be a multivariate Gaussian $\pi_{\mathbf{X},\mathbf{Y}} = \mathcal{N}(\mu, \Sigma)$ with mean and covariance

$$\mu = \begin{bmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{Y}} \end{bmatrix}. \tag{5.18}$$

In this case, the posterior distribution given an observation $\boldsymbol{y}^*$ is also Gaussian, i.e., $\pi_{\mathbf{X}|\boldsymbol{y}^*} = \mathcal{N}(\mu_{\mathbf{X}|\boldsymbol{y}^*}, \Sigma_{\mathbf{X}|\mathbf{Y}})$ with mean and covariance

$$\begin{aligned}\mu_{\mathbf{X}|\boldsymbol{y}^*} &= \mu_{\mathbf{X}} + \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}(\boldsymbol{y}^* - \mu_{\mathbf{Y}}) \\ \Sigma_{\mathbf{X}|\mathbf{Y}} &= \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{YX}}.\end{aligned}$$

To represent Gaussian target distributions as transformations of a standard Gaussian reference $\eta_{\mathbf{Z}_2}$, it is sufficient to consider affine maps $S^{\mathcal{X}}$ of the form

$$S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}) = \mathbf{A}(\boldsymbol{x} - \mathbf{B}\boldsymbol{y} + \mathbf{c}),$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\mathbf{B} \in \mathbb{R}^{d\times m}$ and $\mathbf{c} \in \mathbb{R}^d$ are matrix parameters. Given a finite number of samples from $\pi_{\mathbf{X},\mathbf{Y}}$, we estimate these parameters by solving the optimization problem in (5.4). The corresponding estimator for the map $\widehat{S}^{\mathcal{X}}$ and the composed map $\widehat{T}_{\boldsymbol{y}^*}$ are given by

$$\widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}) = \widehat{\mathbf{A}}(\boldsymbol{x} - \widehat{\mathbf{B}}\boldsymbol{y} + \widehat{\mathbf{c}}), \tag{5.19}$$

$$\widehat{T}_{\boldsymbol{y}^*}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{x} - \widehat{\mathbf{B}}(\boldsymbol{y} - \boldsymbol{y}^*), \tag{5.20}$$

where $\widehat{\mathbf{B}} = \widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-1}$, $\mathbf{c} = \widehat{\mathbf{B}}\widehat{\mu}_{\mathbf{Y}} - \widehat{\mu}_{\mathbf{X}}$, and $\widehat{\mathbf{A}}$ is the inverse Cholesky factor of the conditional covariance $\widehat{\Sigma}_{\mathbf{X}|\mathbf{Y}} \coloneqq \widehat{\Sigma}_{\mathbf{X}} - \widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-1}\widehat{\Sigma}_{\mathbf{XY}}^T$. We use carets to denote the sample average approximations. Hence, as $n \to \infty$, $\widehat{\mathbf{B}}$ converges to the well-known Kalman gain $\Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1}$.

We now compare the estimated covariance of the posterior distribution when using the maps in (5.19) and (5.20) for conditional sampling. Pushing forward the Gaussian reference $\eta_{\mathbf{Z}_2}$ through the inverse of the single map $\boldsymbol{\xi} \mapsto \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{\xi})$ produces a multivariate Gaussian approximate density $\widehat{\pi}_{\mathbf{X}|\mathbf{Y}}$ with covariance $\widehat{\Sigma}^s_{\mathbf{X}|\mathbf{Y}} = \widehat{\mathbf{A}}^{-1}\widehat{\mathbf{A}}^{-T}$. On the other hand, the pushforward of $\pi_{\mathbf{Y},\mathbf{X}}$ through the composed map $\widehat{T}_{\boldsymbol{y}^*}$ produces

the approximate covariance

$$\begin{aligned}\widehat{\Sigma}^c_{\mathbf{X}|\mathbf{Y}} &= \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-T}\widehat{\Sigma}_{\mathbf{XY}}^T - \widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{XY}}^T + \widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}}\widehat{\Sigma}_{\mathbf{Y}}^{-T}\widehat{\Sigma}_{\mathbf{XY}}^T \\ &= \Sigma_{\mathbf{X}|\mathbf{Y}} + (\widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}}^{1/2} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2})(\widehat{\Sigma}_{\mathbf{XY}}\widehat{\Sigma}_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}}^{1/2} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{Y}}^{-1/2})^T \\ &= \Sigma_{\mathbf{X}|\mathbf{Y}} + (\widehat{\mathbf{B}} - \mathbf{B})\Sigma_{\mathbf{Y}}(\widehat{\mathbf{B}} - \mathbf{B})^T.\end{aligned} \tag{5.21}$$

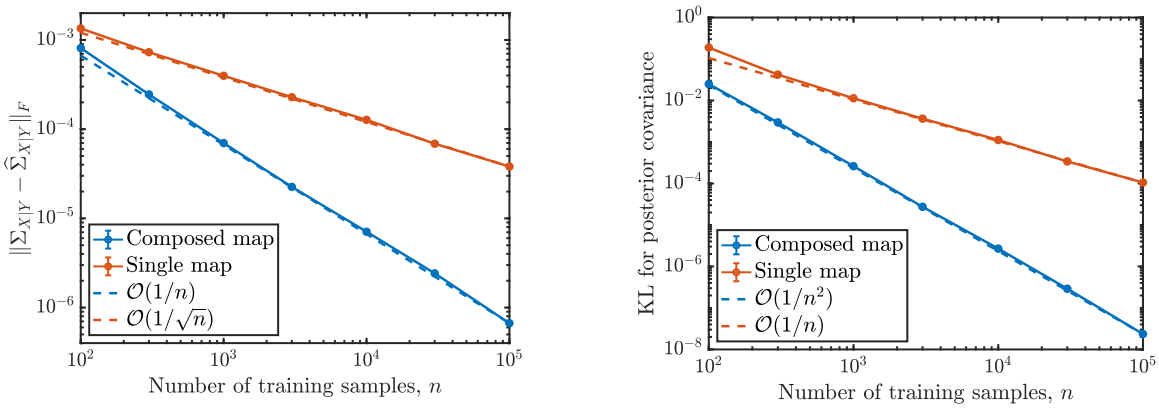
The last equation shows the approximate posterior covariance using $\widehat{T}_{\boldsymbol{y}^*}$ is a perturbation of true posterior covariance. Furthermore, this perturbation only depends on how accurately we estimate the Kalman gain, i.e., $\widehat{\Sigma}^c_{\mathbf{X}|\mathbf{Y}} \to \Sigma_{\mathbf{X}|\mathbf{Y}}$ as $\widehat{\mathbf{B}} \to \mathbf{B}$. In comparison, using the inverse of $\widehat{S}^{\mathcal{X}}$ to sample from $\pi_{\mathbf{X}|\mathbf{Y}}$ requires accurately estimating $\mathbf{A}$, which depends on how well we compute $\mathbf{B}$.

As a numerical example, we consider the posterior of the Bayesian linear regression problem presented in [160]. The prior is the standard Gaussian $\pi_{\mathbf{X}} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and the likelihood is the conditional Gaussian $\pi_{\mathbf{Y}|\mathbf{X}} = \prod_{k=1}^m \mathcal{N}(y^k|\mathbf{u}_k^T\mathbf{x}, \sigma^2)$ where $\mathbf{u}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\sigma = 0.1$. In this experiment we have $d = 6$ parameters and $m = 10$ observations. Using an increasing number of samples n from the joint distribution $\pi_{\mathbf{Y},\mathbf{X}}$, we compute the approximate posterior covariances $\widehat{\Sigma}^s_{\mathbf{X}|\mathbf{Y}}$ and $\widehat{\Sigma}^c_{\mathbf{X}|\mathbf{Y}}$ using the maps in equations (5.19) and (5.20), respectively. Figure 5-13 displays the error between the true and approximate covariances in the Frobenius norm $\|\widehat{\Sigma}_{\mathbf{X}|\mathbf{Y}} - \Sigma_{\mathbf{X}|\mathbf{Y}}\|_F$ and the KL divergence $D_{\mathrm{KL}}(\mathcal{N}(\mu_{\mathbf{X}|\mathbf{Y}}, \Sigma_{\mathbf{X}|\mathbf{Y}})||\mathcal{N}(\mu_{\mathbf{X}|\mathbf{Y}}, \widehat{\Sigma}_{\mathbf{X}|\mathbf{Y}})) = \frac{1}{2}(\log\frac{|\widehat{\Sigma}_{\mathbf{X}|\mathbf{Y}}|}{|\Sigma_{\mathbf{X}|\mathbf{Y}}|} - d + \mathrm{Tr}(\widehat{\Sigma}^{-1}_{\mathbf{X}|\mathbf{Y}}\Sigma_{\mathbf{X}|\mathbf{Y}}))$. We observe that the error using the composed map converges at twice the rate of the single map. A similar difference in convergence rates is observed when using other norms and divergences to measure the error.

The $\mathcal{O}(1/n)$ convergence rate in the Frobenius norm when using the composed map can be shown analytically using the upper bound

$$\left\|\widehat{\Sigma}^c_{\mathbf{X}|\mathbf{Y}} - \Sigma_{\mathbf{X}|\mathbf{Y}}\right\|_F = \left\|(\widehat{\mathbf{B}} - \mathbf{B})\Sigma_{\mathbf{Y}}(\widehat{\mathbf{B}} - \mathbf{B})^T\right\|_F \leq \left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_F^2 \|\Sigma_{\mathbf{Y}}\|_F.$$

Given that $\widehat{\mathbf{B}}$ is a maximum likelihood estimator for $\mathbf{B}$, $\sqrt{n}(\widehat{\mathbf{B}} - \mathbf{B})$ converges asymptotically to a Gaussian random variable with distribution $\mathcal{N}(\mathbf{0}, I^{-1})$, where I de-



(a) (b)

Figure 5-13: The (a) Frobenius norm and (b) KL divergence for the estimated posterior covariance using the single and composed map given n joint samples from $\pi_{\mathbf{X},\mathbf{Y}}$

notes the Fisher information matrix for $\mathbf{B}$. Applying Proposition 1 in [93] we have $\|\widehat{\mathbf{B}} - \mathbf{B}\|_F^2 \leq \frac{1}{n}\left(\mathrm{Tr}(I^{-1}) + 4\|I^{-1}\|_F \log(1/\delta)\right)$ with probability at least $1-\delta$ for $\delta > 0$. Thus with high probability, $\|\widehat{\Sigma}^c_{\mathbf{X}|\mathbf{Y}} - \Sigma_{\mathbf{X}|\mathbf{Y}}\|_F \leq C/n$ for a constant C depending on I, $\Sigma_{\mathbf{Y}}$, and δ. In comparison, the error in the posterior covariance from the single map depends on the accuracy of estimating $\mathbf{A}$, which generally converges at the usual $\mathcal{O}(1/\sqrt{n})$ rate for a maximum likelihood estimator.

5.4.2 Bias using composed map

In this section we compare the bias of statistics computed from the composed map $\widehat{T}_{\boldsymbol{y}^*}$ to the single map $\widehat{S}^{\mathcal{X}}$ in an infinite-sample setting. The stochastic map algorithm begins by learning a function $S^{\mathcal{X}}$ that ideally maps samples from $\pi_{\mathbf{X}|\mathbf{Y}}$ to a standard Gaussian reference density $\eta_{\mathbf{Z}_2}$ by solving

$$\widehat{S}^{\mathcal{X}} = \arg\min_{S^{\mathcal{X}}} \int_{\mathcal{Y}} D_{\mathrm{KL}}\left(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y}) || S^{\mathcal{X}}(\boldsymbol{y},\cdot)^\sharp \eta_{\mathbf{Z}_2}\right) \pi_{\mathbf{Y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}. \tag{5.22}$$

If the objective value is zero at the minimizer, the map $\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)$ pushes forward each conditional $\pi_{\mathbf{X}|\boldsymbol{y}}$ to the same reference $\eta_{\mathbf{Z}_2}$ for each $\boldsymbol{y}$. Here we consider $\widehat{S}^{\mathcal{X}}$ in (5.22) to be found by minimizing over a constrained function space such that the minimum objective value is not zero. In this case, $\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)$ pushes forward the conditional den-

sity $\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})$ to the conditional reference density $\pi_{\mathbf{Z}_2|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) \coloneqq \widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)_\sharp \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})$, which is not equal to the standard Gaussian density $\eta_{\mathbf{Z}_2}$ and may depend on $\boldsymbol{y}$.

Given an approximate map $\widehat{S}^{\mathcal{X}}$, we denote the random variable for the output of the composed map by $\mathbf{X}_c \coloneqq \widehat{T}_{\boldsymbol{y}^*}(\mathbf{Y},\mathbf{X})$ for $(\mathbf{Y},\mathbf{X}) \sim \pi_{\mathbf{Y},\mathbf{X}}$. The distribution of $\mathbf{X}_c$ is found by first marginalizing the conditional reference densities over $\mathbf{Y}$ to define the marginal density for $\mathbf{Z}_2 = \widehat{S}^{\mathcal{X}}(\mathbf{Y},\mathbf{X}) \sim \pi_{\mathbf{Z}_2}$, and second by pulling back $\pi_{\mathbf{Z}_2}$ through the map $\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)$. We let π_c denote the density for the random variable $\mathbf{X}_c$. This density is given by

$$\pi_c(\boldsymbol{x}) = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^\sharp \pi_{\mathbf{Z}_2}(\boldsymbol{x}) \tag{5.23}$$

$$= \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^\sharp \int_{\mathcal{Y}} \pi_{\mathbf{Z}_2|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})\pi_{\mathbf{Y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{y} \tag{5.24}$$

$$= \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^\sharp \int_{\mathcal{Y}} \widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)_\sharp \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})\pi_{\mathbf{Y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \tag{5.25}$$

Let us make two observations. First, if the conditional reference densities are all equal to the standard Gaussian reference, i.e., $\pi_{\mathbf{Z}_2|\boldsymbol{y}} = \eta_{\mathbf{Z}_2}$ and hence $\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x})$ for all $\boldsymbol{y}$, then π_c corresponds to the exact posterior density. From equation (5.24) we have

$$\pi_c(\boldsymbol{x}) = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^\sharp \int \eta_{\mathbf{Z}_2}(\boldsymbol{x})\pi_{\mathbf{Y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{y} = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x}) = \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}^*). \tag{5.26}$$

Second, equation (5.25) shows that the density for the output of the composed map is given by a weighted average of the *true* conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}$ by integrating over the observations $\mathbf{Y}$.

To determine the benefit of using the composed map $T_{\boldsymbol{y}^*}$ to sample from the posterior density, we compare π_c to the density arising from pulling back the Gaussian reference $\eta_{\mathbf{Z}_2}$ through the single map $\widehat{S}^{\mathcal{X}}$, i.e., $\pi_s(\boldsymbol{x}) \coloneqq \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x})$. The following proposition shows that π_c is closer in KL divergence to the true posterior density $\pi_{\mathbf{X}|\boldsymbol{y}^*}$ than π_s on average for any realization of the observations. The proof of this result is provided in Appendix D.

Proposition 11. *Let $\pi_{\mathbf{X},\mathbf{Y}}$ be a joint density and $\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)$ be a transport map that*

approximately pushes forward $\pi_{\mathbf{X}|\boldsymbol{y}}$ *to the standard Gaussian reference* $\eta_{\mathbf{Z}_2}$ *for each* $\boldsymbol{y} \in \mathbb{R}^m$. *Then, the KL divergence from the density of the composed map* π_c *to the posterior density* $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)$ *in expectation over the realizations* $\mathbf{Y}^* \sim \pi_{\mathbf{Y}}$ *satisfies*

$$\mathbb{E}[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_c)] \leq \mathbb{E}[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_s)], \tag{5.27}$$

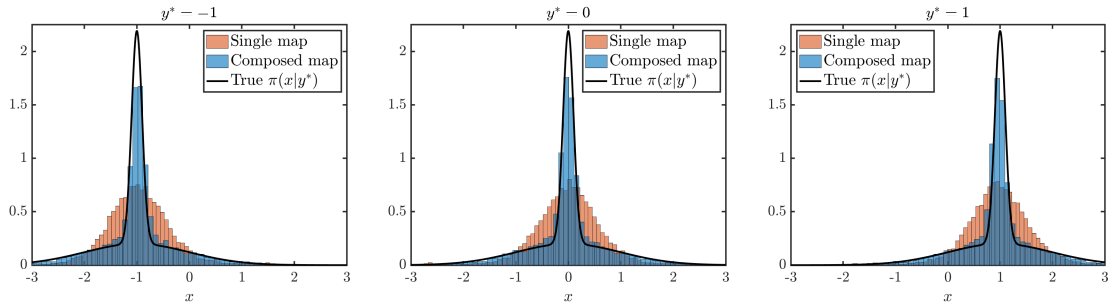
where π_s *denotes the density from the single map. Furthermore, the inequality in* (5.27) *is strict when* $\widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \cdot)$ *does not perfectly pull back* $\eta_{\mathbf{Z}_2}$ *to* $\pi_{\mathbf{X}|\boldsymbol{y}}$.

To numerically demonstrate the reduction in bias coming from the composed map, we consider a one-dimensional example with the Gaussian mixture likelihood model in [197]. This is a common benchmark for ABC algorithms because the multiple length scales in the mixture result in poor nonparametric approximations of the form in (5.2) based on a single threshold parameter. In this example, we consider the uniform prior $\pi_X \sim \mathcal{U}(-10, 10)$ and the likelihood function $\pi_{Y|X} = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(0, 0.01)$. To approximate the posteriors, we use $n = 10^4$ joint samples to learn a monotone map $S^{\mathcal{X}}$ that is parametrized using the degree 5 Hermite polynomials introduced in Chapter 3. Figure 5-14 plots a histogram of posterior samples arising from both the single map and the composed map in comparison to the true posterior density. We observe that the single map using low-degree polynomials does not capture the localized feature in the posterior, while the composed map samples better capture both mixture components by using the same map to push forward joint to posterior samples. Furthermore, this behavior holds for different values of the observation y^*.

5.4.3 Connection to regression adjustment

Regression adjustment is a commonly used method in high-dimensional ABC to correct approximate posterior samples [26]. Furthermore, it can be interpreted on its own as an approximate posterior sampling method. Given samples $\{(\boldsymbol{Y}^i, \boldsymbol{X}^i)\}_{i=1}^n \sim \pi_{\mathbf{Y},\mathbf{X}}$[3], this technique first fits the model $\mathbf{X}^i = \widehat{f}(\mathbf{Y}^i) + \boldsymbol{\epsilon}^i$ for a function $\widehat{f}\colon \mathbb{R}^d \to \mathbb{R}^m$. The

[3]In ABC, these pairs consist of data $\boldsymbol{Y}^i$ and accepted parameter samples $\boldsymbol{X}^i$ that closely match the observation $\boldsymbol{y}^*$.



(a) (b) (c)

Figure 5-14: Posterior densities for observations $y^* \in \{-1, 0, 1\}$ with the mixture of Gaussians likelihood model and a uniform prior for X.

model $\widehat{f}$ is typically found using least-squares regression by assuming the residuals $\boldsymbol{\epsilon}^i$ are independent and Gaussian distributed with variance σ^2. To sample from the predictive distribution for $\boldsymbol{x}$ given an observation $\boldsymbol{y}^*$, regression adjustment uses the training data to generate i.i.d. samples from the residual distribution $\boldsymbol{\epsilon}^i = \mathbf{X}^i - \widehat{f}(\mathbf{Y}^i)$. Let us remark that the distribution for $\boldsymbol{\epsilon}$ may be in general non-Gaussian, regardless of the assumption that is used when fitting the model $\widehat{f}$. Within the statistical estimation literature, this is also known as bootstrap or residual resampling [63]. An approximate sample from the predictive distribution is then given by

$$\mathbf{X}^i_{ra} = \widehat{f}(\boldsymbol{y}^*) + \boldsymbol{\epsilon}^i = \mathbf{X}^i - \widehat{f}(\mathbf{Y}^i) + \widehat{f}(\boldsymbol{y}^*). \tag{5.28}$$

The transformation in (5.28) from the joint distribution of parameters and observations to $\mathbf{X}_{ra}$ has the same form as the prior-to-posterior transformation $T_{\boldsymbol{y}^*}$ in the stochastic map algorithm when the map $S^{\mathcal{X}}$ is an affine function of $\boldsymbol{x}$ with an additive and (possibly) nonlinear dependence on the observation $\boldsymbol{y}$, i.e., $S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}) = -f(\boldsymbol{y}) + \boldsymbol{x}$. Thus, the stochastic map algorithm can be interpreted as a generalization of regression adjustment that is in principle consistent on its own for posterior sampling when there is no bias in the estimator for $T_{\boldsymbol{y}^*}$. Alternatively, the samples from the composed map $\widehat{T}_{\boldsymbol{y}^*}(\boldsymbol{Y}^i, \boldsymbol{X}^i)$ can be used to correct approximate posterior samples in ABC methods, with a more complex model than what is typically used in regression adjustment.

As an alternative to the sampling proposed in regression adjustment, we could

discard the n training samples after fitting $\widehat{f}$ and use the least-squares model to generate independent samples from $\mathcal{N}(\widehat{f}(\boldsymbol{y}^*), \sigma^2)$. This Gaussian approximation to the predictive distribution corresponds to mapping standard normal samples through the inverse map $\boldsymbol{\xi} \to S^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{\xi})$; see Section 5.4.2. As in the stochastic map algorithm, this approach does not account for the distribution of the residual random variables. In addition, sampling from a Gaussian distribution is more sensitive to the distributional assumptions made when learning $\widehat{f}$. Table 5.2 summarizes the steps in regression adjustment and the SM algorithm. The SM algorithm generalizes each step in regression adjustment to sample exactly from the conditional density for $\mathbf{X}|\mathbf{Y}$.

Table 5.2: A comparison of steps in regression adjustment and the stochastic map (SM) algorithm for sampling from the conditional density $\pi_{\mathbf{X}|\boldsymbol{y}^*}$.

Regression adjustment	**SM algorithm**
1. Fit regression model: $\mathbf{X} = \widehat{f}(\mathbf{Y}) + \boldsymbol{\epsilon}$	1. Fit transport map: $\widehat{S}^{\mathcal{X}}(\mathbf{Y}, \mathbf{X}) \sim \eta_{\mathbf{Z}_2}$
2. Evaluate model at $\boldsymbol{y}^*$:	2. Evaluate model at $\boldsymbol{y}^*$:
For Gaussian $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$,	For Gaussian $\eta_{\mathbf{Z}_2} = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$,
then $\mathbf{X}|\boldsymbol{y}^* \sim \mathcal{N}(\widehat{f}(\boldsymbol{y}^*), \sigma^2\mathbf{I}_d)$	then $\mathbf{X}|\boldsymbol{y}^* \sim \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^\sharp \eta_{\mathbf{Z}_2}$
3. Sample residuals: $\boldsymbol{\epsilon}^i = \boldsymbol{X}^i - \widehat{f}(\boldsymbol{Y}^i)$	3. Sample reference: $\boldsymbol{Z}_2^i = \widehat{S}^{\mathcal{X}}(\boldsymbol{Y}^i, \boldsymbol{X}^i)$
4. Sample posterior:	4. Sample posterior:
$\mathbf{X}_{ra}^i = \widehat{f}(\boldsymbol{y}^*) + \boldsymbol{\epsilon}^i$	$\mathbf{X}_c^i = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^{-1}\Big\vert_{\boldsymbol{Z}_2^i}$
$= \widehat{f}(\boldsymbol{y}^*) + \boldsymbol{X}^i - \widehat{f}(\boldsymbol{Y}^i)$	$= \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{Y}^i, \boldsymbol{X}^i)$

5.5 Exploiting structure in composed maps

A natural question to ask is when the composed map $\widehat{T}_{\boldsymbol{y}^*}$ will sample exactly from the posterior distribution. The following proposition shows that this is true even when the map $\widehat{S}^{\mathcal{X}}$ does not pushforward $\pi_{\mathbf{X}|\mathbf{Y}}$ to a standard Gaussian reference $\eta_{\mathbf{Z}_2}$. Instead, $\widehat{S}^{\mathcal{X}}$ only needs to map all conditionals to a distribution that does not depend on $\mathbf{Y}$. The proof of this result is provided in Appendix D.

Proposition 12. *Let $\widehat{S}^{\mathcal{X}}\colon \mathbb{R}^{m+d} \to \mathbb{R}^d$ be a map that satisfies $L_\sharp \eta_{\mathbf{Z}_2} = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \cdot)_\sharp \pi_{\mathbf{X}|\boldsymbol{y}}$ for all $\boldsymbol{y}$ where $L\colon \mathbb{R}^d \to \mathbb{R}^d$ is an arbitrary monotone transformation and $\eta_{\mathbf{Z}_2}$ is the standard Gaussian density on $\mathbb{R}^d$. Then, $\widehat{T}_{\boldsymbol{y}^*}(\mathbf{Y}, \mathbf{X})$ follows the law of the posterior distribution $\pi_{\mathbf{X}|\boldsymbol{y}^*}$.*

As a consequence of Proposition 12, it is sufficient to learn a map $\widehat{S}^{\mathcal{X}}$ that removes the dependence of each conditional distribution $\pi_{\mathbf{X}|\boldsymbol{y}}$ on $\boldsymbol{y}$. This goal can be stated as finding maps $\widehat{S}^{\mathcal{X}}$ such that the reference random variable $\mathbf{Z}_2 \coloneqq \widehat{S}^{\mathcal{X}}(\mathbf{Y}, \mathbf{X})$ is independent of $\mathbf{Y}$. The following example finds such a map in a setting where the conditional dependence is isolated to a single parameter of the distribution.

Example 3 (Location-scale family)**.** *Let the conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ for all $\boldsymbol{y}$ belong to a location-scale family with the same functional form where only the conditional mean depends on $\boldsymbol{y}$. These densities can be expressed as $\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \rho(\boldsymbol{x} - \mu(\boldsymbol{y}))$ in terms of a density $\rho\colon \mathbb{R}^d \to \mathbb{R}_+$ that does not depend on $\boldsymbol{y}$ and the conditional mean $\mu(\boldsymbol{y}) \coloneqq \mathbb{E}[\mathbf{X}|\mathbf{Y} = \boldsymbol{y}]$. In this case, the affine transformation $\widehat{S}^{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} - \mu(\boldsymbol{y})$ is sufficient to remove the dependence on the observations $\mathbf{Y}$. Furthermore, the conditional reference*

$$\pi_{\mathbf{Z}_2|\mathbf{Y}}(\boldsymbol{z}_2|\boldsymbol{y}) = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \cdot)_\sharp \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{z}_2|\boldsymbol{y}) = \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{z}_2 + \mu(\boldsymbol{y})|\boldsymbol{y}) = \rho(\boldsymbol{z}_2),$$

is only a function of $\boldsymbol{z}_2$. Thus, by Proposition 12 the composed map $T_{\boldsymbol{y}^*}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{x} - \mu(\boldsymbol{y}) + \mu(\boldsymbol{y}^*)$ can be used to sample exactly from $\pi_{\mathbf{X}|\boldsymbol{y}^*}$. Let us remark that using the single map alone by pulling back the standard Gaussian density $\eta_{\mathbf{Z}_2}$ through $\widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \cdot)$ will not equal $\pi_{\mathbf{X}|\mathbf{Y}}$ if ρ departs strongly from $\eta_{\mathbf{Z}_2}$. In this case, we require a richer map that characterizes the functional form of ρ to sample from the conditional density $\pi_{\mathbf{X}|\mathbf{Y}}$ correctly using the single map.

Similarly, composing the single map $\widehat{S}^{\mathcal{X}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \mu(\boldsymbol{y}))/\sigma(\boldsymbol{y})$ is sufficient to sample exactly from the conditional densities in the location-scale family where both the conditional mean μ and conditional standard deviation $\sigma(\boldsymbol{y}) \coloneqq \mathbb{V}[\mathbf{X}|\mathbf{Y} = \boldsymbol{y}]$ depend on $\boldsymbol{y}$.

Outside of the location-scale families in Example 3, we consider maps $\widehat{S}^{\mathcal{X}}$ that

are general nonlinear functions of $\boldsymbol{x}$ and $\boldsymbol{y}$ for conditional densities whose shapes and functional forms depend on the observations $\boldsymbol{y}$. To measure the dependence of $\mathbf{Z}_2$ on $\mathbf{Y}$, a popular measure of independence is the mutual information (MI)[4]. The MI $\mathrm{I}(\mathbf{Z}_2;\mathbf{Y})$ is defined as the Kullback-Leibler divergence from the product of marginals $\pi_{\mathbf{Z}_2}$ and $\pi_{\mathbf{Y}}$ to the joint density $\pi_{\mathbf{Z}_2,\mathbf{Y}} = \pi_{\mathbf{Z}_2|\mathbf{Y}}\pi_{\mathbf{Y}}$. That is,

$$\mathrm{I}(\mathbf{Z}_2;\mathbf{Y}) = \int \pi_{\mathbf{Z}_2,\mathbf{Y}}(\boldsymbol{z}_2,\boldsymbol{y}) \log\left(\frac{\pi_{\mathbf{Z}_2,\mathbf{Y}}(\boldsymbol{z}_2,\boldsymbol{y})}{\pi_{\mathbf{Z}_2}(\boldsymbol{z}_2)\pi_{\mathbf{Y}}(\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{z}_2\mathrm{d}\boldsymbol{y}. \tag{5.29}$$

The mutual information satisfies $\mathrm{I}(\mathbf{Z}_2;\mathbf{Y}) = 0$ if and only if $\pi_{\mathbf{Z}_2,\mathbf{Y}}$ factorizes into a product of its marginals, meaning that $\mathbf{Z}_2$ and $\mathbf{Y}$ are independent. Thus, to find a map $\widehat{S}^{\mathcal{X}}$ that removes the dependence of $\mathbf{X}$ on $\mathbf{Y}$, we cast the variational problem

$$\begin{aligned} &\min_{S^{\mathcal{X}}} I(\mathbf{Z}_2;\mathbf{Y}) \\ &\text{s.t. } \mathbf{Z}_2 = S^{\mathcal{X}}(\mathbf{Y},\mathbf{X}) \\ &\qquad S^{\mathcal{X}} \text{ is monotone in } \boldsymbol{x} \text{ for all } \boldsymbol{y}. \end{aligned} \tag{5.30}$$

As compared to the optimization problem for the map in (5.4), the mutual information is a less restrictive objective. For instance, the mutual information $I(\mathbf{Z}_2;\mathbf{Y})$ does not change when adding a constant to the map $S^{\mathcal{X}}$. On the other hand, minimizing the KL divergence will change any additive constants in the map in order to match the means of $\pi_{\mathbf{X}|\mathbf{Y}}$ and $S^{\mathcal{X}}(\boldsymbol{y},\cdot)^\sharp\eta_{\mathbf{Z}_2}$. More generally, the mutual information is invariant to any invertible marginal transformation of $\mathbf{Z}_2$. A consequence of this invariance, however, is that the optimization problem in (5.30) does not have a unique optimal solution. In particular, any map that meets the condition in Proposition 12 will satisfy $\mathbf{Z}_2 \perp\!\!\!\perp \mathbf{Y}$. This includes the triangular map components in the KR rearrangement that push forward $\pi_{\mathbf{X}|\mathbf{Y}}$ to the standard Gaussian reference $\eta_{\mathbf{Z}_2}$. Let us remark that the KR rearrangement also enforces that all components of $\mathbf{Z}_2$ are independent and have a standard Gaussian distribution. The optimization problem

[4]Mutual information is a special case of the conditional mutual information introduced in Chapter 4. Furthermore, the expected information gain introduced in Section 5.3.4 is the mutual information between the parameters $\mathbf{X}$ and observations $\mathbf{Y}$ in a Bayesian inference problem.

in (5.30) relaxes these later two conditions on $\mathbf{Z}_2$. This is important when considering a constrained function space $\mathcal{S}$ for the map. In this case, there may exist a solution for $\widehat{S}^{\mathcal{X}}$ in $\mathcal{S}$ that satisfies $\mathrm{I}(\widehat{S}^{\mathcal{X}}(\mathbf{Y}, \mathbf{X}); \mathbf{Y}) = 0$, even when there is no map $S^{\mathcal{X}}$ that satisfies $S^{\mathcal{X}}(\boldsymbol{y}, \cdot)^\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x}) = \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})$ for all $\boldsymbol{y}$.

Furthermore, the value of the mutual information objective is interpretable for conditional sampling. By rewriting the mutual information in (5.29) in terms of the KL divergence from $\pi_{\mathbf{Z}_2}$ to $\pi_{\mathbf{Z}_2|\mathbf{Y}}$ in expectation over $\mathbf{Y}$, we have

$$\begin{aligned} I(\mathbf{Z}_2; \mathbf{Y}) &= \mathbb{E}_{\pi_{\mathbf{Y}}} \left[D_{\mathrm{KL}}(\pi_{\mathbf{Z}_2|\mathbf{Y}}(\cdot|\mathbf{Y}) || \pi_{\mathbf{Z}_2}) \right] \\ &= \mathbb{E}_{\pi_{\mathbf{Y}}} \left[D_{\mathrm{KL}}(\widehat{S}^{\mathcal{X}}(\mathbf{Y}, \cdot)_\sharp \pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}) || \pi_{\mathbf{Z}_2}) \right] \\ &= \mathbb{E}_{\pi_{\mathbf{Y}}} \left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}) || \widehat{S}^{\mathcal{X}}(\mathbf{Y}, \cdot)^\sharp \pi_{\mathbf{Z}_2}) \right], \end{aligned} \tag{5.31}$$

where in the last two equalities we used the definition of the conditional reference and the invariance of KL to monotone transformations. Recalling the density π_c in (5.23) for the output of the composed map we see that the mutual information in (5.31) gives us the KL divergence from π_c to the conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}$ in expectation over the observations $\mathbf{Y}$. Therefore, solving (5.30) can be seen as minimizing the posterior approximation error from using the composed map $\widehat{T}_{\boldsymbol{y}^*}$.

We can also interpret (5.30) as an optimization problem over the conditional reference densities $\pi_{\mathbf{Z}_2|\mathbf{Y}}$. From this perspective, we can also find maps that pushforward $\pi_{\mathbf{X}|\mathbf{Y}}$ to select reference densities that are independent of $\mathbf{Y}$. Two natural choices for the reference are:

1. The prior density for the parameters, $\pi_{\mathbf{X}}$. If the target conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}$ inherit features from the prior distribution (e.g., they have the same essential support), a map $S^{\mathcal{X}}$ that pulls back a non-Gaussian prior to the conditionals may not need to encode these features.

2. The Wasserstein barycenter of the family of densities $\pi_{\mathbf{X}|\mathbf{Y}}$ is defined by minimizing the transportation distance (in Wasserstein distance) from each conditional to the barycenter in expectation over $\boldsymbol{y}$. We refer to [209, 234] for computa-

tional algorithms to find the barycenter of an infinite number of conditionals for $\mathbf{X}|\mathbf{Y}$ with continuous variables $\mathbf{Y}$.

In this work we will regularize the map optimization problem in (5.30), and implicitly choose a family of reference densities $\pi_{\mathbf{Z}_2|\boldsymbol{y}}$, by seeking simple parametric maps $S^{\mathcal{X}}$. The remainder of this section is organized as follows. In subsection 5.5.1 we present a multivariate Gaussian example and in subsection 5.5.2 we present a general algorithm for solving (5.30) in non-Gaussian settings.

5.5.1 Multivariate Gaussian example

In this section we find an affine map $S^{\mathcal{X}}$ that minimizes the mutual information when $\pi_{\mathbf{X},\mathbf{Y}}$ is a multivariate Gaussian distribution. A general map $S^{\mathcal{X}}$ that pushes forward $\pi_{\mathbf{X}|\mathbf{Y}}$ to the standard normal reference $\eta_{\mathbf{Z}_2}$ has the form $S^{\mathcal{X}}(\boldsymbol{y},\boldsymbol{x}) = \mathbf{A}(\boldsymbol{x} - \mathbf{B}\boldsymbol{y} + \mathbf{c})$ for some invertible lower triangular matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, matrix $\mathbf{B} \in \mathbb{R}^{d\times m}$ and constant $\mathbf{c} \in \mathbb{R}^d$. The MI objective, however, is invariant to the additive constant $\mathbf{c}$ and the matrix $\mathbf{A}$. Thus, it is sufficient to consider maps of the form $\mathbf{Z}_2 = \mathbf{X} - \mathbf{B}\mathbf{Y}$. In this case, the marginal distribution for $\mathbf{Z}_2$ is a multivariate Gaussian with mean $\mu_{\mathbf{Z}_2}$ and covariance $\Sigma_{\mathbf{Z}_2}$ given by

$$\begin{aligned} \mu_{\mathbf{Z}_2} &= \mu_{\mathbf{X}} - \mathbf{B}\mu_{\mathbf{Y}} \\ \Sigma_{\mathbf{Z}_2} &= \Sigma_{\mathbf{X}} - \mathbf{B}\Sigma_{\mathbf{Y},\mathbf{X}} - \Sigma_{\mathbf{X},\mathbf{Y}}\mathbf{B}^T + \mathbf{B}\Sigma_{\mathbf{Y}}\mathbf{B}^T. \end{aligned} \tag{5.32}$$

Our goal is then to optimize $\mathbf{B}$ such that the mutual information $I(\mathbf{Z}_2;\mathbf{Y})$ is minimized. From the decomposition of the mutual information in terms of the marginal and joint entropies we have

$$I(\mathbf{Z}_2;\mathbf{Y}) = H(\mathbf{Z}_2) + H(\mathbf{Y}) - H(\mathbf{Z}_2,\mathbf{Y}). \tag{5.33}$$

The joint entropy is invariant under bijective transformations and thus $H(\mathbf{Z}_2,\mathbf{Y}) = H(\mathbf{X},\mathbf{Y})$. Using the entropy for multivariate Gaussian distributions, (5.33) can be

written as

$$
\begin{aligned}
I(\mathbf{Z}_2;\mathbf{Y}) &= \frac{1}{2}\log|\Sigma_{\mathbf{Z}_2}| + \frac{1}{2}\log|\Sigma_{\mathbf{Y}}| - \frac{1}{2}\log|\Sigma_{\mathbf{Y}}(\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{X},\mathbf{Y}}^T)| \\
&= \frac{1}{2}\log|\Sigma_{\mathbf{Z}_2}| - \frac{1}{2}\log|\Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{X},\mathbf{Y}}^T|.
\end{aligned} \tag{5.34}
$$

Given that the mutual information is lower bounded by zero, the objective is minimized when $\Sigma_{\mathbf{Z}_2} = \Sigma_{\mathbf{X}} - \Sigma_{\mathbf{X},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{X},\mathbf{Y}}^T = \Sigma_{\mathbf{X}|\mathbf{Y}}$, which corresponds to setting $\mathbf{B} = \Sigma_{\mathbf{X},\mathbf{Y}}\Sigma_{\mathbf{Y}}^{-1}$ in (5.32). While the mutual information objective does not indicate how to choose $\mathbf{A}$ and $\mathbf{c}$, we can arrive at the choice $\mathbf{A} = \mathbf{I}_d$ and $\mathbf{c} = \mathbf{0}$ by seeking a map $S^{\mathcal{X}}$ that satisfies $I(\mathbf{Z};\mathbf{Y}) = 0$ *and* that is closest to the identity function with respect to $\boldsymbol{x}$.

5.5.2 Optimizing mutual information

We now minimize the mutual information for general non-Gaussian joint distributions of $(\mathbf{X},\mathbf{Y})$. Minimizing the MI objective in (5.31) requires evaluating the marginal density $\pi_{\mathbf{Z}_2}$ in (5.23). The form of this density depends on the map $S^{\mathcal{X}}$ and the joint density $\pi_{\mathbf{X},\mathbf{Y}}$, which is unavailable in a likelihood-free setting. To circumvent this, we consider a variational upper bound for the mutual information [168]. Let $q_{\mathbf{Z}_2}\colon \mathbb{R}^d \to \mathbb{R}_+$ be a density that approximates $\pi_{\mathbf{Z}_2}$. The mutual information is then bounded as

$$
\begin{aligned}
\mathrm{I}(\mathbf{Z}_2;\mathbf{Y}) = \mathbb{E}_{\pi_{\mathbf{Z},\mathbf{Y}}}\left[\log\frac{\pi_{\mathbf{Z}_2|\mathbf{Y}}(\mathbf{Z}_2|\mathbf{Y})}{\pi_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right] &= \mathbb{E}_{\pi_{\mathbf{Z}_2,\mathbf{Y}}}\left[\log\frac{\pi_{\mathbf{Z}_2|\mathbf{Y}}(\mathbf{Z}_2|\mathbf{Y})q_{\mathbf{Z}_2}(\mathbf{Z}_2)}{\pi_{\mathbf{Z}_2}(\mathbf{Z}_2)q_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right] \\
&= \mathbb{E}_{\pi_{\mathbf{Z}_2,\mathbf{Y}}}\left[\log\frac{\pi_{\mathbf{Z}_2|\mathbf{Y}}(\mathbf{Z}_2|\mathbf{Y})}{q_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right] - D_{\mathrm{KL}}(\pi_{\mathbf{Z}_2}||q_{\mathbf{Z}_2}) \\
&\leq \mathbb{E}_{\pi_{\mathbf{Z}_2,\mathbf{Y}}}\left[\log\frac{\pi_{\mathbf{Z}_2|\mathbf{Y}}(\mathbf{Z}_2|\mathbf{Y})}{q_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right],
\end{aligned} \tag{5.35}
$$

where the last inequality follows from the positivity of the KL divergence. The inequality in (5.35) holds for any $q_{\mathbf{Z}_2}$ such that the expectations above are finite, and it is an equality when $q_{\mathbf{Z}_2} = \pi_{\mathbf{Z}_2}$. Furthermore, to find the tightest upper bound for the mutual information, we can minimize the KL divergence $D_{\mathrm{KL}}(\pi_{\mathbf{Z}_2}||q_{\mathbf{Z}_2})$ over $q_{\mathbf{Z}_2}$.

Given a map $S^{\mathcal{X}}$, we evaluate it at samples $(\boldsymbol{X}^i, \boldsymbol{Y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$ to generate samples $\mathbf{Z}_2^i = S^{\mathcal{X}}(\boldsymbol{Y}^i, \boldsymbol{X}^i)$ from the marginal distribution for $\mathbf{Z}_2$. We can then use these samples to define $q_{\mathbf{Z}_2}$. For instance, we can learn a monotone triangular transport map $Q\colon \mathbb{R}^d \to \mathbb{R}^d$ using the framework in Chapter 3 such that $q_{\mathbf{Z}_2} = Q^\sharp \eta = \pi_{\mathbf{Z}_2}$, where η is the standard Gaussian density on $\mathbb{R}^d$. We then find $S^{\mathcal{X}}$ that minimizes (5.35), or equivalently $\mathbb{E}_{\pi_{\mathbf{Y}}}[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||S^{\mathcal{X}}(\mathbf{Y},\cdot)^\sharp q_{\mathbf{Z}_2})]$, by solving the optimization problem

$$\arg\min_{s} \frac{1}{n}\sum_{i=1}^{n}\left[-\log q_{\mathbf{Z}_2}(s(\boldsymbol{Y}^i, \boldsymbol{X}^i_{\leq k})) - \log|\partial_{m+k} s(\boldsymbol{Y}^i, \boldsymbol{X}^i_{\leq k})|\right] \tag{5.36}$$

over the space of lower triangular maps $s\colon \mathbb{R}^{m\times d} \to \mathbb{R}^d$ with the monotonicity constraints $\partial_k s(\boldsymbol{y}, \boldsymbol{x}_{\leq k}) > 0$ for the map components $k = 1, \ldots, d$. Let us remark that this optimization problem is only convex with respect to s if $q_{\mathbf{Z}_2}$ is a log-concave density. This result closely follows the proof of convexity for the standard Gaussian reference in Appendix A. Unlike the optimization problem in (5.4) that is separable over the d map components, however, the map components must be found jointly in (5.36) when $q_{\mathbf{Z}_2}$ is not a product of independent marginal densities.

Each map $S^{\mathcal{X}}$ defines a new upper bound for the mutual information. Thus we can repeatedly compute and refine the upper bound as we update the map. An iterative approach to minimize an objective function using a sequence of upper bounds is the majorize-minimize (MM) framework [117]. Let $\mathrm{I}(s(\mathbf{Y},\mathbf{X});\mathbf{Y})$ be the mutual information with respect to the map s. For a map s^t at iteration t, we construct a surrogate function $J(s; q_{\mathbf{Z}_2}) = \mathbb{E}_{\pi_{\mathbf{Y}}} D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||s(\mathbf{Y},\cdot)^\sharp q_{\mathbf{Z}_2})$ by choosing $q_{\mathbf{Z}_2} = \pi_{\mathbf{Z}_2} = \int s^t(\boldsymbol{y},\cdot)_\sharp \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})\pi_{\mathbf{Y}}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}$ so that $D_{\mathrm{KL}}(q_{\mathbf{Z}_2}||\pi_{\mathbf{Z}_2}) = 0$. The surrogate satisfies $I(s(\mathbf{Y},\mathbf{X});\mathbf{Y}) \leq J(s; q_{\mathbf{Z}_2})$ for all invertible maps s and $J(s^t; q_{\mathbf{Z}_2}) = \mathrm{I}(s^t(\mathbf{Y},\mathbf{X});\mathbf{Y})$. A surrogate satisfying these properties is known as a majorizer. We then solve (5.36) to define the new map s^{t+1} and repeat these steps until convergence. [231] showed that the related expectation-maximization algorithm—an instance of the MM framework—will either converge to a local optimum or a saddle point of the objective. We define convergence as the value of the objective (5.36) at the minimizing map across two iterations being small. The complete procedure is summarized in

Algorithm 6.

Algorithm 6: Majorization-Minimization algorithm for MI minimization

Input : i.i.d. samples $\{(\boldsymbol{X}^i, \boldsymbol{Y}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Y}}$, observation $\boldsymbol{y}^*$

Output: Approximate posterior samples $\{\boldsymbol{X}_c^i\}_{i=1}^n$

Initialize $q_{\mathbf{Z}_2} = \pi_{\mathbf{X}}$

while *Not Converged* **do**

  Compute $\widehat{S}^{\mathcal{X}}$ given joint samples to push forward $\pi_{\mathbf{X}|\mathbf{Y}}$ to $q_{\mathbf{Z}_2}$

  Evaluate residual samples $\boldsymbol{Z}_2^i = \widehat{S}^{\mathcal{X}}(\boldsymbol{Y}^i, \boldsymbol{X}^i)$ for $i = 1, \dots, n$

  Compute $q_{\mathbf{Z}_2}$ given residual samples

Generate posterior samples using composed map $\boldsymbol{X}_c^i = \widehat{T}_{\boldsymbol{y}^*}(\boldsymbol{X}^i, \boldsymbol{Y}^i)$

For a numerical example, we consider sampling from a one-dimensional density in the location-scale family (see Example 3) where the conditional mean is given by $\mu(y) = \sin(y)$ and each conditional is a Laplace density, i.e., $\pi_{X|y} = \mathcal{L}(\sin(y), 0.5)$. We specify the marginal density for the observation to be $\pi_Y = \mathcal{N}(0, 1)$ and generate samples as $Y^i \sim \pi_Y$ and $X^i \sim \pi_{X|Y}(\cdot|Y^i)$. An approximation to the conditional density given by $S^{\mathcal{X}}(y, \cdot)^\sharp \eta$ for the Gaussian reference density η requires a map that can represent the heavy Laplace tails. On the other hand, the map $S^{\mathcal{X}}(y, x) = x - \sin(y)$ can remove the mean dependence on y and be used to derive the composed transformation $T_{y^*}(y, x) = x - \sin(y) + \sin(y^*)$ that can sample exactly from the conditional density. Furthermore, the map $S^{\mathcal{X}}(y, x) = x - \sin(y)$ yields the marginal reference $\pi_{Z_2|y} = \mathcal{L}(z_2; 0, 0.5)$, which satisfies the condition in Proposition 12 of $Z_2 = S^{\mathcal{X}}(Y, X)$ being independent of Y.

Figures 5-15a plots the joint density $\pi_{X,Y}$ and the conditional density, corresponding to the slice at $y^* = -1$. In this experiment, we consider parametric maps with the monotone representation in Chapter 3. For all computations, we consider maps of the form $S^{\mathcal{X}}(y, x) = f(y, 0) + \int_0^x g(\partial_t f(y, t))dt$ where $f : \mathbb{R}^2 \to \mathbb{R}$ is a linear expansion of univariate Hermite polynomials of maximum degree 3 (see Chapter 3) that only depend on either x or y. These separable maps of high-enough degree are sufficient to approximate the map $S^{\mathcal{X}}(y, x) = x - \sin(y)$. Figure 5-15b plots the one-dimensional Laplace density for $\pi_{X|y^*=-1}$ and the pullback density $S^\sharp \eta$ of a single map through

a one-dimensional Gaussian reference. We observe that a single map is unable to represent the heavy tails and sharp cusp in the density near $x = -1$.

We apply Algorithm 6 to optimize the map with two classes of reference densities $q_{\mathbf{Z}_2}$: a measure transport approximation and a non-parametric kernel approximation. For the measure transport approach, we use i.i.d. samples $Z_2^i = \widehat{S}^{\mathcal{X}}(Y^i, X^i)$ for $i = 1, \ldots, n$ to learn a monotone map $Q\colon \mathbb{R} \to \mathbb{R}$ such that $Q^\sharp \eta \approx \pi_{Z_2}$ where Q is represented using an expansion of linearized Hermite polynomials; see Chapter 3. We use 5-fold cross-validation to select the maximum degree of the polynomial expansion.

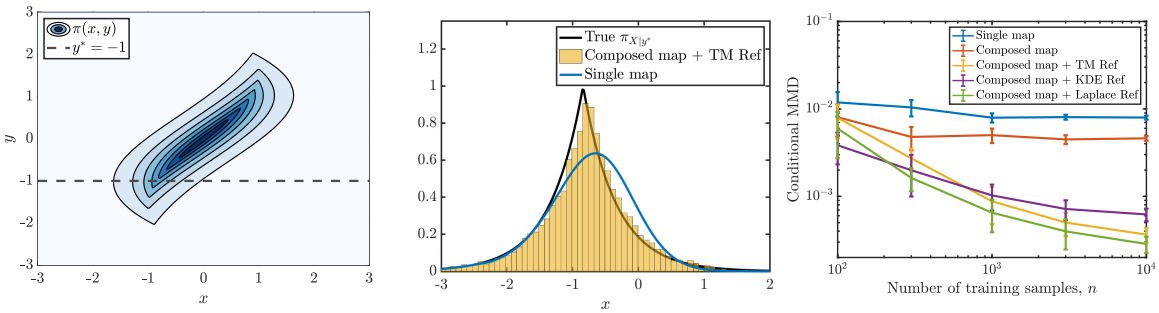
Alternatively, we also consider the kernel density estimator (KDE) for the marginal density $\pi_{\mathbf{Z}_2}$ given by

$$\widehat{\pi}_{Z_2}(z_2) \coloneqq \frac{1}{n} \sum_{i=1}^{n} K(\|z_2 - Z_2^i\|_\sigma), \tag{5.37}$$

where $K\colon \mathbb{R} \to \mathbb{R}$ denotes a kernel function and σ is the bandwidth or smoothing parameter. One can generalize the approximation in (5.37) for $d > 1$ by seeking a positive definite matrix H instead of the scalar σ. In our experiments we use the Gaussian kernel $K(r) = \exp(-r^2)$ and use 5-fold cross-validation to select the bandwidth. Let us remark that even though a kernel approximation is used when solving (5.36), we still find a parametric form for the map and the kernel density estimator in (5.37) is not used in the subsequent step for conditional sampling. Instead, we compose the resulting map $\widehat{S}^{\mathcal{X}}$ to derive $\widehat{T}_{y^*}$.

Given (approximate) posterior samples from either the single map or the composed map, we estimate the maximum mean discrepancy (MMD) between the approximate and the true conditional density $\pi_{X|y^*=-1}$. The MMD is an integral probability metric that is commonly used in the LFI literature to compare distributions given only samples [136]. Figure 5-15c plots the MMD for different map estimators using an increasing number of training samples $n \in [10^2, 10^4]$ and 10^4 independent samples from the true conditional density. The error-bars indicate the standard error over 30 replications of each experiment. We include the MMD for using the composed map that does not optimize the reference (i.e., the SM algorithm alone), as well as an estimator that uses the ideal centered Laplace reference density $\mathcal{L}(0, 0.5)$ in place

of the standard Gaussian in (5.4). As expected from Proposition 11, the composed map improves upon the error of the single map, but remains biased with increasing sample size n from the limited capacity of the map. In comparison, we observe a decreasing error when optimizing the reference with either the measure transport or non-parametric reference, labelled as TM Ref and KDE Ref, respectively. Furthermore, the composed maps with the optimized reference densities closely match the error of the composed map that is computed using the Laplace reference density.



(a) (b) (c)

Figure 5-15: (a) Joint density for the Laplace model with the conditional slice at $y^* = -1$ denoted by a dashed line. (b) The true and approximate conditional densities using a single map with maximum polynomial degree of 10. (c) The maximum mean discrepancy from the true conditional density to various approximations with an increasing number of training samples.

5.6 Conditional sampling via GANs

In this section we present an alternative approach for finding monotone transport maps $G^{\mathcal{X}}(\boldsymbol{y}, \cdot)\colon \mathbb{R}^d \to \mathbb{R}^d$ that push forward a reference density $\eta_{\mathbf{Z}_2}$ (e.g., a standard Gaussian) to the conditional density $\pi_{\mathbf{X}|\boldsymbol{y}}$ for each $\boldsymbol{y} \in \mathbb{R}^m$. We can use these maps to cheaply sample from the conditional density $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y}^*)$ given an observation $\boldsymbol{y}^*$ by generating samples $\mathbf{Z}_2^i \sim \eta_{\mathbf{Z}_2}$ and evaluating the map $G^{\mathcal{X}}(\boldsymbol{y}^*, \mathbf{Z}_2^i)$. These posterior samples are obtained without inverting the map, as compared to the approach in Section 5.2 that constructs the inverse transport map $S^{\mathcal{X}}(\boldsymbol{y}, \cdot)$ pushing forward $\pi_{\mathbf{X}|\boldsymbol{y}}$ to $\eta_{\mathbf{Z}_2}$. This makes $G^{\mathcal{X}}$ a so-called *generative model* that can be used for likelihood-free inference. To easily learn the map without requiring the inversion of $G^{\mathcal{X}}(\boldsymbol{y}, \cdot)$,

we will use an adversarial training objective that compares evaluations of the map to samples from the target density. This does not require computing the pushforward density or evaluating the determinant of the map. Hence, we are free to consider block-triangular maps, meaning that each component of $G^{\mathcal{X}}$ is a function of both parameters and data. This structure generalizes the triangular structure of the KR rearrangement, at the expense of the unique solution to the optimization problem.

In subsection 5.6.1, we present the adversarial training framework for learning a monotone map $G^{\mathcal{X}}$, that is commonly used for learning generative adversarial networks (GAN)s [78]. Subsection 5.6.2 discusses the properties of the minimizer and its connection to optimal transport maps; see Chapter 1. Lastly, subsections 5.6.3-5.6.5 present numerical results that the highlight the benefit of block-triangularity and the feasibility of the resulting maps for likelihood-free inference problems.

5.6.1 Adversarial training of transport maps

Let $D\colon \mathbb{P}(\mathbb{R}^{m+d}) \times \mathbb{P}(\mathbb{R}^{m+d}) \to \mathbb{R}$ denote a functional that compares probability measures on $\mathbb{R}^{m+d}$. In this work, we consider the Kantorovich-Rubinstein type divergences

$$D(\mu_1, \mu_2; \mathcal{F}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{\mu_1} f - \mathbb{E}_{\mu_2} f,$$

where $\mathcal{F}$ is an appropriate class of functions from $\mathbb{R}^{d+m} \to \mathbb{R}$ for computing the expectations with respect to each measure. For instance, if $\mathcal{F}$ is the class of Lipschitz-1 functions, D corresponds to the dual representation of the Wasserstein-1 metric. Alternatively, if $\mathcal{F}$ is a reproducing Kernel Hilbert space, D corresponds to the maximum mean discrepancy metric. Both of these choices have been used to find generative models [9].

The divergences above can asses the quality of a transport map for approximating conditional distributions. Let G be the block-triangular transport map

$$G(\boldsymbol{y}, \boldsymbol{z}_2) = \begin{bmatrix} \mathrm{I}_m(\boldsymbol{y}) \\ G^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{z}_2) \end{bmatrix}. \tag{5.38}$$

This map is partitioned similarly to (5.3), with the exception that the first block is kept as the identity function. This choice avoids approximating the marginal of $\mathbf{Y}$. We let the reference distribution have the product measure $\boldsymbol{\nu}_{\mathbf{Y}} \otimes \boldsymbol{\nu}_{\mathbf{Z}_2}$ with density $\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}$, where $\eta_{\mathbf{Z}_2}$ is the standard Gaussian density on $\mathbb{R}^d$. Our goal is to find $G\colon \mathbb{R}^{m+d} \to \mathbb{R}^{m+d}$ by solving

$$\arg\min_{G} D(G_\sharp(\boldsymbol{\nu}_{\mathbf{Y}} \otimes \boldsymbol{\nu}_{\mathbf{Z}_2}), \boldsymbol{\nu}_{\mathbf{Y},\mathbf{X}}; \mathcal{F}), \tag{5.39}$$

where $\boldsymbol{\nu}_{\mathbf{Y},\mathbf{X}}$ denotes the joint measure of $\mathbf{Y}$ and $\mathbf{X}$. For a block-triangular map of the form in (5.38) that satisfies $G_\sharp(\boldsymbol{\nu}_{\mathbf{Y}} \otimes \boldsymbol{\nu}_{\mathbf{Z}_2}) = \boldsymbol{\nu}_{\mathbf{Y},\mathbf{X}}$ it can be shown that $G^{\mathcal{X}}(\boldsymbol{y}, \cdot)_\sharp \eta_{\mathbf{Z}_2} = \pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ for $\boldsymbol{y} \in \operatorname{supp}(\pi_{\mathbf{Y}})$ almost everywhere; see [113]. For the remainder of this section, we abuse the notation above by assuming densities exist and using them in place of measures.

In general, there exists an infinite number of block-triangular maps that couple $\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}$ and $\pi_{\mathbf{Y},\mathbf{X}}$. For example, the Knothe-Rosenblatt re-arrangement is the monotone and lower triangular map in (2.3) that pushes forward one density to another and achieves a value of zero for the objective in (5.39). In this work, we regularize the optimization problem above by looking for a map G that is monotone. The map G in (5.38) is monotone in the vector sense if $\langle G^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{z}_2) - G^{\mathcal{X}}(\boldsymbol{y}', \boldsymbol{z}_2'), \boldsymbol{z}_2 - \boldsymbol{z}_2' \rangle \geq 0$ for all $(\boldsymbol{y}, \boldsymbol{z}_2), (\boldsymbol{y}', \boldsymbol{z}_2') \in \operatorname{supp}(\pi_{\mathbf{Y}}\pi_{\mathbf{Z}_2})$, where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Let us remark that the first block of G is monotone by construction. To easily differentiate the objective, we consider the average monotonicity penalty

$$R(G) = \mathbb{E}_{(\mathbf{Y},\mathbf{Z}_2)\sim\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}} \mathbb{E}_{(\mathbf{Y}',\mathbf{Z}_2')\sim\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}} \langle G^{\mathcal{X}}(\mathbf{Y}, \mathbf{Z}_2) - G^{\mathcal{X}}(\mathbf{Y}', \mathbf{Z}_2'), \mathbf{Z}_2 - \mathbf{Z}_2' \rangle. \tag{5.40}$$

We then look for the map G by solving the optimization problem

$$\min_{G\in\mathcal{G}} \max_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}} f(\mathbf{Y}, G^{\mathcal{X}}(\mathbf{Y}, \mathbf{Z}_2)) - \mathbb{E}_{\pi_{\mathbf{Y},\mathbf{X}}} f(\mathbf{Y}, \mathbf{X}) - \lambda R(G), \tag{5.41}$$

where $\lambda > 0$ is a positive constant. We refer to the solution of (5.41) as a Monotone GAN, or MGAN.

In practice, we solve (5.41) using an alternating gradient descent procedure that

is commonly used for training GANs [78]. This procedure repeats the following two steps: (1) update G holding f fixed; and (2) update f holding G fixed. Informally, the first step improves the map so that the push-forward density of G is closer to $\pi_{\mathbf{Y},\mathbf{X}}$, and the second step updates the function f, often referred to as a discriminator, to better distinguish "real" versus "fake" samples from $\pi_{\mathbf{Y},\mathbf{X}}$ and $G_\sharp(\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2})$, respectively.

In the experiments below, we parameterize the functions G and f using deep neural networks. We update the weights and biases of these networks using the Adam optimizer with the parameters $\beta_1 = 0.5$, and $\beta_2 = 0.999$. At each gradient descent step, we replace the expectations in (5.41) with empirical averages over mini-batches of samples of size 100. We note that since the reference contains the marginal of $\mathbf{Y}$, we sample from the training data each time to estimate the first expectation in (5.41), independently from sampling from $\pi_{\mathbf{Y},\mathbf{X}}$ to estimate the second expectation in (5.41). We update the parameters of G and f over multiple epochs (i.e., passes through the training set), until the values for the divergence and the regularization term converge. All of the examples use three-layer, fully-connected neural networks with hidden layer sizes specified below and the Leaky ReLU non-linearity with parameter 0.2. In our experiment, we use the WGAN-GP functional introduced in [82] to define the discrepancy. This functional considers Lipschitz-1 functions f by adding the average gradient penalty $\gamma\mathbb{E}_{\pi_{\mathbf{Y},\mathbf{X}}}(\|\nabla f(\mathbf{X},\mathbf{Y})\|^2 - 1|)^2$ to the objective, where $\gamma > 0$ is a positive constant. Additional numerical results using MGAN for image in-painting problems can be found in [113].

5.6.2 Connection to optimal transport

Various forms of regularization have been proposed when finding general transport maps G, primarily for the purpose of stabilizing the numerical optimization procedure. These techniques include penalizing the gradients of G, penalizing neural network weights, and adding normalization layers to the networks; see [115, 144] for recent overviews and comparisons of these techniques for GANs. The choice of regularization that provides the best empirical performance on a range of datasets remains an open problem. Furthermore, to the best of our knowledge, little is known

about the minimizers of (5.41) with different forms of regularization, i.e., for different choices of the penalty R. In this section, we show that a monotonicity constraint yields transport maps G that are *optimal* in a specific sense.

Given two measures $\boldsymbol{\nu}_\eta$ and $\boldsymbol{\nu}_\pi$ on $\mathbb{R}^d$, optimal transport looks for maps G that satisfy $G_\sharp \boldsymbol{\nu}_\eta = \boldsymbol{\nu}_\pi$ and minimize $\int c(\boldsymbol{z}, G(\boldsymbol{z}))\mathrm{d}\boldsymbol{\nu}_\eta$ where $c(\boldsymbol{z}, \boldsymbol{x})$ measures the cost of transporting one unit of mass from $\boldsymbol{z}$ to $\boldsymbol{x}$. In general, finding an optimal transport map is challenging and, in fact, is not guaranteed to always exist for general measures $\boldsymbol{\nu}_\eta$ and $\boldsymbol{\nu}_\pi$. One setting, however, where the map is well known is for the quadratic cost $c(\boldsymbol{z}, \boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm of $\mathbb{R}^d$. In this case, a celebrated result of Brenier that was later generalized by McCann [142], showed that under mild assumptions on $\boldsymbol{\nu}_\eta$ and $\boldsymbol{\nu}_\pi$ (stated in Theorem 1.22 of [189]) there exists a convex function ψ such that the map $G = \nabla\psi$ pushes forward $\boldsymbol{\nu}_\eta$ to $\boldsymbol{\nu}_\pi$ and is unique for $\boldsymbol{z} \in \mathrm{supp}(\boldsymbol{\nu}_\eta)$ almost everywhere. The next proposition shows that this result also extends to the conditional setting, i.e., for maps $G^{\mathcal{X}}(\boldsymbol{y}, \cdot)$ that push forward $\boldsymbol{\nu}_{\eta_2}$ to the conditional measure $\boldsymbol{\nu}_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ for each $\boldsymbol{y}$. This result and its proof is given in Theorem 2.3 of [32].

Proposition 13. *Let $\mathbf{Z}_2$ be a random variable that has density with respect to the Lebesgue measure on $\mathbb{R}^d$ with convex support (e.g., the standard Gaussian distribution on $\mathbb{R}^d$). Let $\boldsymbol{z}_2 \mapsto G^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{z}_2) = \nabla_{\boldsymbol{z}_2}\psi(\boldsymbol{y}, \boldsymbol{z}_2)$ be a transport map that pushes forward $\eta_{\mathbf{Z}_2}$ to the conditional measure $\boldsymbol{\nu}_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathcal{Y}$ where ψ is a convex function. Then, $G^{\mathcal{X}}$ is the unique monotone map among all such maps that are the gradients of a convex function and it minimizes the transport cost $\mathbb{E}_{\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}}\|\mathbf{Z}_2 - G(\mathbf{Y}, \mathbf{Z}_2)\|^2$.*

Remark. *The authors in [32] also impose the constraint that $\mathbf{Z}_2$ is conditionally independent of $\mathbf{Y}$ in the optimal transport problem to find the map $G^{\mathcal{X}}$, i.e., $\mathbf{Z}_2|\mathbf{Y} \sim \eta_{\mathbf{Z}_2}$. This property is automatically imposed for the tensor-product reference measure $\boldsymbol{\nu}_{\mathbf{Y}} \otimes \boldsymbol{\nu}_{\mathbf{Z}_2}$ we consider in this section.*

A consequence of Proposition 13 is that solving (5.41) over a class of maps where $G^{\mathcal{X}}$ is the gradient of a convex function will have a unique minimizer for the map. The next example provides the unique monotone map for the case of a multivariate

joint Gaussian measure $\boldsymbol{\nu}_{\mathbf{X},\mathbf{Y}}$.

Example 4 (Gaussian random variables)**.** *The monotone transport map that pushes forward* $\eta_{\mathbf{Z}_2}$ *to the conditional measure* $\boldsymbol{\nu}_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y})$ *is the affine map* $G^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{z}_2) = \mu_{\mathbf{X}|\boldsymbol{y}} + \Sigma^{1/2}_{\mathbf{X}|\boldsymbol{y}}\boldsymbol{z}_2$ *where* $\Sigma^{1/2}_{\mathbf{X}|\boldsymbol{y}}$ *the unique positive-definite square root of the conditional covariance matrix* $\Sigma_{\mathbf{X}|\boldsymbol{y}}$ *and* $\mu_{\mathbf{X}|\boldsymbol{y}}$ *is the conditional mean of* $\mathbf{X}|\mathbf{Y} = \boldsymbol{y}$*. The map* $G^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{z}_2) = \nabla_{\boldsymbol{z}_2}\psi(\boldsymbol{y}, \boldsymbol{z}_2)$ *is realized as the gradient of the convex function* $\psi(\boldsymbol{y}, \boldsymbol{z}_2) = \mu^T_{\mathbf{X}|\boldsymbol{y}}\boldsymbol{z}_2 + \frac{1}{2}\boldsymbol{z}_2^T\Sigma^{1/2}_{\mathbf{X}|\boldsymbol{y}}\boldsymbol{z}_2$.

Ideally, the unique map satisfying the pushforward condition in Proposition 13 could be found by searching in the space of all monotone maps. In general, however, the space of multivariate maps that are gradients of convex functions is a subset of the space of monotone maps. A famous result of Rockafellar (Theorems 24.8 and 24.9 in [183]) states that G is uniquely determined by the gradient of a proper convex function if and only if G is a maximal cyclically monotone mapping. A function G is cyclically monotone if for any M and samples $\{\boldsymbol{W}^j\}_{j=1}^{M+1}$ where $\boldsymbol{W}^j \coloneqq (\boldsymbol{Y}^j, \boldsymbol{Z}_2^j) \sim \pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}$ such that $\boldsymbol{W}_{M+1} \equiv \boldsymbol{W}_1$, it holds that $\sum_{j=1}^{M}\langle G(\boldsymbol{W}^j) - G(\boldsymbol{W}^{j+1}), \boldsymbol{W}^j\rangle \geq 0$. For $M = 2$ this condition ensures that $\langle G(\boldsymbol{W}^1) - G(\boldsymbol{W}^2), \boldsymbol{W}^1\rangle + \langle G(\boldsymbol{W}^2) - G(\boldsymbol{W}^1), \boldsymbol{W}^2\rangle = \langle G(\boldsymbol{W}^1) - G(\boldsymbol{W}^2), \boldsymbol{W}^1 - \boldsymbol{W}^2\rangle \geq 0$ for all $\boldsymbol{W}^1, \boldsymbol{W}^2$, which is equivalent to G being monotone in the vector sense above. In practice one can parameterize the class of mappings that are gradients of convex functions by using input convex neural networks [3]. This approach was used in [94] to find the transport map between unconditional measures for density estimation tasks.

In the next example, we show that MGAN recovers the optimal transport map $G^{\mathcal{X}}$ minimizing the Wasserstein distance $\mathbb{E}_{\pi_{\mathbf{Y}}\eta_{\mathbf{Z}_2}}\|\mathbf{Z}_2 - G^{\mathcal{X}}(\mathbf{Y}, \mathbf{Z}_2)\|^2$, despite only penalizing average deviations from monotonicity. Given $n = 10^4$ samples from a multivariate Gaussian density $\pi_{Y,\mathbf{X}}$ with non-trivial correlations between $\mathbf{X} \in \mathbb{R}^5$ and $Y \in \mathbb{R}$, we estimate a transport map $G^{\mathcal{X}}$ for three increasing values of the monotonicity penalty λ. Figure 5-16 plots the transportation distance of the estimated maps (left), and the expected L^2 error between the estimated and the monotone map given in Example 4 (right). We observe that the maps found with the average monotonicity penalty

converge to the unique monotone map in Proposition 13 when increasing the penalty parameter λ. Future work will compare the minimizer(s) of (5.41) with the average monotonicity penalty to the solution of a constrained problem that uses maps $G^{\mathcal{X}}$ given by the gradients of convex functions.

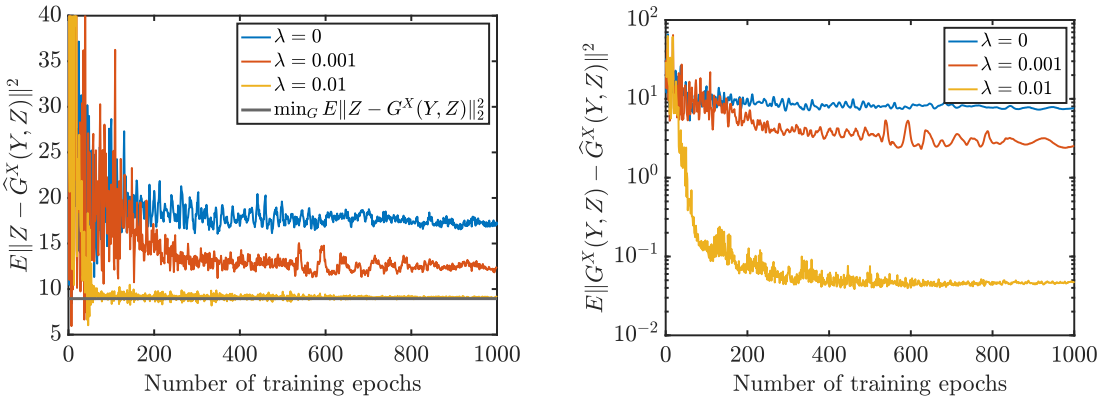


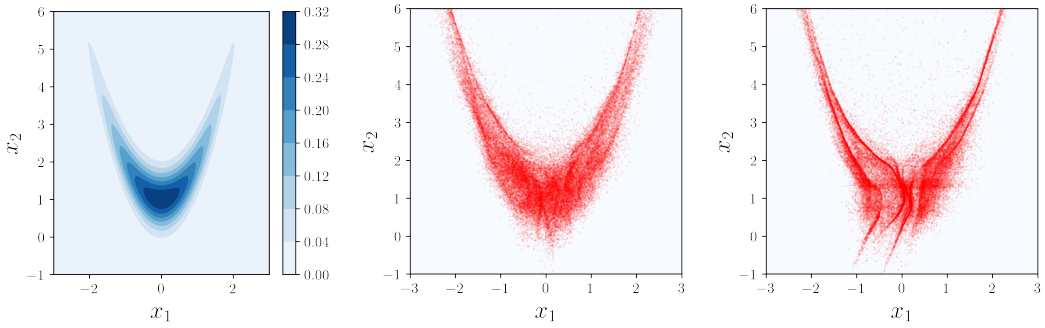
Figure 5-16: The estimated map $\widehat{G}^{\mathcal{X}}$ converges to the monotone transport map that minimizes the transportation distance $\mathbb{E}\|\mathbf{Z}_2 - G^{\mathcal{X}}(\mathbf{Y}, \mathbf{Z}_2)\|^2$ with increasing penalty parameter λ. The map converges in (a) transportation distance and in (b) the expected L^2_π error with respect to the affine monotone map in Example 4. The minimum transportation distance is computed exactly for the multivariate Gaussian joint distribution and is denoted by the black line in the left plot.

Let us remark that an alternative approach to penalizing average monotonicity is to parameterize maps $G^{\mathcal{X}}$ that are monotone by construction. Several neural network parameterizations for one-dimensional monotone increasing functions have been used to build triangular maps where each component is monotone with respect to its last variable [95, 99, 16, 62]; see Chapter 3. To increase the expressiveness of transformations built from simple triangular maps, it is often necessary to *compose* many layers of maps with variable permutations between the layers. In contrast, the MGAN framework relaxes the triangular constraint by defining block triangular maps that are less dependent on the ordering of the $\boldsymbol{x}$ variables.

5.6.3 Block triangular versus triangular maps

We now study the benefit of using block triangular rather than triangular maps. In this example, we consider a two-dimensional target distribution for $\mathbf{X} = (X_1, X_2)$ with no conditioning variables where $X_1 \sim \mathcal{N}(0, 1)$ and $X_2|x_1 = x_1 \sim \mathcal{N}(x_1^2 + 1, 0.5^2)$. The joint density of $\mathbf{X}$ can be represented exactly as the push-forward of the standard Gaussian density $\eta_{\mathbf{Z}}$ through the map $G(\boldsymbol{z}) = [z_1; z_1^2 + 1 + 0.5z_2]$. Hence, G can be easily approximated by a triangular map of this form. However, when the ordering of z_1, z_2 is reversed—i.e., when the first component of G depends on z_2 instead of z_1—the map is more challenging to approximate. We refer the reader to [162] for a similar application. We demonstrate that by using a block triangular parameterization we can avoid issues pertaining to the ordering of the variables and achieve a more robust map in practice.

We use $n = 10^4$ training samples and $\lambda = 0.01$ to train a MGAN with either fully triangular or block triangular structure. We use three-layer, fully-connected neural networks with hidden layer sizes $32 - 64 - 32$ for the block triangular maps and neural networks with hidden layer sizes $22 - 46 - 22$ for each component of the triangular map. In total, the block triangular and triangular maps have about the same number of parameters. Figure 5-17 compares the samples generated by the triangular and block triangular maps to the true density. We observe that the block triangular map is able to capture the target density independent of the ordering of the variables, unlike the triangular map. Table 5.3 reports the KL divergence between the true and approximate distributions for both variable orderings. While the block triangular map has similar performance under both the favorable and reverse orderings, the performance of the triangular map degrades significantly depending on the ordering. This suggests that block triangular maps are less sensitive to variable order. This is a major advantage of MGANs over autoregressive models where it is necessary to specify a variable ordering in advance.



(a) Target density (b) Block triangular (c) Triangular

Figure 5-17: (a) The true density of (X_1, X_2). (b) Samples generated by MGAN using a block triangular map with the reverse ordering of the variables. (c) Samples generated by a fully triangular MGAN also with reverse ordering.

	Block triangular	Triangular
Favorable order	0.056 ± 0.003	0.039 ± 0.002
Reverse order	0.058 ± 0.002	0.102 ± 0.004

Table 5.3: The KL divergence for block triangular and triangular MGANs based on $n = 10^4$ training samples. The approximate densities are estimated using KDE with an optimal bandwidth that is chosen using 5-fold cross-validation. The KL divergence is computed using 5×10^4 test samples and is reported with its 95% standard error.

5.6.4 Parameter inference for stochastic ODE models

Next, we apply use the MGAN framework to infer the parameters in both deterministic and stochastic Lotka–Volterra population models. These models describe two interacting populations such as predators and prey, but can also be applied to other biological settings. Lotka–Volterra models are defined using nonlinear ODEs where the rates of change of the two populations depend on parameters that describe the interactions between the two species. Our goal is to solve the Bayesian inference problem for the underlying parameters given limited observations of the population time-series, i.e., the system states. The stochastic Lotka–Volterra model is a classic example for LFI because the dynamical model for the states is a stochastic Markov jump process. Therefore, the forward model mapping the parameters to the states is not deterministic and evaluating the likelihood requires integrating over these random variables. Furthermore, only a narrow region of the parameter space in these models

produces realistic observations, which results in very concentrated posterior densities.

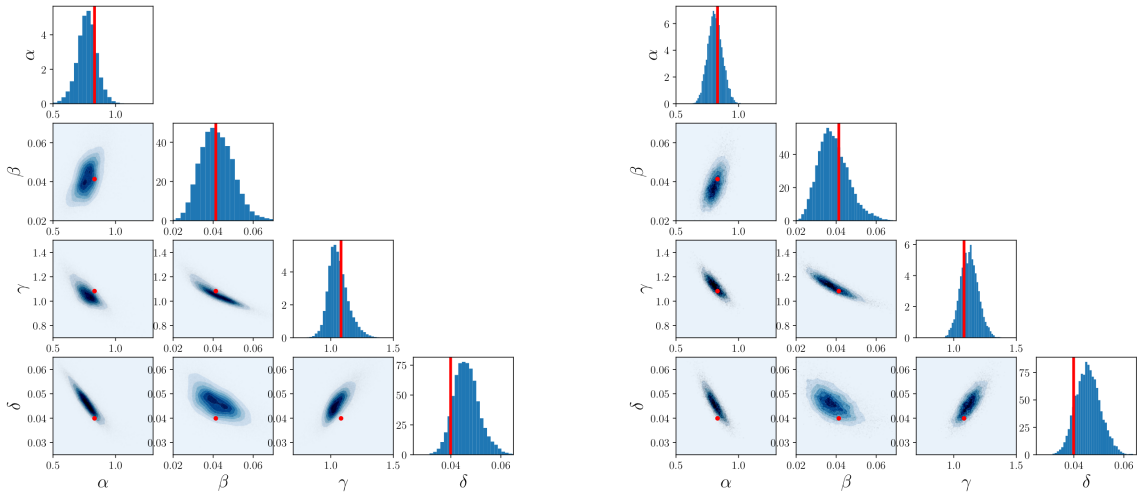
As a validation study, we first consider the posterior distribution of the parameters $\mathbf{X} = (\alpha, \beta, \gamma, \delta) \in \mathbb{R}^4$ in the deterministic Lotka–Volterra model. The states in this model $Z(t) \in \mathbb{R}^2_+$ represent the populations of predators and prey, respectively, and evolve according to the coupled ODEs

$$\begin{aligned}\frac{\mathrm{d}Z_1}{\mathrm{d}t} &= \alpha Z_1(t) - \beta Z_1(t) Z_2(t),\\ \frac{\mathrm{d}Z_2}{\mathrm{d}t} &= -\gamma Z_2(t) + \delta Z_1(t) Z_2(t),\end{aligned}$$

with the initial condition $Z(0) = (30, 1)$. We simulate these ODEs for $T = 20$ time units and collect noisy observations of the state every $\Delta t_{\text{obs}} = 2$ time units. The observations are corrupted with log-normal noise, i.e., $\log \mathbf{Y}_k \sim \mathcal{N}(Z(k\Delta t_{\text{obs}}), \sigma \mathrm{I}_2)$ for $k = 1, \ldots, 9$, with standard deviation $\sigma = 0.01$. For inference, we use an independent log-normal prior distribution for the parameters. Figure 5-18 displays the time-series $Z(t)$ (solid line) for the parameter $\boldsymbol{x}^* = (0.92, 0.05, 1.50, 0.02)$ and an observation $\boldsymbol{y}^* \in \mathbb{R}^{18}$ drawn from the likelihood model $\pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\boldsymbol{x}^*)$.

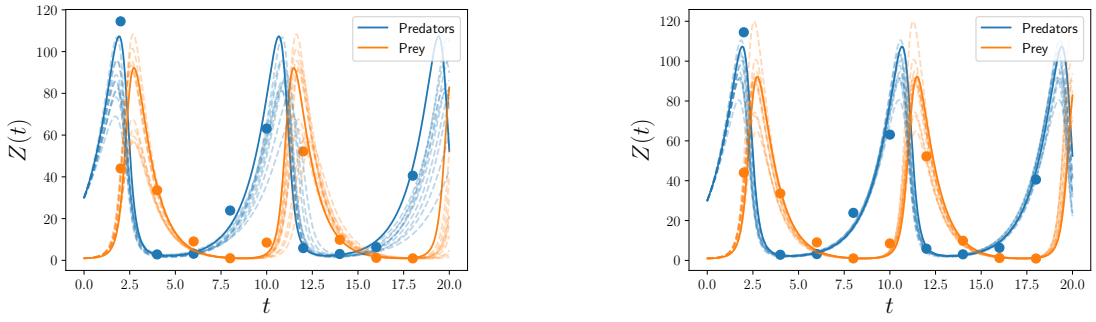
We now sample from the posterior density for $\mathbf{X}|\mathbf{Y} = \boldsymbol{y}^*$ given $n = 10^5$ training samples from the joint density $\pi_{\mathbf{Y},\mathbf{X}}$ using MGAN and MCMC. First, we train the MGAN network with the gradient penalty parameter $\gamma = 1.0$ and the monotonicity penalty parameter $\lambda = 0.1$. The left and right plots of Figure 5-18 display $100,000$ parameter samples from MGAN, i.e., $\boldsymbol{X}^i = G^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{Z}_2^i)$ for $\boldsymbol{Z}_2^i \sim \eta_{\mathbf{Z}_2}$ after learning the MGAN $G^{\mathcal{X}}$, and from an adaptive Metropolis MCMC sampler for the target density $\pi_{\mathbf{X}|\boldsymbol{y}^*}$, respectively. We observe similar marginal distributions and correlations using both methods. The true parameter $\boldsymbol{x}^*$ that generated the data (denoted in red) is contained in the bulk of the posterior distributions, and it appears like a representative sample. Lastly, we integrate the ODEs for sample realizations of the posterior parameters to sample from the predictive distribution for the states $Z(t)$. The dashed lines in Figure 5-18 plot ten posterior predictive samples for both MGAN and MCMC. While samples from both methods concentrate around the true states, we observe that the predictions from MCMC samples have less spread, especially at

later times.



(a) (b)

Figure 5-18: Posterior samples of the parameters in the deterministic Lotka–Volterra model using (a) the MGAN framework, and (b) the Adaptive Metropolis algorithm.



(a) (b)

Figure 5-19: Posterior predictive samples of the states $Z(t)$ given ten posterior samples from the (a) MGAN network and the (b) MCMC procedure.

In the stochastic Lotka–Volterra model, the populations Z_1, Z_2 evolve according to a random rate of change starting from $Z(0) = (50, 100)$. Given the model parameters $\boldsymbol{\theta} \in \mathbb{R}^4$, the populations are simulated using Gillespie's algorithm by drawing the time to the next reaction from an exponential distribution with rate $(\theta_1 + \theta_4)Z_1(t)Z_2(t) +$

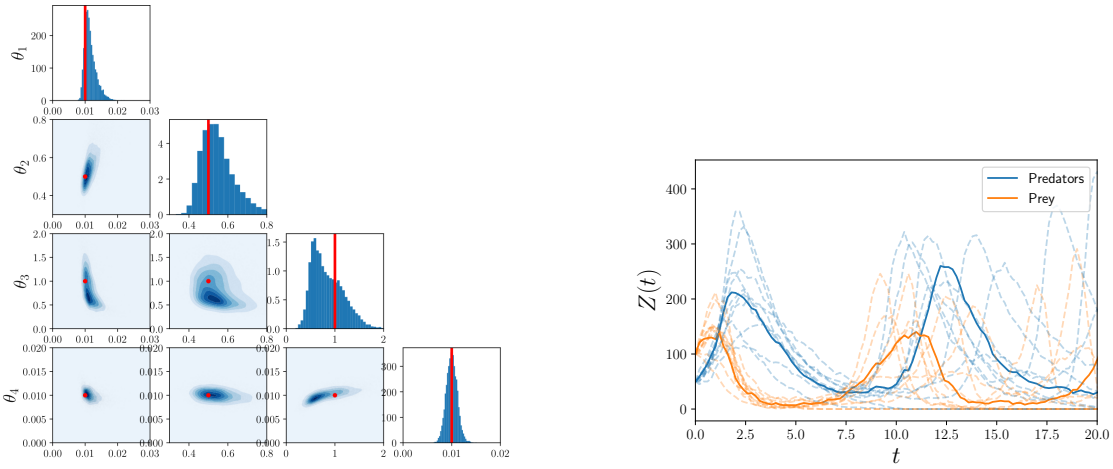
$\theta_2 Z_1(t) + \theta_3 Z_2(t)$, and simulating one of the four possible state changes

$$
\begin{aligned}
Z_1(t) &\leftarrow Z_1(t) + 1 \text{ with rate } \theta_1 Z_1(t) Z_2(t) \\
Z_1(t) &\leftarrow Z_1(t) - 1 \text{ with rate } \theta_2 Z_2(t) \\
Z_2(t) &\leftarrow Z_2(t) + 1 \text{ with rate } \theta_1 Z_1(t) Z_2(t) \\
Z_2(t) &\leftarrow Z_2(t) - 1 \text{ with rate } \theta_1 Z_1(t) Z_2(t),
\end{aligned}
$$

with probability proportional to its rate. To infer $\boldsymbol{\theta}$, we record the values of $Z(t)$ at every 0.2 time units, and compute the following statistics: the mean of the two time series, the log variance of the time series, the auto-correlation of the time series at lag 1 and 2 and the cross-correlation coefficient between the two time series. These statistics result in an observation $\mathbf{Y}$ with 9 entries. We draw $n = 50,000$ samples of $\log(\boldsymbol{\theta})$ from a uniform prior distribution and observations $\mathbf{Y}$ by simulating the stochastic model and computing the statistics above. Figure 5-20 plots the approximate posterior samples from MGAN along with the posterior predictive samples given an observation $\boldsymbol{y}^*$ sampled from the likelihood model for the parameter $\boldsymbol{\theta}^* = (0.01, 0.5, 1, 0.01)$. We observe that MGAN concentrates around the true parameter value in red and predicts the time-series well near the initial time $t = 0$. Let us remark that MCMC is not available as a baseline computational inference algorithm for this stochastic forward model.

5.6.5 Inference of permeability in Darcy flow model

In this section, we consider a benchmark elliptic inverse problem of recovering the subsurface permeability from noisy pressure measurements. Let $a \colon \mathcal{D} \to \mathbb{R}_+$ denote the permeability coefficient in the domain $\mathcal{D} = [0,1]^2$. The pressure field of the subsurface flow p is then modeled as the solution to the elliptic partial differential



(a) (b)

Figure 5-20: (a) The posterior distribution for the parameters of the stochastic Lotka–Volterra model using MGAN, and (b) the posterior predictive samples for the model states.

equation (PDE)

$$-\nabla \cdot (a(\boldsymbol{s})\nabla p(\boldsymbol{s})) = f(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \tag{5.42}$$

$$p(\boldsymbol{s}) = 0, \quad \boldsymbol{s} \in \partial\mathcal{D}, \tag{5.43}$$

under the fixed forcing $f : \mathcal{D} \to \mathbb{R}$. We introduce a log-normal random field for the permeability given by $\mathbf{X} = \log(a) \sim \mathcal{N}(0, (-\Delta + 9I)^{-2})$ where Δ is a Laplacian operator with zero Neumann boundary conditions. The inverse problem is to recover the random field $\mathbf{X}$ given noisy measurements $\mathbf{Y}$ of the pressure at 64 regularly spaced locations, i.e., $\mathbf{Y} = (p(s_1), \ldots, p(s_{64})) + \boldsymbol{\mathcal{E}}$ where $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, 10^{-6} I_{64})$. Figure 5-21 plots a realization of $\log(a)$ along with the solution to the PDE on a grid of size 256×256. In practice, we solve the PDE using finite differences and use splines to interpolate the solution at the measurement locations.

To recover the permeability a, we learn a MGAN network with the gradient penalty parameter $\gamma = 1.0$ and the monotonicity penalty parameter $\lambda = 0.01$. We use $n = 10^5$ training samples $\{(\boldsymbol{X}^i, \boldsymbol{Y}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Y}}$ given by sampling the prior $\mathbf{X}^i = \log(a^i)$ and computing measurements $\mathbf{Y}^i$ by solving the PDE with $f(\boldsymbol{s}) = 1$ for all $\boldsymbol{s} \in \mathcal{D}$. For a random field a, that is in principle infinite-dimensional, we use

principal component analysis (PCA) to make the input of the network $G^{\mathcal{X}}$ finite-dimensional. Given samples $\{\boldsymbol{X}^i\}_{i=1}^n$ in a Hilbert space $\mathcal{X}$, PCA projects the samples onto a d-dimensional subspace $V_d \subset \mathcal{X}$ that minimizes the L^2 projection error. This subspace is defined by the span of the d leading eigenvectors of the empirical covariance operator $C_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}^i \otimes \boldsymbol{X}^i$. That is, $V_d = \text{span}\{\phi_1, \ldots, \phi_d\}$ where $C_n\phi_i = \lambda_i\phi_i$ for a sequence of decreasing eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$. The orthogonal projection of a sample $\boldsymbol{x} \in \mathcal{X}$ onto V_d is then given by the mapping $F(\boldsymbol{x}) = (\langle \boldsymbol{x}, \phi_1\rangle, \ldots, \langle \boldsymbol{x}, \phi_d\rangle) \in \mathbb{R}^d$.

Given $50,000$ training samples, we construct the empirical covariance operator and project each prior sample $\boldsymbol{X}^i$ onto $d = 25$ leading PCA modes. These modes are sufficient to capture more than 96% of the energy in the first 1000 modes of the spectrum, i.e., $\sum_{i=1}^{25}\lambda_i/(\sum_{i=1}^{1000}\lambda_i) \geq 0.96$. We then construct a MGAN network with $64+25$ inputs and 25 outputs to sample from the conditional distribution of the PCA coefficients given the observations. For each (approximate) posterior sample from the trained network, we use the eigenvectors of C_n to reconstruct the random field a in the full-dimensional space. Figure 5-21 plots the posterior mean and the standard deviation for the permeability field. To avoid any *inverse crime* [44, 104], we generate the data $\boldsymbol{y}^*$ from a mesh that is twice as fine as the mesh used to generate the training data. We observe that the posterior mean from MGAN samples closely match the posterior mean from using the preconditioned Crank-Nicolson MCMC algorithm [45] with a tuned parameter β (to achieve an acceptance rate greater than 0.2) for 10^6 steps. While the standard deviation from MGAN is on a similar order of magnitude to the MCMC results, we believe the under-estimate in uncertainty in the center of the domain is a result of the dimension reduction procedure, which ignores some variance in the log-permeability.

5.7 Discussion and extensions

This chapter introduced several algorithms for likelihood-free Bayesian inference by estimating transport maps that depend on both parameters and data. While the

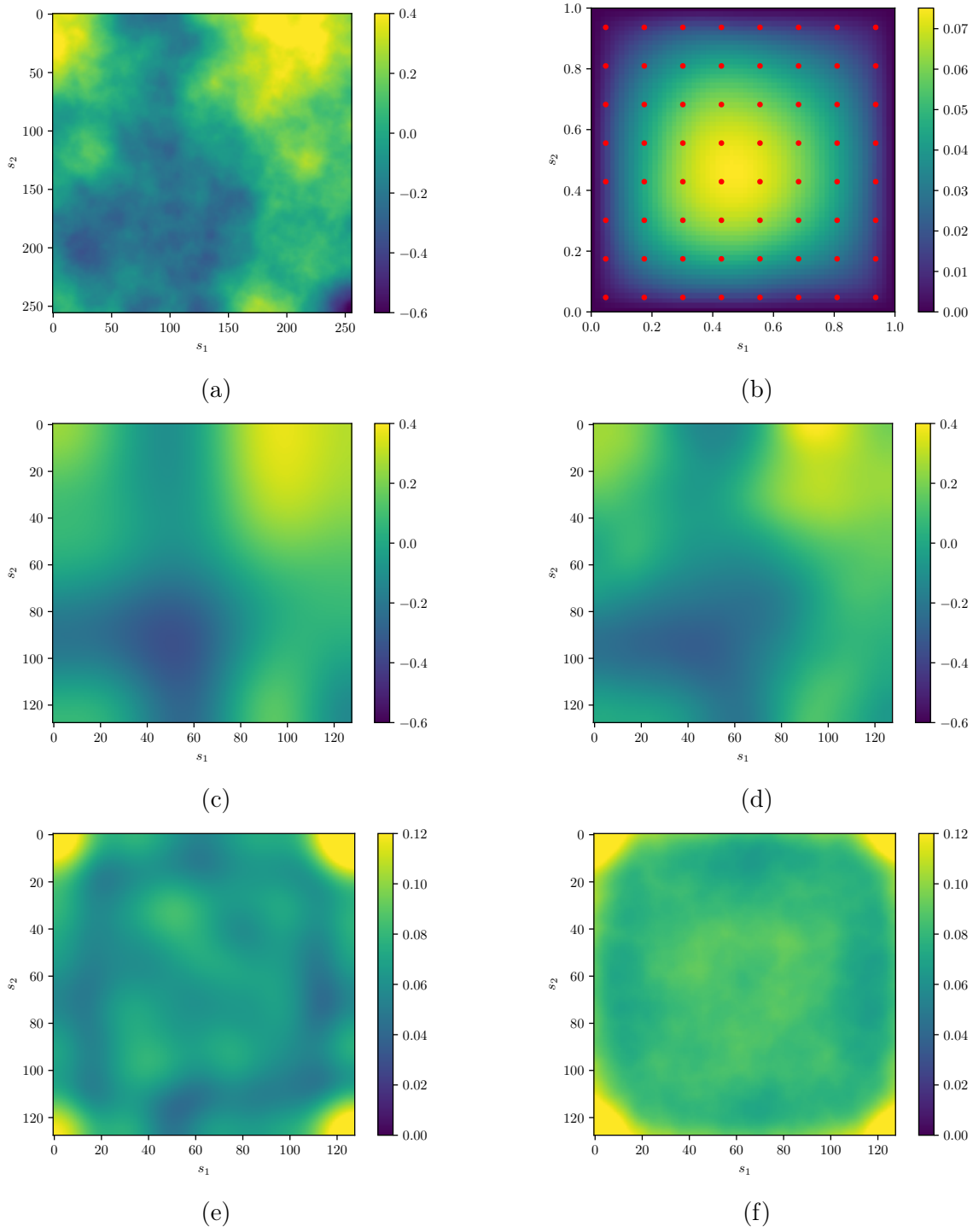


Figure 5-21: A comparison of conditional sampling for the Darcy Flow model using MGAN and MCMC. (a) A realization of the Gaussian random field $\log(a)$. (b) The solution to the PDE with the measurement locations denoted in red. (c) and (d) The mean of the posterior samples from MGAN and MCMC, respectively. (e) and (f) The standard deviation of the posterior samples from MGAN and MCMC, respectively.

measure transport approach in Chapter 3 immediately enables conditional sapling and density estimation for the posterior distributions in LFI problems, this chapter proposed alternative methods to design transport maps for the purpose of conditional sampling. First, we demonstrated that given any estimate of $S^{\mathcal{X}}$, the prior-to-posterior transformation $T_{\boldsymbol{y}^*}$ given by the stochastic map algorithm from composing $S^{\mathcal{X}}$ with its partial inverse yields lower bias and variance than using the inverse of $S^{\mathcal{X}}$ alone. Taking advantage of this construction we proposed an information-theoretic objective function that is tailored for finding $S^{\mathcal{X}}$, when it is used to build $T_{\boldsymbol{y}^*}$. Second, we showed how to consistently extend triangular to block-triangular maps and presented an adversarial optimization algorithm to tractably learn such maps for nonlinear inverse problems. We demonstrated the ability of block-triangular maps to estimate the parameters of coupled ODEs and to infer subsurface permeability from pressure measurements.

We outline a few additional avenues for future work.

Dimension-independent algorithms. Similarly to ABC algorithms, the complexity of finding transport maps that are functions of parameters and observations will depend on the dimensions of these variables. While adaptive algorithms, such as `ATM` from Chapter 3, can alleviate the dimension dependence, it will be interesting to develop algorithms where the map error (i.e., variance and bias of the estimators) is independent of the problem's dimensions. For inverse problems where the data is only informative about a subset of the input variables, one promising path is to identify a low-dimensional subspace of parameters $\mathbf{X}$ that is informed by the data $\mathbf{Y}$. The posterior distribution departs from the prior primarily in the directions that span this subspace, and it is thus sufficient for the map to characterize the posterior change within this subspace. An approach to identify this subspace outside of the likelihood-free setting will be presented in Chapter 7.

Alternative objective functions for measuring independence. So far, we have primarily considered mutual information (MI) as an independence criterion for

finding the transport map $S^{\mathcal{X}}(\mathbf{Y}, \mathbf{X})$ whose output random variable is independent of $\mathbf{Y}$. While the value for the MI is interpretable, it is in general challenging to compute. Hence, it will be interesting to identify other measures for independence that have better computational scalability. For example, the Hilbert-Schmidt independence criterion (HSIC) is a popular metric for assessing independence of two random variables based on the cross-covariance of functions of these variables, when restricting these functions to a reproducing kernel Hilbert space [80].

Validating transport maps for conditional sampling. An important step after constructing maps for arbitrary non-Gaussian distributions is to validate their quality for estimating or sampling conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})$ for any $\boldsymbol{y}$. Given samples from the target distribution, we can estimate various metrics that quantify the accuracy of the pullback densities $\widehat{S}^{\mathcal{X}}(\boldsymbol{y}, \cdot)^\sharp \eta$ on a validation set, e.g., the negative log-likelihood, which is related to the KL divergence from the pullback to the target density (up to an unknown constant). Another practical diagnostic is to compare the reference to the push-forward of the target distribution through the approximate map, i.e., the distribution for the random variable $\mathbf{Z}_2 = \widehat{S}^{\mathcal{X}}(\mathbf{Y}, \mathbf{X})$. If the reference is standard Gaussian, we can compute the empirical quantiles of one-dimensional projections of $\mathbf{Z}_2$ and compare these to the quantiles of a standard Gaussian. While passing these tests is necessary, they do not guarantee that the map accurately models the conditionals uniformly over $\mathbf{Y}$. Hence, it is useful to develop rigorous tests that provide theoretical guarantees on the trustworthiness of approximate transport maps. This is particularly important to detect if a class of transport maps is not sufficiently rich to capture the conditional densities and to provide insights on how to improve such maps. We refer the reader to hypothesis tests and other diagnostics for validating approximate conditional densities in [53, 54, 242]. It would also be interesting to develop rigorous diagnostics on the quality of the composed transport maps presented in this chapter for conditional sampling.

Sampling with parametric flows. An alternative to finding one map $S^{\mathcal{X}}$ is to greedily build a composition of simple maps that push forward samples from the conditionals $\pi_{\mathbf{X}|\mathbf{Y}}$ to the reference density. That is, we seek maps $S_1, \ldots, S_d$ in a class $\mathcal{S}$ such that $(S_1 \circ \cdots \circ S_d)_\sharp \pi_{\mathbf{X}|\mathbf{Y}} = (S_d)_\sharp \cdots (S_1)_\sharp \pi_{\mathbf{X}|\mathbf{Y}} = \eta_{\mathbf{Z}_2}$. When the reference density $\eta_{\mathbf{Z}_2}$ is Gaussian and the maps (S_i) are found simultaneously, this composition is referred to as normalizing flow in the machine learning literature [161]. The advantage of flows over single maps is that we can often consider simple classes of parametric functions $\mathcal{S}$ and produce rich transformations by composition; see the classes $\mathcal{S}$ in [211] given by perturbations of the identity function with respect to $\boldsymbol{x}$. Furthermore, by searching for one triangular transport map S_i at a time, we can preserve the nice properties of the map optimization problem derived in Chapter 3. See [210] for a related greedy construction of nonparametric maps based on optimal transport.

Chapter 6

Data assimilation via couplings

6.1 Introduction

This chapter focuses on probabilistically estimating the states of dynamical systems. This procedure is known as sequential Bayesian inference, and is commonly referred to as *data assimilation* (DA) in geophysical problems. DA is a ubiquitous and enormously important part of many geophysical applications, such as operational aspects of numerical weather prediction (NWP) [10]. The principles used in DA, however, are applicable in any field that combines predictive models with partial and noisy observations to make state predictions, including finance, and epidemiology.

In many application areas, characterizing the probability distributions arising in DA is challenging. This is particularly true when (i) the states are very high-dimensional, (ii) the dynamics are nonlinear and complex, and (iii) the observations are sparse in space and time [137]. As an example, NWP tracks properties of the atmosphere including temperature, wind, and humidity on a fine grid at various locations. This results in state dimensions of $\mathcal{O}(10^9)$ [35]. This high dimensionality renders many traditional sampling-based sequential inference methods, such as sequential Monte Carlo (SMC), infeasible for these systems.

Nevertheless, research driven to a large extent by practitioners, has lead to the development of many algorithmic innovations to enable operational data assimilation. One of the primary approaches in the geosciences for online filtering, i.e., estimating

the current state over time, is the ensemble Kalman filter (EnKF). Similarly to SMC, the EnKF provides an empirical approximation to the distribution of interest using a collection of samples, that is known as an ensemble. Instead of tracking sample weights and resampling as in SMC, however, the EnKF updates the ensemble by constructing affine transformations to move the samples at each assimilation step [65]. The restriction of the EnKF to affine transformations results in a robust approach for high-dimensional problems when using small ensemble sizes.

Despite its robustness, the EnKF does not yield consistent estimates in non-Gaussian systems. The transformation used in the EnKF is derived under Gaussian assumptions, and hence the EnKF is intrinsically biased. This bias means that the accuracy of the EnKF is fundamentally limited and the performance of this algorithm does not improve when increasing the ensemble size beyond a certain limit. As increasing computational resources enable data assimilation studies with larger ensembles [149], it becomes necessary to address this limitation. Recently, several algorithms have been proposed to obtain more consistent approximations to the filtering distribution; see Sections 6.2.1 and 6.2.2. Many of these ensemble-based approaches, however, still suffer from particle degeneracy by relying on importance sampling, or from using nonlinear transformations that are inexact in the large-sample limit.

In this chapter, we present a generalization of the EnKF to nonlinear transformations that we call the stochastic map filter. This algorithm builds upon the stochastic map algorithm introduced in Section 5.4 for posterior sampling. At each filtering step, the algorithm builds a prior-to-posterior transformation that pushes forward samples from the forecast (i.e., the prior) distribution to samples from the filtering (i.e., the posterior) distribution. Unlike the EnkF, these transformations are, in principle, consistent for posterior sampling in the general non-Gaussian setting. In practice, we construct parametric estimators for these transformations that trade-off bias for variance by gradually enriching the map's complexity with more samples. Using nonlinear prior-to-posterior transformations, we show that the stochastic map filter outperforms the biased EnKF for tracking the states of chaotic dynamical systems.

The remainder of this chapter is organized as follows. We introduce the data

assimilation problem and existing methods in Section 6.2. The stochastic map filtering algorithm is introduced in Section 6.3 and its performance is exhibited on the Lorenz-63 system in Section 6.3.3. We show how to exploit approximate conditional independence to extend the algorithm to high-dimensional problems using sparse maps in Section 6.4. Lastly, we discuss extensions to smoothing and other DA problems in Section 6.5.

6.2 Problem Setup

To formulate the sequential inference problem, we consider a (possibly non-Gaussian) state-space model for the time dependent state of a dynamical system that is not directly observable. We model the latent stochastic process as a discrete-time Markov chain $(\mathbf{X}_t)_{t\geq 0}$ taking values in $\mathbb{R}^d$ that evolves in time according to the transition kernels $\pi_{\mathbf{X}_{t+1}|\mathbf{X}_t}$. We model the noisy observations of the state $(\mathbf{Y}_t)$ as an observed process taking values in $\mathbb{R}^m$ that are conditionally independent given the states at all times. The observations are sampled from the likelihood functions $\pi_{\mathbf{Y}_t|\mathbf{X}_t}$. In this chapter, we assume the observations have the form $\mathbf{Y}_t = h(\mathbf{X}_t, \boldsymbol{\mathcal{E}}_t)$ where h is a (possibly nonlinear) observation operator. This operator is a function of the state and a "noise" random variable $\boldsymbol{\mathcal{E}}_t$ that is independent of $\mathbf{X}_t$.

Together with the distribution for the initial condition of the latent process $\mathbf{X}_0$, the transition kernels and likelihood model describe the joint law of the Markov process $(\mathbf{X}_t, \mathbf{Y}_t)_{t\geq 0}$. The model for the state and observation variables at discrete times $t \in \mathbb{N}$ is known as a *hidden Markov model* and its dependencies are illustrated by the graphical model in Figure 6-1.

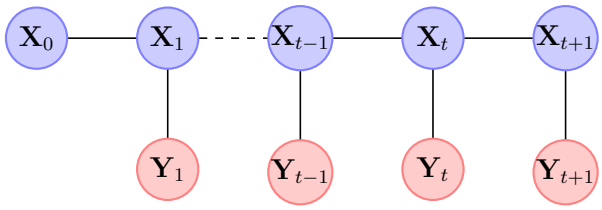


Figure 6-1: Markov structure for a latent process $\mathbf{X}_t$ and an observed process $\mathbf{Y}_t$ for $t \in \mathbb{N}$

Given a sequence of realizations $(\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_t^*)$ of the observed process, our goal is to infer the distributions for a marginal of the hidden state at time k conditioned on the observations, i.e., $\pi_{\mathbf{X}_k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}(\cdot) \coloneqq \pi_{\mathbf{X}_k|\mathbf{Y}_1,\ldots,\mathbf{Y}_t}(\cdot|\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_t^*)$. This task can be categorized into three problems: (i) *filtering* if $k = t$, (ii) *smoothing* if $k < t$, and (iii) *prediction* if $k > t$.

In an *online* setting, we are often interested in assimilating data as soon as it becomes available and updating our current estimate for the state given all collected measurements. For instance, we might be interested in estimating atmospheric variables in a meteorological model given real-time weather station measurements, or in tracking the current position of a satellite given GPS measurements. In this chapter, we focus on the filtering problem for nonlinear dynamical systems.

Using the Markov properties of the state-space model, the filtering distribution $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}$ can be computed sequentially as an update of the filtering distribution at time t given by $\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$. This update is defined by two-steps: *forecast* which predicts the state at the time $t+1$ conditioned on observations up to time t using the forecast distribution $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$, and *analysis* which assimilates the data at time $t+1$. In the forecast step, the transition kernel is used to update the previous filtering distribution using

$$\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}(\boldsymbol{x}_{t+1}) = \int_{\mathbb{R}^d} \pi_{\mathbf{X}_{t+1}|\mathbf{X}_t}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}(\boldsymbol{x}_t)\mathrm{d}\mathbf{x}_t. \tag{6.1}$$

In the analysis step, the latest measurement $\boldsymbol{y}_{t+1}^*$ is assimilated using Bayes' formula to yield the updated filtering distribution

$$\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}(\boldsymbol{x}_{t+1}) = \frac{\pi_{\boldsymbol{y}_{t+1}|\mathbf{X}_{t+1}}(\boldsymbol{x}_{t+1})\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}(\boldsymbol{x}_{t+1})}{\pi_{\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}}, \tag{6.2}$$

where $\pi_{\boldsymbol{y}_{t+1}|\mathbf{X}_{t+1}}(\boldsymbol{x}_{t+1}) \coloneqq \pi_{\mathbf{Y}_{t+1}|\mathbf{X}_{t+1}}(\boldsymbol{y}_{t+1}^*|\boldsymbol{x}_{t+1})$. The composition of the forecast and analysis steps provides a mapping of the filtering distribution from time t to time $t+1$. The following example is a setting where this mapping is available in closed-form.

Example 5 (Linear Gaussian Model)**.** *Let the latent process* $\mathbf{X}_t \in \mathbb{R}^d$ *follow the linear*

stochastic difference equation

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \boldsymbol{\mathcal{W}}_t, \tag{6.3}$$

where $\boldsymbol{\mathcal{W}}_t \in \mathbb{R}^d$ *is an independent Gaussian random variable with distribution* $\mathcal{N}(\mathbf{0}, \mathbf{Q})$*, and the initial condition is also drawn from a Gaussian distribution, i.e.,* $\mathbf{X}_0 \sim \mathcal{N}(\mathbf{m}_0, \boldsymbol{\Sigma}_0)$*. If the observation operator is a linear function of the state with additive noise, i.e.,* $\mathbf{Y}_t = h(\mathbf{X}_t, \boldsymbol{\mathcal{E}}_t) = \mathbf{H}\mathbf{X}_t + \boldsymbol{\mathcal{E}}_t$ *where* $\boldsymbol{\mathcal{E}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$*, the state* $\mathbf{X}_t$ *(before and after conditioning on any observations) follows a Gaussian distribution for all time* t*. In particular, if* $\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}(\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_t; \mathbf{m}_t, \boldsymbol{\Sigma}_t)$*, the mean and covariance of the filtering distribution at time* $t+1$*, following the forecast and analysis steps above, are given by*

$$\begin{aligned} \boldsymbol{\Sigma}_{t+1}^{-1} &= (\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T + \mathbf{Q})^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} \\ \boldsymbol{\Sigma}_{t+1}^{-1}\mathbf{m}_{t+1} &= (\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T + \mathbf{Q})^{-1}\mathbf{A}\mathbf{m}_t + \mathbf{H}^T\mathbf{R}^{-1}\boldsymbol{y}_{t+1}^*. \end{aligned} \tag{6.4}$$

These updates for the mean and covariance are known as the Kalman Filter, and their derivation can be found in [101].

If the model dynamics and observation operators are nonlinear or the additive noise is non-Gaussian, the update to the filtering distribution does not have a closed-form expression in terms of finite-dimensional parameters. For nonlinear models, the extended Kalman filter linearizes the model about the current mean and applies the recursion in (6.4). However, for geophysical applications that deviate strongly from linear Gaussian models, this approximate filter can easily fail to estimate the state. Instead, two algorithms that can approximate non-Gaussian distributions in these systems are the ensemble Kalman filter and the particle filter, which are described in Sections 6.2.1 and 6.2.2, respectively. We refer the reader to [120, 10] for other filtering algorithms that are used in the data assimilation community.

6.2.1 Ensemble Kalman filter

The *ensemble Kalman filter* (EnKF) approximates the filtering distribution by a set of n particles $\{\boldsymbol{X}_t^i\}_{i=1}^n$, known as an ensemble. These particles are updated in each fil-

tering step by first sampling from the Markov transition kernel $\overline{\boldsymbol{X}}^i_{t+1} \sim \pi_{\mathbf{X}_{t+1}|\mathbf{X}_t}(\cdot|\boldsymbol{X}^i_t)$ to generate an empirical approximation to the forecast distribution (6.1), and second by applying a linear transformation to perform the analysis step. The analysis transformation is derived by approximating the forecast distribution $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$ with a multivariate Gaussian that has the sample mean and covariance

$$\overline{\mathbf{m}}_{t+1} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{X}}^i_{t+1},$$
$$\overline{\boldsymbol{\Sigma}}_{t+1} \coloneqq \frac{1}{n-1}\sum_{i=1}^{n} (\overline{\boldsymbol{X}}^i_{t+1} - \overline{\mathbf{m}}_{t+1})(\overline{\boldsymbol{X}}^i_{t+1} - \overline{\mathbf{m}}_{t+1})^T.$$

For a Gaussian forecast (i.e., prior) and linear-Gaussian model for the observations (i.e., $\mathbf{Y}_{t+1} = \mathbf{H}\mathbf{X}_{t+1} + \boldsymbol{\mathcal{E}}_{t+1}$ for $\boldsymbol{\mathcal{E}}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$), the posterior distribution is also Gaussian with the mean $\mathbf{m}_{t+1}$ and covariance $\boldsymbol{\Sigma}_{t+1}$ given in (6.4). In this case, an update formula can be derived and applied to each member of the forecast ensemble $\{\overline{\boldsymbol{X}}^i_{t+1}\}_{i=1}^n$ to generate the analysis ensemble $\{\boldsymbol{X}^i_{t+1}\}_{i=1}^n$. This update is chosen so that the analysis samples have the desired posterior mean $\mathbf{m}_{t+1}$ and covariance $\boldsymbol{\Sigma}_{t+1}$ under the Gaussian approximation to the forecast distribution. The "perturbed observation" form of the EnKF update is given by

$$\begin{aligned} \boldsymbol{Y}^i_{t+1} &= \mathbf{H}\overline{\boldsymbol{X}}^i_{t+1} + \boldsymbol{\mathcal{E}}^i_{t+1}, \quad \boldsymbol{\mathcal{E}}^i_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \\ \boldsymbol{X}^i_{t+1} &= \overline{\boldsymbol{X}}^i_{t+1} - \mathbf{K}_{t+1}(\boldsymbol{Y}^i_{t+1} - \boldsymbol{y}^*_{t+1}) \end{aligned} \tag{6.5}$$

for $i = 1, \ldots, n$, where $\mathbf{K}_{t+1} \coloneqq \overline{\boldsymbol{\Sigma}}_{t+1}\mathbf{H}^T(\mathbf{H}\overline{\boldsymbol{\Sigma}}_{t+1}\mathbf{H}^T + \mathbf{R})^{-1} \in \mathbb{R}^{d\times m}$ is known as the Kalman Gain. The analysis ensemble is used to approximate the filtering distribution at time $t+1$, and these steps are repeated at each assimilation cycle.

Since the development of the EnKF algorithm by Evensen [65], several variants have been proposed. These primarily include deterministic square-root EnKFs that avoid sampling the observational noise $\boldsymbol{\mathcal{E}}_{t+1}$ in (6.5) to reduce sampling error [24, 6, 229], and related optimization-based methods for deriving the forecast-to-analysis transformation [158]. These EnKF algorithms have been successfully applied and are

used operationally in various domains including atmospheric modeling [105], oceanography [18], reservoir engineering [1], and wildfire monitoring [139].

In practice, there are several modifications of the EnKF that enable it to be applied on high-dimensional systems with ensemble sizes $n \ll d$. Two primary statistical regularization techniques are *inflation* and *localization*. Inflation replaces the forecast sample covariance by using the map $\overline{\mathbf{\Sigma}}_{t+1} \mapsto (1 + \lambda_1)\overline{\mathbf{\Sigma}}_{t+1} + \lambda_2 \mathbf{I}_d$ with two scalar parameters $\lambda_1, \lambda_2 > 0$ to increase the rank and to account for possible model errors in the dynamics. A common implementation of localization applies an elementwise mask to the forecast covariance to reduce spurious correlations between states that are separated by large physical distances, i.e., $\overline{\mathbf{\Sigma}}_{t+1} \mapsto \rho \circ \overline{\mathbf{\Sigma}}_{t+1}$ where ρ is a Gaspari-Cohn tapering function. This localization technique relies on the assumption of approximate marginal independence between spatially separated states, at least on short time scales. Another form of localization assimilates local observations by only updating nearby state variables to the observation location [10].

Despite its empirical success, the EnKF can be seen as a Monte Carlo approximation of the Kalman filter and is inconsistent for capturing the filtering distribution in non-Gaussian state-space models. With increasing sample sizes $n \to \infty$, the empirical approximations from the EnKF only converge to the true filtering distributions in linear Gaussian systems [122, 121]. To address this, several algorithms have been proposed to improve the EnKF for non-Gaussian forecast distributions including the rank histogram filter [5], the moment matching filter [125], and Gaussian mixture approximations [71]. Lastly, an approach based on couplings, known as the ensemble transform particle filter [179], uses likelihood weights to derive a discrete optimal transport map that pushes the forecast to the analysis distribution. Another measure transport approach that generalizes the EnKF is presented in Section 6.3.

6.2.2 Particle filter

Consistent approaches to estimate non-Gaussian filtering distributions are *sequential Monte Carlo* algorithms, often referred to as *particle filters* (PF) [57]. Particle filters use an empirical measure to approximate the filtering distribution as

$\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t} \approx \frac{1}{n}\sum_{i=1}^n w_t^i \delta_{\boldsymbol{X}_t^i}$ where $w_t^i \in \mathbb{R}_{>0}$ are normalized particle weights that sum to one. Similarly to the EnKF, in each forecast step the particle filter predicts the state at time $t+1$ by sampling from the Markov transition kernel $\overline{\boldsymbol{X}}_{t+1}^i \sim \pi_{\mathbf{X}_{t+1}|\mathbf{X}_t}(\cdot|\boldsymbol{X}_t^i)$ and keeping the same weights for each particle.

In the analysis step, the particle filter applies Bayes' formula in (6.2) to obtain the empirical approximation

$$\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}} \approx \frac{1}{n}\sum_{i=1}^n w_{t+1}^i \delta_{\boldsymbol{X}_{t+1}^i}, \tag{6.6}$$

where the normalized weights are given by $w_{t+1}^i = \widetilde{w}_{t+1}^i/(\sum_{i=1}^n \widetilde{w}_{t+1}^i)$ for $\widetilde{w}_{t+1}^i = \pi_{\boldsymbol{y}_{t+1}^*|\boldsymbol{X}_{t+1}^i} w_t^i$. Unlike the EnKF, however, the analysis step of the particle filter does not change the positions of the forecast particles. The analysis step only reweighs each particle according to the likelihood of generating the observation $\boldsymbol{y}_{t+1}^*$. After some assimilation steps, this algorithm suffers from a phenomenon known as particle degeneracy where all, but one, of the weights approach 0. This results in an empirical approximation with an effective sample size of 1, that in practice fails to track the true state of the system.

To alleviate particle degeneracy, it is common to introduce a resampling step [79]. Following each analysis step, the algorithm samples a set of n particles from the weighted empirical distribution (6.6). This has the effect of introducing more particles with higher likelihood weights and dropping particles with low weights. This algorithm is known as *sequential importance resampling* (SIR) or the *bootstrap particle filter*. The consistency of SIR for sampling from the filtering distribution in the large-particle limit is proven in Theorem 4.5 of [120].

Despite its consistency for non-Gaussian problems, the bootstrap particle filter is not yet a practical algorithm for high-dimensional systems. Even with re-sampling, the filter typically degenerates over time when one particle dominates the approximation of $\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$ by having a weight near 1. To reduce the variance of the weights, [200] showed that the number of samples n needs to increase exponentially with the state dimension d due to the near mutual singularity of the forecast distri-

bution and the likelihood.

Several improvements have been proposed to apply the particle filter in geophysical applications. These include using proposal distributions that incorporate the observation to generate samples near high probability regions of the filtering distribution [41], and the introduction of stochastic perturbations to increase particle diversity [58]. Although these methods improve the effective number of samples in PF approximations, they do not reduce the exponential scaling of the ensemble size n with the state dimension [201]. This unavoidable scaling is a result of the importance sampling step that is used to compute the particle weights. On the other hand, non-consistent particle filters that alleviate the exponential scaling include: the equal-weight particle filter [244], hybrid EnKF and PF methods that confine the PF to small subspaces of the state [198], and the introduction of localization techniques that are similar in spirit to the ones used in the EnKF [177, 169].

Nevertheless, an active research direction in data assimilation is to develop general algorithms that maintain the robustness of the EnKF for high-dimensional systems and avoid the degeneracy issues of the PF that result from importance sampling.

6.3 Stochastic map filtering algorithm

The goal of every assimilation cycle in a sequential Bayesian inference problem is to transform samples for the state at time t into samples conditioned on a new observation at time $t+1$. In ensemble filtering, this consists of propagating the samples according to the dynamical model (in the *forecast step*), followed by solving a Bayesian inference problem where prior samples are updated into posterior samples (in the *analysis step*) as seen in Figure 6-2. In this section we propose an approach to carry out the Bayesian inference problem in the analysis step using measure transport.

The analysis step of each assimilation cycle can be seen as a "static" Bayesian inference problem where the prior $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\dots,\boldsymbol{y}_t}$ is represented by the forecast ensemble $\{\overline{\boldsymbol{X}}^i_{t+1}\}_{i=1}^n$, the likelihood function is prescribed by the observation model $\pi_{\mathbf{Y}_{t+1}|\mathbf{X}_{t+1}}$, and our goal is to sample from the posterior $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\dots,\boldsymbol{y}_{t+1}}$ given a new observation

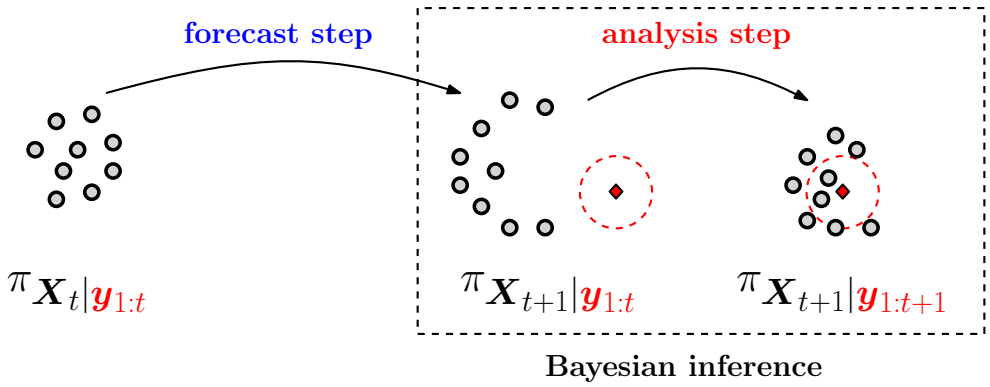


Figure 6-2: Typical ensemble filtering algorithms consist of two steps: forecast and analysis

$\boldsymbol{y}^*_{t+1}$. This is the setting of the stochastic map algorithm presented in Section 5.4.

To generate approximate samples from the filtering distribution at each assimilation cycle, we propose to follow the steps in Algorithm 5. For each forecast sample, we generate a sample from the likelihood model $\boldsymbol{Y}^i_{t+1} \sim \pi_{\mathbf{Y}_{t+1}|\mathbf{X}_{t+1}}(\cdot|\overline{\boldsymbol{X}}^i_{t+1})$ and use joint samples $\{(\boldsymbol{Y}^i_{t+1}, \overline{\boldsymbol{X}}^i_{t+1})\}_{i=1}^n \sim \pi_{\mathbf{Y}_{t+1},\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$ to estimate the block $S^{\mathcal{X}}$ of a triangular and monotone map that pushes forward the joint distribution to a standard Gaussian reference distribution with density η. The estimated block $\widehat{S}^{\mathcal{X}}$ is composed with its partial inverse to construct the prior-to-posterior transformation $\widehat{T}_{\boldsymbol{y}^*_{t+1}}$ in (5.17). To approximately sample from the filtering distribution at time $t+1$, we evaluate the transformation $\widehat{T}_{\boldsymbol{y}^*_{t+1}}$ at the joint samples that are used to learn $\widehat{S}^{\mathcal{X}}$. This procedure closely follows the EnKF (see Section 6.2.1), which computes the Kalman Gain in (6.5) using joint samples and applies an affine transformation to the same joint samples to generate the analysis ensemble. The main distinction with the EnKF is that the transformation $\widehat{T}_{\boldsymbol{y}^*_{t+1}}$ applied to the joint samples may be non-linear. We denote this generalization of the EnKF to nonlinear transformations as the *stochastic map filter*. The procedure for each assimilation cycle is summarized in Algorithm 7.

Algorithm 7: Stochastic map (SM) filtering algorithm

Input : Analysis samples $\boldsymbol{X}_t^i \sim \pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$ at time t, observation $\boldsymbol{y}_{t+1}^*$

Output: Analysis samples $\boldsymbol{X}_{t+1}^i$ at time $t+1$

1 Apply dynamics to generate forecast ensemble $\overline{\boldsymbol{X}}_{t+1}^i \sim \pi_{\mathbf{X}_t|\boldsymbol{X}_t^i}$ for $i = 1, \ldots, n$

2 Sample observations $\boldsymbol{Y}_{t+1}^i \sim \pi_{\mathbf{Y}_{t+1}|\overline{\boldsymbol{X}}_{t+1}^i}$ using forecast samples

3 Estimate block $S^{\mathcal{X}}$ of lower-triangular map that couples $\pi_{\mathbf{X}_{t+1},\mathbf{Y}_{t+1}}$ to η

4 Evaluate composed map to generate analysis ensemble
$\boldsymbol{X}_{t+1}^i = \widehat{T}_{\boldsymbol{y}_{t+1}^*}(\boldsymbol{Y}_{t+1}^i, \overline{\boldsymbol{X}}_{t+1}^i)$ (see Algorithm 5)

6.3.1 Connection with the EnKF

When constraining the estimator for $S^{\mathcal{X}}$ to a class of affine functions, the optimal estimator according to (5.4) has the form

$$\widehat{S}^{\mathcal{X}}(\boldsymbol{y}_t, \boldsymbol{x}_t) = \mathbf{A}(\boldsymbol{x}_t - \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t,\mathbf{Y}_t}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}^{-1}\boldsymbol{y}_t), \tag{6.7}$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is the inverse of the Cholesky factor of the conditional covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t|\mathbf{Y}_t} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t,\mathbf{Y}_t}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_t}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t,\mathbf{Y}_t}^T$ for $(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1}) \sim \pi_{\boldsymbol{x}_{t+1},\boldsymbol{y}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}$, and where the carets denote sample covariances. We refer the reader to Section 5.4.2 for a discussion on estimating the map in (6.7) for multivariate Gaussian target densities. For an observation $\boldsymbol{y}_t^*$, the prior-to-posterior transformation $\widehat{T}_{\boldsymbol{y}_t^*}$ is then given by

$$\widehat{T}_{\boldsymbol{y}_t^*}(\boldsymbol{y}_t, \boldsymbol{x}_t) = \boldsymbol{x}_t - \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t,\mathbf{Y}_t}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_t}^{-1}(\boldsymbol{y}_t - \boldsymbol{y}_t^*). \tag{6.8}$$

This transformation has the same form as transformation (6.5) applied by the ensemble Kalman filter to each forecast sample where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t,\mathbf{Y}_t}\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_t}^{-1}$ corresponds to the Kalman gain. For a linear-Gaussian likelihood model (see Example 5) given by the observation $\mathbf{Y}_t = \mathbf{H}\mathbf{X}_t + \boldsymbol{\mathcal{E}}_t$ for $\boldsymbol{\mathcal{E}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, the covariances in (6.8) are given by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t,\mathbf{Y}_t} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t}\mathbf{H}^T$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}} = \mathbf{H}\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}_t}\mathbf{H}^T + \mathbf{R}$. Using these expressions, we recover the form of the perturbed observation EnKF presented in Section 6.2.1. Thus, the SM filter reverts to the EnKF for affine transport maps $S^{\mathcal{X}}$. By seeking nonlinear functions $S^{\mathcal{X}}$, the SM filter offers a framework to generalize the linear *ansatz* of the

EnKF to more closely capture non-Gaussian features in the filtering distributions.

6.3.2 Processing observations incrementally

While the stochastic map filter can condition the forecast distribution on any observation $\boldsymbol{y}_t^* = (y_{t,1}^*, \ldots, y_{t,m}^*)$, in this section we propose a sequential approach for processing observations that are conditionally independent given the state. This means that the likelihood model factorizes into m components as

$$\pi_{\mathbf{Y}_t|\mathbf{X}_t} = \prod_{k=1}^{m} \pi_{Y_{t,k}|\mathbf{X}_t}. \tag{6.9}$$

For example, the conditional independence in (6.9) is present when each observation corresponds to point-wise measurement of one element of the state vector $\mathbf{X}_t$ with an independent additive error. If the likelihood model factorizes into d terms, we can sequentially iterate Algorithm 7 to process each scalar observation $y_{t,k}^*$. In this case, the analysis ensemble associated with $y_{t,k}^*$ is used as the prior ensemble when assimilating $y_{t,k+1}^*$.

Processing observations incrementally has two computational advantages. First, instead of computing a single prior-to-posterior transformation $T_{\boldsymbol{y}_t^*} \colon \mathbb{R}^{m+d} \to \mathbb{R}^d$ of dimension $d+m$, the analysis step is performed by computing m prior-to-posterior transformations $\widehat{T}_{y_{t,k}^*} \colon \mathbb{R}^{1+d} \to \mathbb{R}^d$ for $k = 1, \ldots, m$ with lower input dimension $d+1$. In the context of the EnKF, this avoid the storage and inversion of matrices (such as $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}_t}$), that will be expensive for large m [92]. Second, processing observations serially enables us to more easily take advantage of sparsity in $S^{\mathcal{X}}$ arising from conditional independence structure; see Theorem 4.4.1 in Chapter 4 for a relation between the Markov properties of a target distribution and the sparsity of a corresponding lower triangular transport map. For instance, if observation $Y_{t,k}$ is only a function of a subset $\mathcal{K} \subset \{1, \ldots, d\}$ of the state variables $\mathbf{X}$, i.e., $Y_{t,k} \perp\!\!\!\perp \mathbf{X}_{t,\mathcal{K}^c} \,|\, \mathbf{X}_{t,\mathcal{K}}$ where $\mathcal{K}^c =$

$\{1, \ldots, d\} \setminus \mathcal{K}$, then the lower triangular map of the form

$$S^{\mathcal{X}}(y_{t,k}, \boldsymbol{x}_t) = \begin{bmatrix} S_{\mathcal{K}}(y_{t,k}, \boldsymbol{x}_{t,\mathcal{K}}) \\ S_{\mathcal{K}^c}(\boldsymbol{x}_{t,\mathcal{K}}, \boldsymbol{x}_{t,\mathcal{K}^c}) \end{bmatrix}, \tag{6.10}$$

will exactly represent the conditional distribution of interest.

For data assimilation problems that use point-wise measurements of the state (as in Sections 6.3.3 and 6.4.1), $\mathcal{K}$ is of size 1 for each observation, i.e., each observation on depends on one scalar state variable. In this case, after a permutation of the state variables in each prior-to-posterior map $T_{y^*_{t,k}}$, only the first component of the map depends on $y_{t,k}$ and the remaining components only depend on $\boldsymbol{x}_t$. Let us remark that evaluating a map with this form can be interpreted as first updating the observed state element $\mathbf{X}_{t,\mathcal{K}}$ and second propagating the information from the observation to the remaining states $\mathbf{X}_{t,\mathcal{K}^c}$ according on the conditional dependence of $\mathbf{X}_{t,\mathcal{K}}$ alone. In our numerical experiments in the following sections, we sequentially process observations from the local likelihood model in (6.9) and take advantage of the resulting sparse structure in the maps.

6.3.3 Lorenz-63 dynamical system

In this section we present the performance of the stochastic map filter on the Lorenz-63 system. The Lorenz-63 model was introduced in [133] and describes the natural convection of a heated-fluid. The state at time s is a three-dimensional vector $\mathbf{X}(s) = (X_1(s), X_2(s), X_3(s))$ whose dynamic is described by the coupled ODEs

$$\begin{aligned} \frac{dX_1}{ds} &= -\sigma X_1 + \sigma X_2 \\ \frac{dX_2}{ds} &= -X_1 X_3 + \rho X_1 - X_2 \\ \frac{dX_3}{ds} &= X_1 X_2 - \beta X_3, \end{aligned} \tag{6.11}$$

where the parameters $\beta = 8/3, \rho = 28, \sigma = 10$ are set to make the dynamics chaotic. We integrate the ODEs with a fourth-order explicit Runge-Kutta method with a

step-size of $\Delta s = 0.05$ and observe the full state every $\Delta s_{\text{obs}} = 0.1$ time units. We index the state at discrete observation times $t \in \mathbb{N}$ where $\mathbf{X}_t = \mathbf{X}(t\Delta s_{\text{obs}})$. For any observation step t, the likelihood model is specified by

$$\mathbf{Y}_t = \mathbf{X}_t + \boldsymbol{\mathcal{E}}_t, \tag{6.12}$$

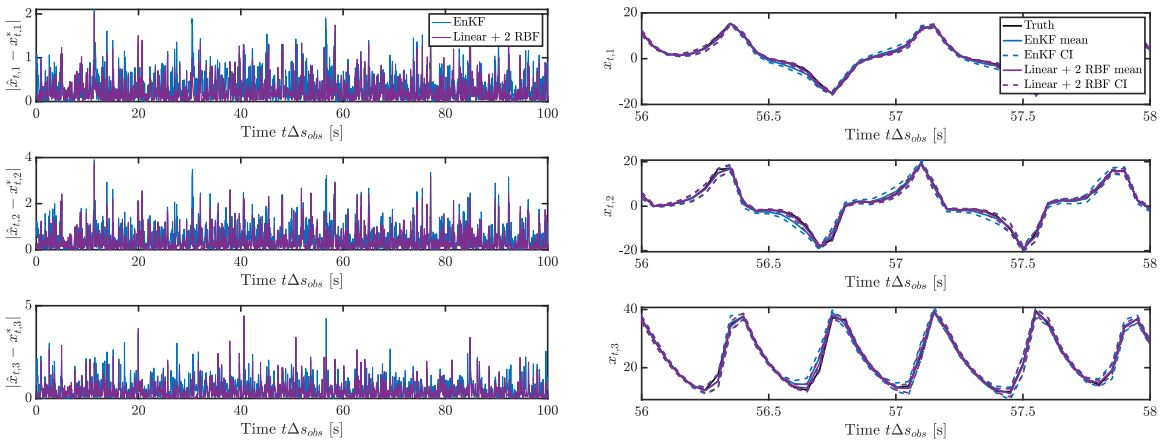
where $\boldsymbol{\mathcal{E}}_t \sim \mathcal{N}(\mathbf{0}, 4\mathbf{I}_3)$ is drawn independently of $\mathbf{X}_t$.

The filtering setup we follow is an "identical twin experiment". Given a random initial condition $\mathbf{X}_0$, we generate a sequence of true hidden states $(\boldsymbol{x}_t^*)$ by integrating the ODEs and a sequence of synthetic observations $(\boldsymbol{y}_t^*)$. We then use these observations to recover the hidden states using the same forecast model. In all of the numerical results, we generate an initial ensemble through a spin-up phase (using a perturbed observation EnKF). The filtering algorithms are run for 4000 assimilation steps. At each step, we compute the ensemble mean $\widehat{\boldsymbol{X}}_t = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_t^i$ as a point estimate of the true state, and the ensemble covariance $\widehat{\boldsymbol{\Sigma}}_t$. We evaluate the performance of the filter for tracking the hidden state via the root-mean-squared error (RMSE), which is defined at assimilation step t as $\text{RMSE}_t := \|\widehat{\boldsymbol{X}}_t - \boldsymbol{x}_t^*\|_2/\sqrt{d}$. The statistics of the filter's performance is based on the average RMSE over the last 2000 assimilation steps.

In this study, we consider a parameterization for triangular transport maps that gradually departs from a linear representation in order to demonstrate the benefit of nonlinear prior-to-posterior transformations $T_{\boldsymbol{y}_t^*}$. In particular, we consider maps $S^{\mathcal{X}}$ where each component has a separable representation of the form $S_k^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}_{1:k}) = u_k^0(\boldsymbol{y}) + \sum_{i=1}^k u_k^i(\boldsymbol{x}_i)$ for $k = 1, \ldots, d$. For the linear version of the stochastic map filter (labeled as "Linear" in the figures), the univariate functions u_k^0 and u_k^i are restricted to be linear functions of their inputs. To increase the complexity of the map, we add radial basis functions (RBFs) to the approximation spaces for u_k^0, u_k^i. We use a monotone parameterization based on sigmoids (i.e., integrals of RBFs) for u_1^1 and keep u_k^k linear for $k > 1$. We label the results for maps with linear terms plus p RBFs/sigmoids in each univariate function as "Linear + p RBF". In all results, we

also add inflation to the forecast by tuning the spread of the samples.

Figure 6-3a displays the absolute error in the prediction of each state element over time for the EnKF and the transport maps with Linear + 2 RBFs when using an ensemble of size $n = 600$. Figure 6-3b displays the ensemble mean and a 95% confidence interval for the mean at times $s \in [56, 58]$. We observe that the nonlinear stochastic map filter yields an ensemble mean that is less biased relative to the true state $\boldsymbol{x}_t^*$. For instance, the estimates of states X_1 and X_2 at time $s = 56.6$ using "Linear + 2 RBF" are closer to the truth $\mathbf{x}_t^*$. In comparison, the truth falls outside the confidence interval for the EnKF ensemble at this time. This is also reflected in higher values for the average RMSE over time: 0.51 ± 0.02 for the EnKF (i.e., linear maps) as compared to 0.36 ± 0.02 for the map with linear terms and 2 RBFs (i.e., nonlinear transformations).



(a) (b)

Figure 6-3: (a) Absolute error in the ensemble mean prediction of each state element for the Lorenz-63 model with $\Delta s_{\text{obs}} = 0.1$ and $n = 600$ samples. (b) The ensemble mean of two filters with their 95% confidence interval for predicting the true hidden state $\boldsymbol{x}_t^*$.

Figure 6-4 plots the RMSE with increasing ensemble sizes $n \in [10, 600]$. For small n, the EnKF and linear maps have the most robust performance. However, with an increasing ensemble size, we observe an improvement in tracking by gradually adding nonlinearities to the map's approximation space. Specifically, the map with linear terms and 1 RBF has better tracking performance for $n \geq 40$, and the map with

linear terms and 2 RBFs has the lowest RMSE for $n \geq 200$. The minimum RMSE also tends towards the RMSE of a "reference" solution obtained with a consistent SIR particle filter (see Section 6.2.2) using $n = 10^6$ samples. Let us remark that the RMSE of the EnKF closely matches that of a linear map, as discussed in Section 6.3.1. The difference between the stochastic EnKF and the linear map results for small n can be attributed to the local likelihood structure that we exploit; see Section 6.3.2. In particular, for scalar observations that depends on one component of the state, only the first component of each lower triangular map (6.10) depends on the observation.

The improvement in RMSE arising from nonlinear maps is nearly constant when increasing the time Δs_{obs} between observations. The right of Figure 6-4 plots the average RMSE of various algorithms for $\Delta s_{\text{obs}} \in [0.1, 0.5]$ with a fixed ensemble size of $n = 1000$. This result demonstrates that the stochastic map filter remains stable with increasing map complexity and continues to provide accurate tracking even as the state evolution becomes more nonlinear. We remark that this behavior is not expected with particle filters that often suffer from particle degeneracy when increasing the inter-observation time Δs_{obs}. We refer the reader to Figure 4 in [152], which presents the performance of many filtering and smoothing algorithms for tracking the state of the Lorenz-63 system as a function of Δs_{obs}.

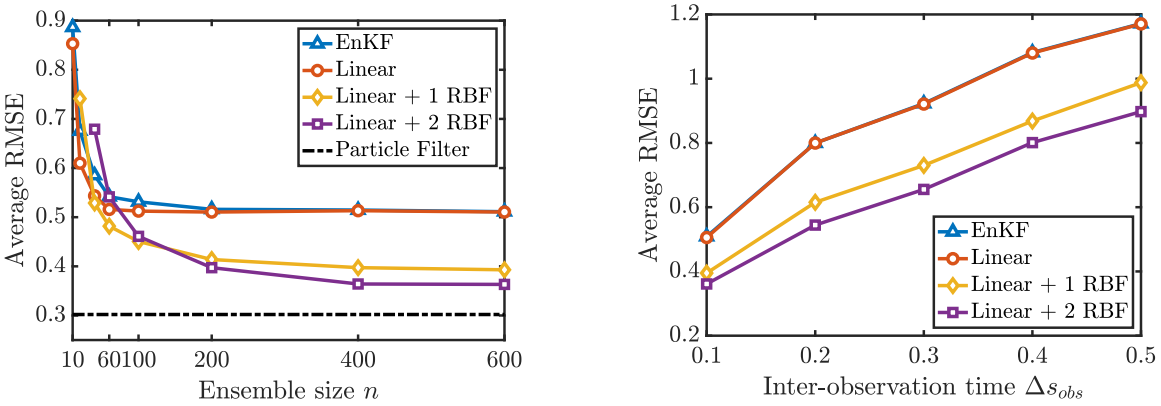


Figure 6-4: Average RMSE of the Lorenz-63 model for $\Delta s_{\text{obs}} = 0.1$ as a function of n (*left*), and for $n = 1000$ as a function of the inter-observation time Δs_{obs} (*right*).

For this low-dimensional problem, we also compare the posterior approximations from the stochastic map filter to the true filtering distribution. We first use a con-

sistent SIR particle filter with 10^6 particles to estimate the posterior mean $\boldsymbol{x}_t^{\text{PF}}$ and covariance matrix $\boldsymbol{\Sigma}_t^{\text{PF}}$ at each observation time t with $\Delta s_{\text{obs}} = 0.1$. Figure 6-5 plots the normalized L_2 error in the posterior mean, defined as $\|\widehat{\boldsymbol{X}}_t - \boldsymbol{x}_t^{\text{PF}}\|_2/\sqrt{d}$, and the normalized Frobenius norm in the posterior covariance matrix, defined as $\|\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t^{\text{PF}}\|_F/d$. Similarly to the RMSE, we observe an improvement in the approximation of the posterior statistics from gradually introducing nonlinear functions in the prior-to-posterior transformation. Lastly, the variability of the particle filter estimates are plotted for reference using dotted lines in Figure 6-5.

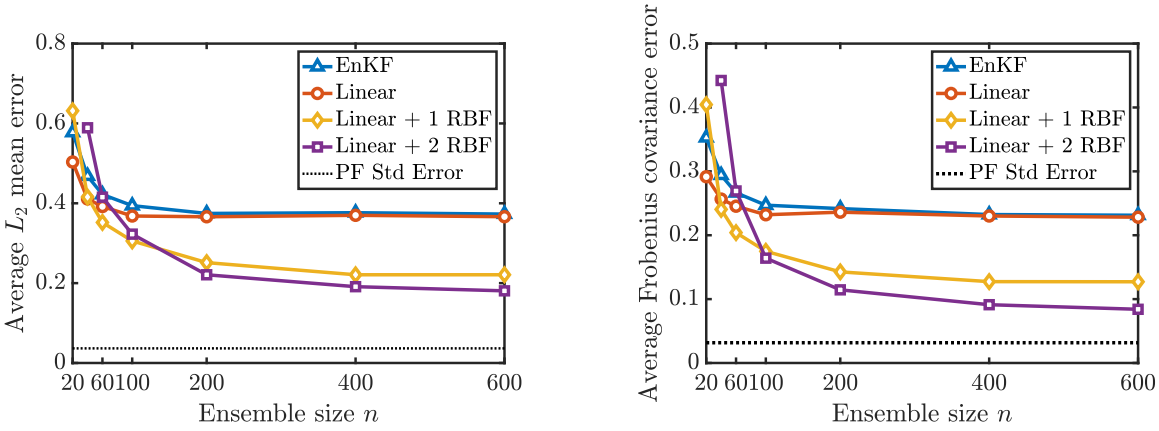


Figure 6-5: Normalized L_2 error of the posterior mean (*left*) and Frobenius error of the posterior covariance matrix (*right*) for the Lorenz-63 model.

To assess the predictive quality of each ensemble, we also compute the continuous ranked probability score (CRPS). The CRPS for the kth marginal component of the state at observation step t compares the ensemble's empirical CDF $\widehat{F}_{t,k}(x)$ with a Heaviside function centered at the true state $H(x - x_{t,k}^*)$, as given by $\text{CRPS}_{t,k} = \int(\widehat{F}_{t,k}(x) - \mathbb{1}(x_{t,k}^* \leq x))^2 \text{d}x$. We average $\text{CRPS}_{t,k}$ over state components k and 2000 assimilation cycles t. The left of Figure 6-6 plots the mean CRPS for the Lorenz-63 model with increasing ensemble size n. For $n > 60$ samples, the nonlinear stochastic map filter yields lower values of CRPS, indicating that the ensemble is better calibrated and has higher sharpness, as defined in [77]. For increasing nonlinearity in the prior-to-posterior transformation, the average CRPS for the Lorenz-63 model approaches that of a SIR particle filter.

Lastly, we comment on the choice of parameterization used in our data assimilation

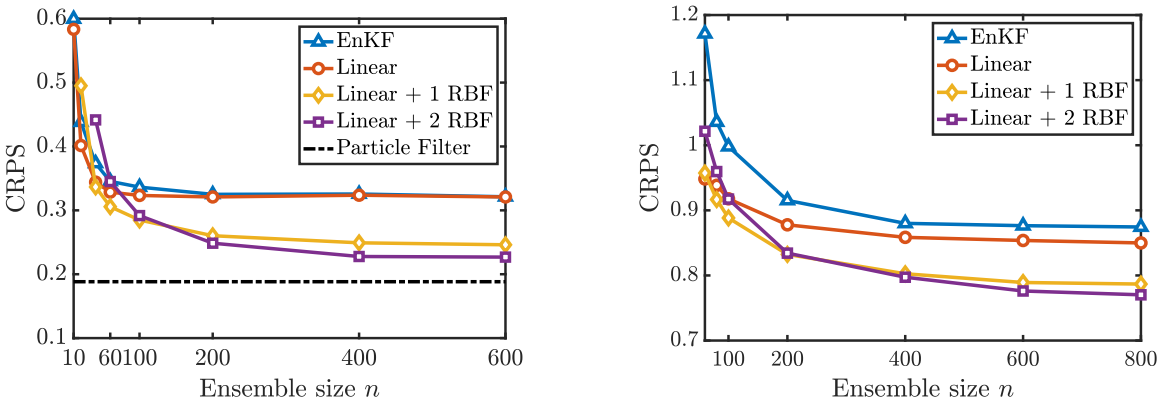


Figure 6-6: Average continuous ranked probability score (CRPS) of the analysis ensemble for the Lorenz-63 model (*left*) and the "hard" Lorenz-96 configuration (*right*).

experiments. Separable map components have a smaller number of parameters (as compared to non-separable expansions; see Chapter 3) and thus can be estimated with lower variance using moderate ensemble sizes. The form of the separable expansions, however, limits them to approximating conditional distributions where *only* the conditional mean depends non-linearly on the observation $\boldsymbol{y}$. Hence, these expansions may not reliably capture the posterior distributions for general inverse problems. As an example, we consider sampling from the posterior distribution for a one-dimensional parameter X with a standard Gaussian prior and the nonlinear observation operator $h(x) = \log|x|$ that was considered in [4]. The first row of Figure 6-7 plots approximations to the joint density $\pi_{X,Y}$ using (a) separable and (b) non-separable expansions with a combination of linear and $p \in \{3, 7\}$ RBFs given $n = 600$ i.i.d. samples from $\pi_{X,Y}$. We observe that separable expansions yield approximate conditional densities (i.e., normalized horizontal slices of the joint density) with unimodal shapes for all y^*. In contrast, non-separable expansions capture both unimodal (e.g., for $y^* = -3$) and bimodal conditional densities (e.g., for $y^* = 2$) arising from the non-identifiability of X's sign. The lower half of Figure 6-7 plots the analysis samples from the SM algorithm. The analysis samples for $y^* = 2$ with non-separable parameterizations better match the bi-modal posterior density. Characterizing these bimodal densities is important, especially when there are multiple states that are consistent with the observations. For the remainder of our experiments with local and monotone obser-

vation operators, however, we found that separable parameterizations are sufficient for tracking the states of the Lorenz-63 and Lorenz-96 dynamical systems.

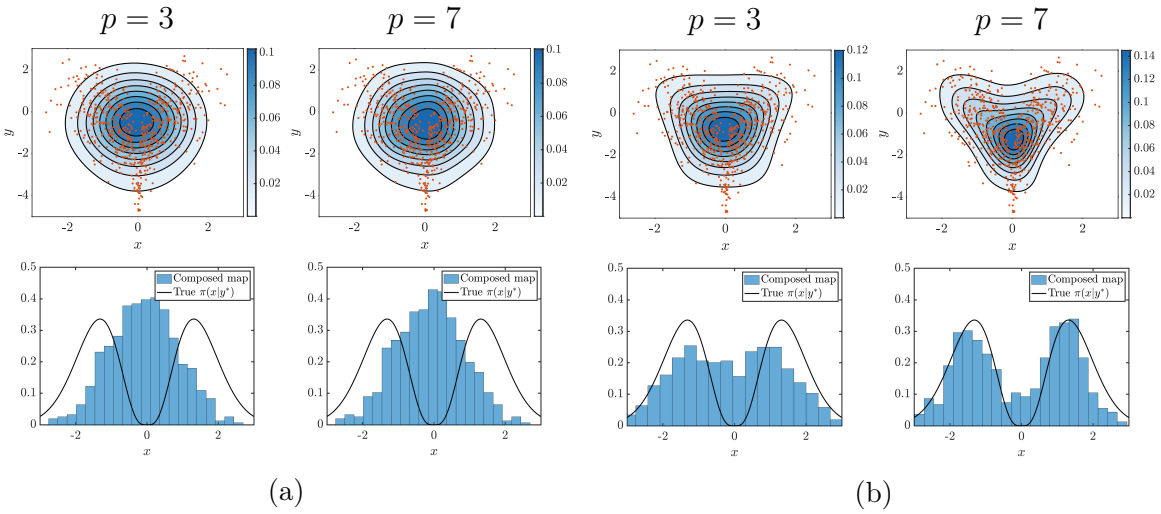


Figure 6-7: Joint density approximations (*top row*) and approximate posterior samples (*bottom row*) using (a) separable expansions and (b) tensor-product expansions of linear functions and $p = 3$ (*left*), and $p = 7$ (*right*) RBFs in each transport map component. The $n = 600$ samples used for approximating the transport maps are denoted in orange.

6.4 Sparse couplings for high-dimensional states

To extend the stochastic map filtering algorithm to high-dimensional problems with limited ensemble sizes, it is necessary to regularize the estimation of the prior-to-posterior map $\widehat{T}_{\boldsymbol{y}_t^*}$. One source of structure that we can exploit is sparsity in the KR rearrangement for $\widehat{S}^{\mathcal{X}}$, as discussed in Chapter 3. In the context of Bayesian inference problems, sparsity in $S^{\mathcal{X}}$ arises from conditional independence properties of the filtering distributions and from decay of dependence between state variables. These properties typically exist in spatial processes with local interactions [178].

One example of a spatial process on a periodic circular domain is the Lorenz-96 model (described in Sections 4.5.5 and 6.4.1). Figure 6-8 displays the values and sparsity of the average inverse covariance matrix for the forecast distributions of this 40-dimensional system. The covariance matrix at each observation time is estimated using 10^4 samples and the matrices are averaged over 1000 assimilation

cycles. Although the forecast distributions are non-Gaussian, the sparsity of the inverse covariance matrix indicates that the forecast, and potentially the posterior distribution, may be well approximated by a distribution where each state is only conditionally dependent on neighboring states in the circular grid. A similar result was obtained in Chapter 4 with only free model runs of the dynamical system, i.e., without data assimilation. Sparse triangular maps $\widehat{S}^{\mathcal{X}}$ define approximate densities $\widehat{\pi}_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t} = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}_t^*, \cdot)^\sharp \eta_{\mathbf{Z}_2}$ that can satisfy these conditional independence properties, or equivalently that represent sparse Markov random fields [204].

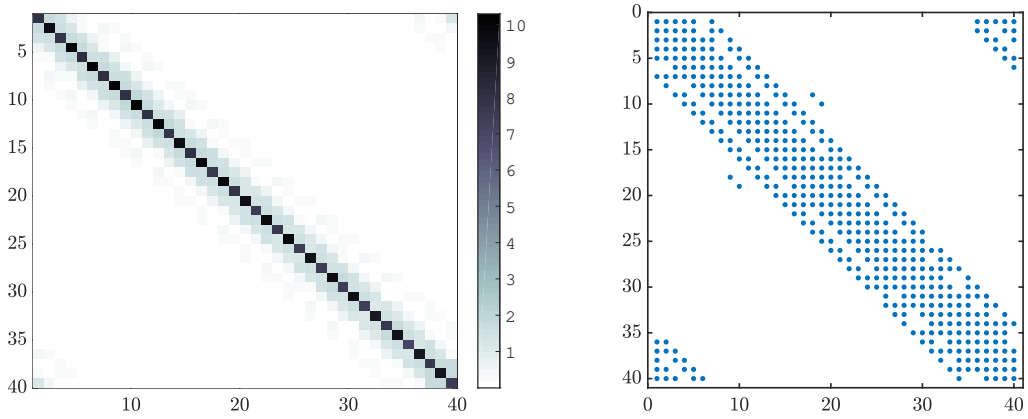


Figure 6-8: Average values (*left*) and non-zero entries (*right*) of the average inverse covariance matrix of the forecast distributions $\overline{\boldsymbol{\Sigma}}_{t+1}^{-1}$

There are several approaches to exploit sparsity when computing $\widehat{S}^{\mathcal{X}}$. These include learning sparse parameterizations as in Chapters 3 and 4 or imposing sparsity constraints directly on $\widehat{S}^{\mathcal{X}}$. In the statistics literature, a common regularization technique is to band or taper the inverse covariance matrix or its Cholesky factor according to a tuned correlation length parameter [20]. In this section, we follow the natural generalization of this approach by seeking nonlinear map components that depend on few variables. In particular, we consider estimators for the kth map component S_k given by $S_k(\boldsymbol{y}_t, \boldsymbol{x}_{t,\mathcal{A}_{k,r}})$ where $\mathcal{A}_{k,r}$ indexes the neighboring state variables that are within a distance r from $\boldsymbol{x}_{t,k}$ according to a distance metric d, i.e., $\mathcal{A}_{k,r} := \{i \in (1,\ldots,k) \text{ s.t. } d(i,k) \leq r\}$. Given a natural ordering for the state variables based on their physical location on a grid, each value of r defines a banded estimator for $S^{\mathcal{X}}$. Let us remark that finding sparse maps is analogous to localization

in the EnKF from tapering the covariance matrix. The difference is that sparsity in $S^{\mathcal{X}}$ assumes conditional independence between the states variable, while localization in the EnKF commonly assumes the stronger property of marginal independence.

Remark. *For linear maps $S^{\mathcal{X}}$, the localized estimators above correspond to banded estimators for the Cholesky factor of the precision matrix. The analysis of these estimators given multivariate Gaussian samples has been studied in [20, 232].*

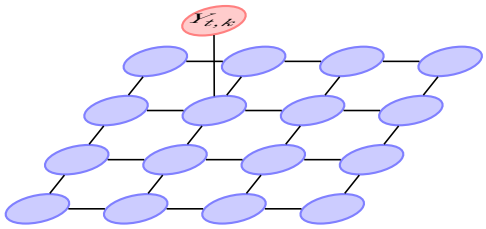


Figure 6-9: For local likelihood models, the scalar observation $Y_{t,k}$ corresponds to an noisy observation of a single component of $\mathbf{X}_t$. Typically, only marginal components of $\mathbf{X}_t$ that are physically close to $Y_{t,k}$ are affected by the observation.

A second source of structure is sparsity in the prior-to-posterior transformation $T_{\boldsymbol{y}_t^*}$ from assimilating local observations that depend on a single element of the state; see Figure 6-9. When the underlying process $\mathbf{X}_t$ has a short spatial correlation length, a scalar observation $Y_{t,k} = X_{t,\mathcal{K}} + \epsilon$ that is a function of a scalar state variable $X_{t,\mathcal{K}}$ for $\mathcal{K} \in \{1, \ldots, d\}$ only affects state variables close to $X_{t,\mathcal{K}}$. As a result, the map $T_{y_{t,k}^*}$ is expected to revert to the identity function for components that update state variables far from the observation location. To enforce this sparsity in the estimator for $T_{y_{t,k}^*}$, we order the state variables based on their spatial distance from $Y_{t,k}$ with respect to the distance metric d. We then look for a KR rearrangement $S^{\mathcal{X}}$ of the form

$$S^{\mathcal{X}}(y_{t,k}, \boldsymbol{x}_t) = \begin{bmatrix} s(y_{t,k}, \boldsymbol{x}_{t,1:l}) \\ x_{t,l+1} \\ \vdots \\ x_{t,d} \end{bmatrix}, \tag{6.13}$$

where $s \colon \mathbb{R}^{1+l} \to \mathbb{R}^{l+1}$ is a lower triangular and monotone function and l defines a localization radius. The composed maps, $T_{y_{t,k}^*}(y_{t,k}, \boldsymbol{x}_t) = S^{\mathcal{X}}(y_{t,k}^*, \cdot)^{-1} \circ S^{\mathcal{X}}(y_{t,k}^*, \boldsymbol{x}_t)$

with $S^{\mathcal{X}}$ given by (6.13) will then have only l non-identity components. The remaining components of $T_{y^*_{t,k}}$ correspond to state variables that are unaffected by the kth observation. Let us remark that the pullback distribution of the Gaussian reference through the map in (6.13) represents the marginal conditional distributions for $(X_{t,l+1}, \ldots, X_{t,d})$ as standard Gaussian. While a map of this form alone yields a poor estimator for the conditional distributions of $\mathbf{X}_t$, it is justified when we use it to estimate a prior-to-posterior transformation $T_{y^*_{t,k}}$ that has sparse structure.

The two localization parameters, the neighborhood size r and the number of non-identity components l, define a set of sparse triangular transport maps. The optimal values for these parameters for each problem can be identified by minimizing a measure for the quality of map estimation or the performance of the sequential inference procedure. One measure that is commonly used in data assimilation experiments is the average RMSE over time, as defined in 6.3.3. We adopt this measure to choose the regularization parameters l and r for each ensemble size and map parameterization in the following experiments. The EnKF results are also tuned by choosing the optimal localization radius for the Gaspari-Cohn tapering function, that is applied to each forecast covariance matrix.

6.4.1 Lorenz-96 dynamical system

In this section we present the performance of the localized stochastic map filter on the Lorenz-96 system. Lorenz-96 is a representative model for the mid-latitude atmosphere and is commonly used as a testbed for numerical weather prediction algorithms [134]. The model defines the evolution of a 40-dimensional periodic state vector $\mathbf{X}(s) = (X_1(s), \ldots, X_{40}(s))$ according to the coupled system of ODEs

$$\frac{dX_k}{ds} = (X_{k+1} - X_{k-2})X_{k-1} + X_k + F, \quad k = 1, \ldots, 40 \tag{6.14}$$

where F is a forcing term and the state is periodic, i.e., $X_0 = X_{40}$, $X_{-1} = X_{39}$ and $X_1 = X_{41}$. In our numerical experiments we set $F = 8$ to simulate chaotic behavior and discretize the ODEs using a Runge-Kutta method with a step size of $\Delta s = 0.01$.

Following the "hard case" setup of [125], we indirectly observe the state sparsely in space and time with $m = 20$ (i.e., observing every other component of the state) and $\Delta s_{\text{obs}} = 0.4$. Each set of observations has independent and additive Gaussian noise with variance $0.5\mathbf{I}_m$. The large inter-observation time corresponds approximately to observing the state every 2 days in a global weather model [137]. These parameters result in highly non-Gaussian forecast distributions that require localizing the maps $S^{\mathcal{X}}$ to have stable filtering performance with small ensemble sizes. In this study, we use the same separable parameterizations used for the maps $S^{\mathcal{X}}$ in Section 6.3.3.

Figure 6-10 shows the sensitivity of the average RMSE to the localization parameters r and l when using nonlinear maps. For large ensemble sizes, RMSE improves when increasing the number of non-identity map components l and the neighborhood size r. For each setting of these parameters, the filter with $p = 2$ RBFs (right) offers a slight benefit over the filter with $p = 1$ RBFs (left) for large ensemble sizes. We emphasize that increasing l, r, and p, for any $p \geq 1$, *all* comprise ways of increasing the nonlinearity of the map: not just via the choice of the approximation space for $\mathfrak{u}$, but also by expanding where in the map—and in which variables—nonlinearity appears.

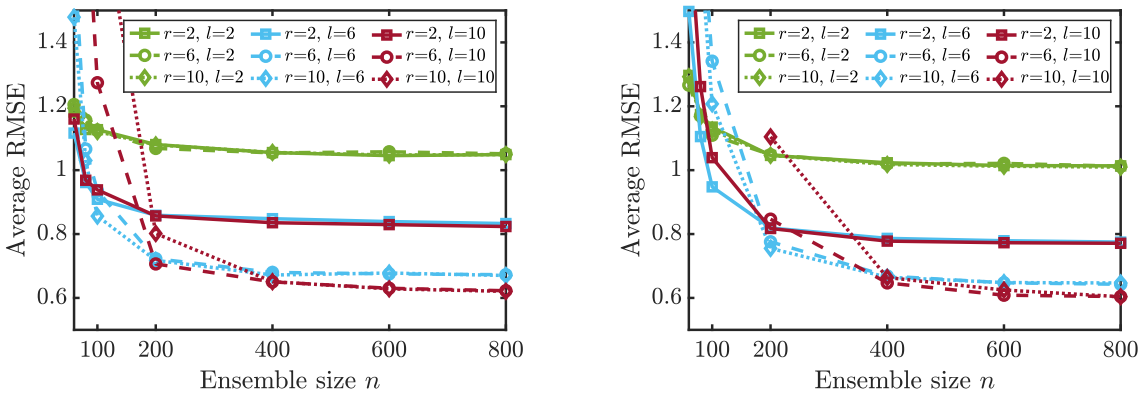


Figure 6-10: Average RMSE (over 2000 assimilation cycles) for the "hard" Lorenz-96 configuration using 'Linear + 1 RBF' maps (*left*) and 'Linear + 2 RBF' maps (*right*) for different localization parameters: the number of non-identity components l in the map $S^{\mathcal{X}}$, and the neighborhood size r defining the input dependence of each map component.

Figure 6-11 plots the average RMSE as a function of the ensemble size n with the optimal choice of localization parameters. With increasing nonlinearity in the map

$S^{\mathcal{X}}$, the optimal estimators for the stochastic map filter outperform the biased EnKF given sufficient samples. The regularized estimators for the map with local nonlinear terms capture the non-Gaussian forecast statistics and improve the tracking of the hidden state. The average spread of the ensemble, i.e., $[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_t)/d]^{1/2}$ where $\widehat{\boldsymbol{\Sigma}}_t$ is the ensemble covariance matrix for the filtering distribution at each step, also decreases with added nonlinearity as displayed on the right of Figure 6-11. These two results indicate that the ensemble from the stochastic map filter with nonlinear maps is tracking and concentrating around the true hidden state $\boldsymbol{x}_t^*$. Figure 6-12 produces similar observations for the median RMSE and average coverage probability.

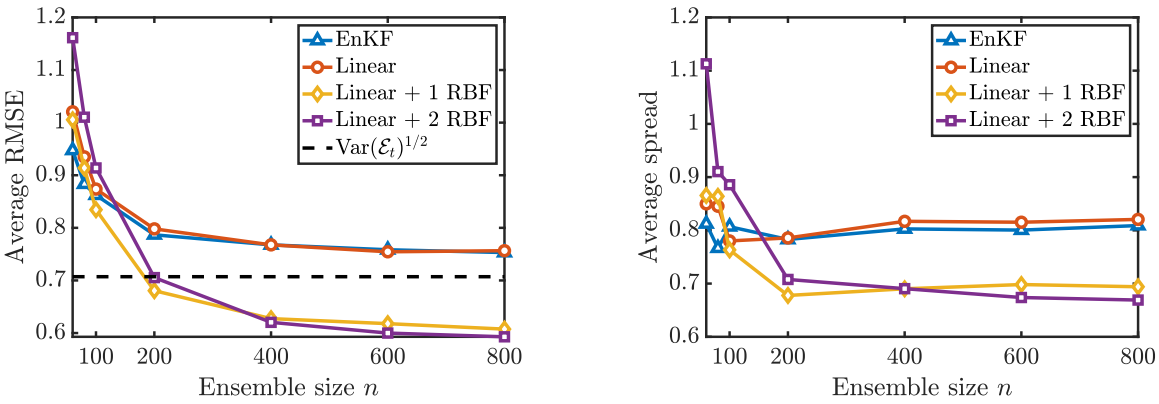


Figure 6-11: Average RMSE (*left*) and ensemble spread (*right*) over 2000 assimilation cycles for the "hard" Lorenz-96 configuration with $\Delta s_{\mathrm{obs}} = 0.4$ and sparse observations.

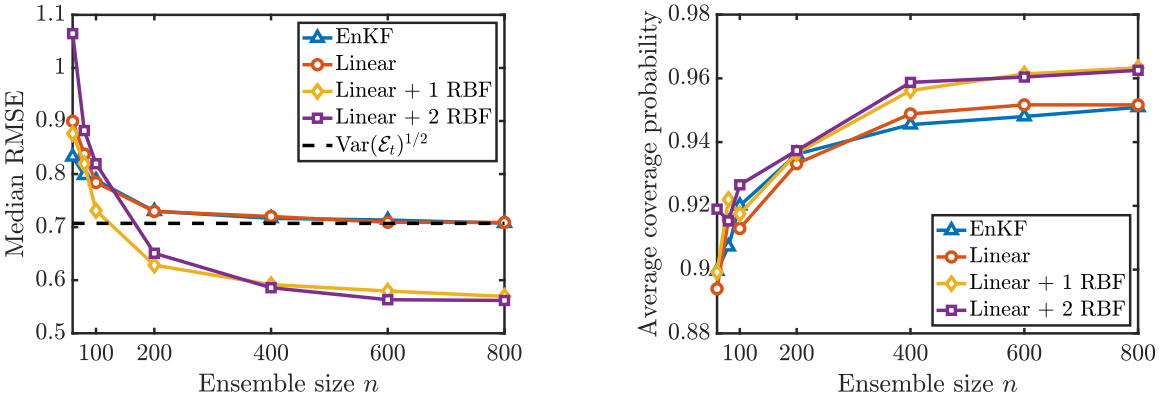


Figure 6-12: Median RMSE (*left*) and coverage probability (*right*) over 2000 assimilation cycles for the "hard" Lorenz-96 configuration with $\Delta s_{\mathrm{obs}} = 0.4$ and sparse observations.

6.4.2 Other configurations

To demonstrate the performance of the stochastic map filter with more complex likelihood models, we consider two "hard" configurations of the Lorenz-96 model with heavy-tailed observational noise and nonlinear observation operators.

First, we consider sparse and *heavy-tailed* observations with $\Delta s_{\text{obs}} = 0.1$, $m = 10$ (i.e., observing every fourth component of the state), and additive Laplace (i.e., double-exponential)-distributed observational noise. The noise has variance $2\theta^2\mathbf{I}_m$, where $\theta = 1$ is the scale parameter of the Laplace distribution. The large variance of the noise and limited number of observations make filtering difficult; for instance, large RMSE values are obtained when using the EnKF with ensemble sizes smaller than $n = 100$.

Similar to the results in Section 6.4.1, we observe improved tracking performance with increasing nonlinearity in the map $S^{\mathcal{X}}$, provided n is not too small. Figure 6-13 shows the average and median RMSE for the EnKF and different stochastic map parameterizations. As n increases, more complex transport maps yield the lowest value of RMSE, while the EnKF and linear maps offer more robust tracking for small ensemble sizes n. Furthermore, Figure 6-14 illustrates that more nonlinearity in the maps (i.e., higher numbers of RBFs in each component) increases the average coverage probability of the ensemble across 2000 assimilation cycles without increasing the ensemble's average spread. These results suggest that the stochastic map filter with more complex transformation can better capture the true state $\boldsymbol{x}_t^*$ with a sharp (i.e., concentrated) ensemble for problems with non-Gaussian forecast and analysis distributions.

Second, we consider a Lorenz-96 system with *nonlinear observations* of the state using the observation operator considered in [4]. We begin with the "hard case" configuration of the Lorenz-96 system in Section 6.4.1, with $\Delta s_{\text{obs}} = 0.4$ (long time between observations) and $m = 20$ observations at each assimilation step (i.e., observing every other component of the state), but we modify the observation model. Now, each component of $\mathbf{Y}_t \in \mathbb{R}^m$ results from a nonlinear square root transforma-

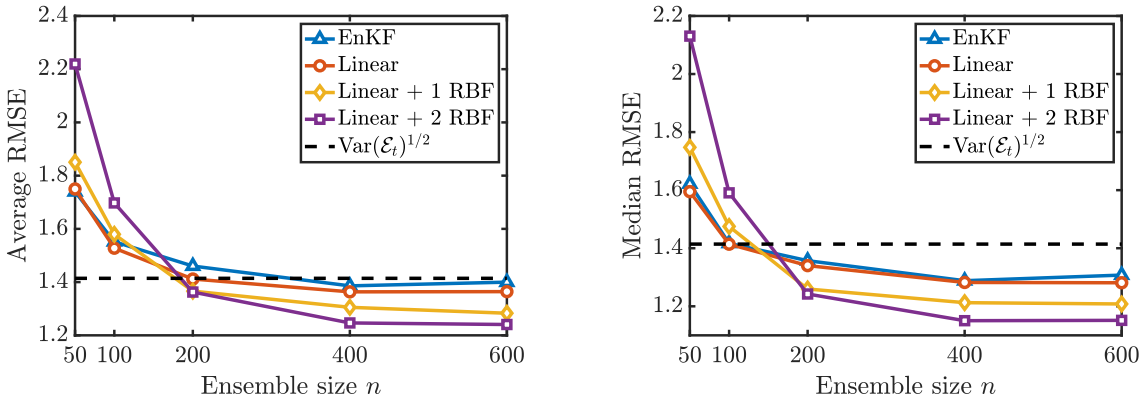


Figure 6-13: Average (*left*) and median (*right*) RMSE (over 2000 assimilation cycles) for the "hard" Lorenz-96 configuration of Section 6.4.1, with $\Delta s_{obs} = 0.4$, and $m = 10$ heavy-tailed observations of the state.

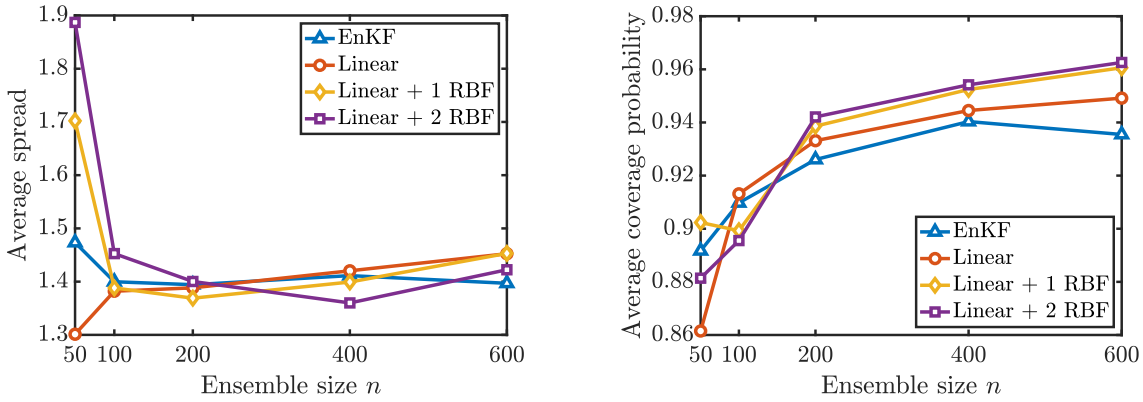


Figure 6-14: Average ensemble spread (*left*) and average coverage probability of the $[2.5\%, 97.5\%]$ empirical intervals of each state marginal (*right*) for the "hard" Lorenz-96 configuration with heavy-tailed observations of the state.

tion of the corresponding state component, perturbed by additive Gaussian noise. In other words, each component of $\mathbf{Y}_t$ is given by $Y_{t,k} = \mathrm{sign}(X_{t,2k-1})\sqrt{|X_{t,2k-1}|} + \mathcal{E}_{t,k}$, where $\{\mathcal{E}_{t,k}\}_{k=1}^{d} \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.25)$ are independent of $\mathbf{X}_t$. The observational noise variance is selected to maintain a similar signal-to-noise ratio as in the direct observation case. In this experiment, we use the separable map parameterization described in Section 6.3.3 with component functions S_k that depend linearly on the state variable $x_{t,k}$ and only depend nonlinearly on the other variables, for all k.

Figure 6-15 plots the average and median RMSE for the stochastic EnKF and the stochastic map filters. Similarly to the results in 6.4.1, the stochastic EnKF and linear maps have the lowest average RMSE for small n, while the richer maps result in

lower RMSE for larger ensemble sizes because they have less bias. Figure 6-16 shows that the marginal coverage probability increases monotonically with increasing map complexity, meaning that the ensemble derived from richer maps covers more often the true hidden state.

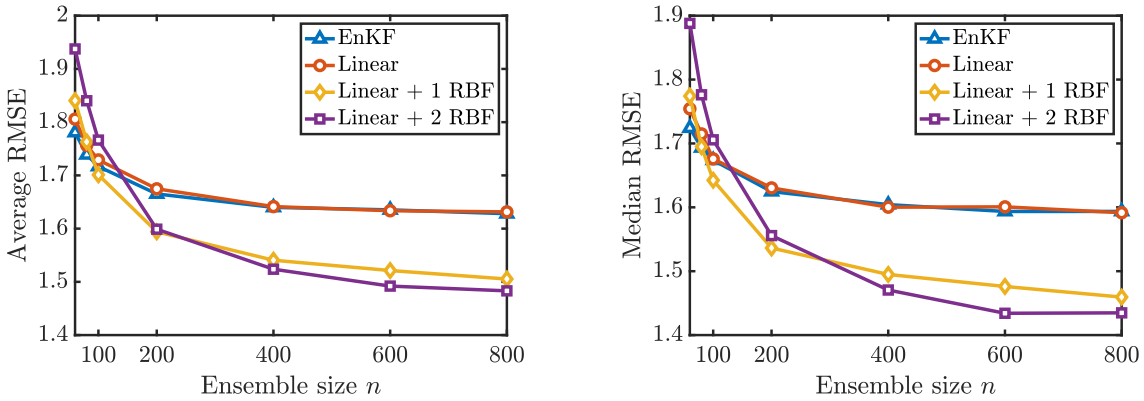


Figure 6-15: Average (*left*) and median (*right*) RMSE (over 2000 assimilation cycles) for the "hard" Lorenz-96 configuration of Section 6.4.1, with $\Delta s_{obs} = 0.4$, and $m = 20$ square-root observations of the state.

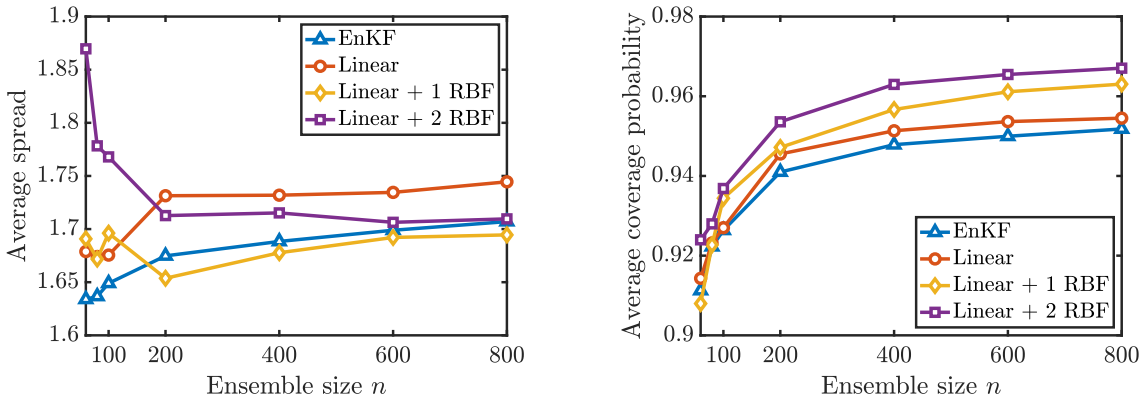


Figure 6-16: Average ensemble spread (*left*) and average coverage probability of the $[2.5\%, 97.5\%]$ empirical intervals of each state marginal (*right*) for the "hard" Lorenz-96 configuration with square-root observations of the state.

6.5 Discussion and extensions

In this chapter we introduced a nonlinear filtering algorithm that generalizes the ensemble Kalman filter (EnKF) to robustly estimate the states of dynamical systems,

while approaching the asymptotic consistency of particle filters. At each observation time, this algorithm estimates a triangular transport map using forecast samples and observations sampled from the likelihood model, and constructs a prior-to-posterior transformation to sample from each filtering distribution. These transformations reduce to the EnKF when the transport maps are constrained to a class of affine functions. By seeking nonlinear maps, the algorithm generalizes the EnKF using nonlinear prior-to-posterior transformations.

Our numerical experiments show that nonlinear maps can outperform the EnKF for tracking the states of chaotic dynamical systems, such as the Lorenz-63 and Lorenz-96 models. The improvement, however, depends on having a sufficient number of ensemble members to reliably learn complex maps. For small ensemble sizes n, richer maps yield estimators with less bias but higher variance. The EnKF and linear maps, on the other hand, yield more biased but lower variance estimators that offer more stable tracking performance for small n. In this work, we explore various sources of sparse structure in triangular maps for regularizing the estimation of high-dimensional prior-to-posterior transformations. These estimators result in a sequence of nonlinear filters (with an increasing number of parameters) that more accurately estimate the state given a modest increase in n. Overall, the stochastic filter provides a flexible framework for adjusting the complexity of the map to build good estimators for the prior-to-posterior transformations in sequential inference problems.

In what follows, we briefly outline some directions for future work.

Smoothing Another common data assimilation problem is smoothing, which improves estimates of past states given new observations. Similarly to filtering, the stochastic map algorithm can extend linear ensemble-based smoothers to sample consistently from the the joint smoothing distribution $\pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T}$ where $\mathbf{X}_{1:T} = (\mathbf{X}_1, \ldots, \mathbf{X}_T)$, or any of its marginals $\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T}$ for $t \leq T$. To simplify our notation we omit the conditioning on previous data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{T-1}$ and without loss of generality only consider conditioning on a new observation at time $t = T$, as seen in Figure 6-17. In an ensemble setting, our goal is then to sample from $\pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_T}$ after observing the

data $\boldsymbol{y}_T$ at time $t = T$, given a collection of samples from $\pi_{\mathbf{X}_{1:T}}$.

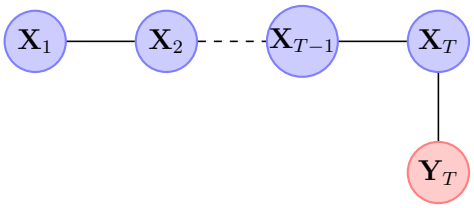


Figure 6-17: Markov structure for a local observation in time

Similarly to the filtering problem, we can construct triangular transport maps that push-forward an ensemble from the distribution $\pi_{\mathbf{X}_{1:T}}$ to the smoothed distribution conditioned on new data $\pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_T}$. To do so, we define a triangular map $S\colon \mathbb{R}^m \times \mathbb{R}^{Td} \to \mathbb{R}^m \times \mathbb{R}^{Td}$ of the form

$$S(\boldsymbol{y}_T, \boldsymbol{x}_{1:T}) = \begin{bmatrix} S^{\mathcal{Y}}(\boldsymbol{y}_T) \\ S^{\mathcal{X}}(\boldsymbol{y}_T, \boldsymbol{x}_{1:T}) \end{bmatrix}, \tag{6.15}$$

where S *pulls-back* the $Td + m$-dimensional standard Gaussian density η to the joint density of $\mathbf{X}_{1:T}$ and $\mathbf{Y}_T$, i.e., $S^\sharp \eta = \pi_{\mathbf{Y}_T, \mathbf{X}_{1:T}}$. While the first component of the map $S^{\mathcal{Y}}\colon \mathbb{R}^m \to \mathbb{R}^m$ characterizes the marginal distribution for the data, $S^{\mathcal{X}}\colon \mathbb{R}^m \times \mathbb{R}^{Td} \to \mathbb{R}^{Td}$ pulls-back the Td-dimensional standard Gaussian reference density to the conditional density $\pi_{\mathbf{X}_{1:T}|\mathbf{Y}_T}$. While any map of the form in (6.15) can be used for smoothing, we outline an algorithm based on a specific choice of ordering for the state variables $\boldsymbol{x}_{1:T}$ in the triangular map.

Let $S^{\mathcal{X}}$ be a map that updates the states backwards in time using the ordering: $\boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_1$. This map has the form

$$S^{\mathcal{X}}(\boldsymbol{y}_T, \boldsymbol{x}_{1:T}) = \begin{bmatrix} S_1(\boldsymbol{y}_T, \boldsymbol{x}_T) \\ S_2(\boldsymbol{y}_T, \boldsymbol{x}_T, \boldsymbol{x}_{T-1}) \\ \vdots \\ S_T(\boldsymbol{y}_T, \boldsymbol{x}_T, \boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_1) \end{bmatrix}. \tag{6.16}$$

From the structure of the hidden Markov model for the state and observation variables, the states satisfy the conditional independence properties $\mathbf{X}_k \perp\!\!\!\perp (\mathbf{X}_j, \mathbf{Y}_T) \,|\, \mathbf{X}_{k+1}$ for all $k < T$ and $k + 1 < j \leq T$. Applying Theorem 3 in [204] for the sparsity of

the KR rearrangement that pushes forward $\pi_{\mathbf{X}_{1:T},\mathbf{Y}_T}$ to a Gaussian reference density η, the map in (6.16) has the equivalent sparse representation

$$S^{\mathcal{X}}(\boldsymbol{y}_T, \boldsymbol{x}_{1:T}) = \begin{bmatrix} S_1(\boldsymbol{y}_T, \boldsymbol{x}_T) \\ S_2(\boldsymbol{x}_T, \boldsymbol{x}_{T-1}) \\ S_3(\boldsymbol{x}_{T-1}, \boldsymbol{x}_{T-2}) \\ \vdots \\ S_T(\boldsymbol{x}_2, \boldsymbol{x}_1) \end{bmatrix}. \tag{6.17}$$

This map has T components that all depend on at most two states or observation variables. From the properties of lower triangular transports, the mapping $\boldsymbol{\xi} \mapsto S^{\mathcal{X}}(\boldsymbol{y}_T^*, \boldsymbol{\xi})$ pushes-forward the conditional $\pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_T^*}$ to the standard normal reference; see Chapter 2. Therefore, we can invert the triangular map in (6.17) to push-forward Gaussian samples to sample from the conditional density $\pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_T^*}$ using the inverse map

$$S^{\mathcal{X}}(\boldsymbol{y}_T^*, \cdot)^{-1} := \begin{bmatrix} S_1(\boldsymbol{y}_T^*, \cdot)^{-1} \\ S_2(\widetilde{\boldsymbol{x}}_T, \cdot)^{-1} \\ \vdots \\ S_T(\widetilde{\boldsymbol{x}}_2, \cdot)^{-1} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{x}}_T \\ \widetilde{\boldsymbol{x}}_{T-1} \\ \vdots \\ \widetilde{\boldsymbol{x}}_1 \end{bmatrix}, \tag{6.18}$$

where each component is evaluated at the previous component's output and a reference variable, i.e., $\widetilde{\boldsymbol{x}}_k = S_{T-k+1}(\widetilde{\boldsymbol{x}}_{k+1}, \cdot)^{-1}$. The output of the previous component can be interpreted as the smoothed random variable conditioned on the data $\boldsymbol{y}_T^*$.

By composing (6.18) and (6.17), we can define a prior-to-posterior transformation $T_{\boldsymbol{y}_T^*} \colon \mathbb{R}^m \times \mathbb{R}^{Td} \to \mathbb{R}^{Td}$ given by

$$T_{\boldsymbol{y}_T^*}(\boldsymbol{y}_T, \boldsymbol{x}_{1:T}) = S^{\mathcal{X}}(\boldsymbol{y}_T^*, \cdot)^{-1} \circ S^{\mathcal{X}}(\boldsymbol{y}_T, \boldsymbol{x}_{1:T}). \tag{6.19}$$

Evaluating the map $T_{\boldsymbol{y}_T^*}$ at i.i.d. samples $\{(\boldsymbol{Y}_T^i, \boldsymbol{X}_{1:T}^i)\}_{i=1}^n \sim \pi_{\mathbf{Y}_T,\mathbf{X}_{1:T}} = \pi_{\mathbf{Y}_T|\mathbf{X}_T}\pi_{\mathbf{X}_{1:T}}$ yields posterior samples $\widetilde{\boldsymbol{x}}_{1:T}^i = T(\boldsymbol{Y}_T^i, \boldsymbol{X}_{1:T}^i) \sim \pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_T^*}$. Let us remark that the first component of the transformation in (6.19) is equivalent to the composed map

derived by the stochastic map filtering algorithm in Section 6.3 at time T. The remaining components propagate the information from the last observation to smooth the previous states.

Remark. *To sample from $\pi_{\mathbf{X}_{1:T}|\boldsymbol{y}_{1:T-1}}$, we can run a smoother forward in time. Starting from analysis samples $\{\boldsymbol{X}_1^i\}_{i=1}^n \sim \pi_{\mathbf{X}_1|\boldsymbol{y}_1^*}$ at time $t = 1$, we can apply the forward model to generate samples $\boldsymbol{X}_2^i \sim \pi_{\mathbf{X}_2|\mathbf{X}_1,\mathbf{Y}_1}(\cdot|\boldsymbol{X}_1^i, \boldsymbol{y}_1^*)$ for $i = 1, \ldots, n$. Augmenting the analysis with the forecast ensemble yields samples from the joint distribution $\pi_{\mathbf{X}_1,\mathbf{X}_2|\boldsymbol{y}_1^*}$. We can then apply the procedure above to condition both states on a new observation $\boldsymbol{y}_2^*$ to sample from $\pi_{\mathbf{X}_{1:2}|\boldsymbol{y}_{1:2}}$. This process can be repeated recursively for the next smoothing distributions up to time T.*

Remark. *When the map components in (6.17) are affine functions, the prior-to-posterior transformation in (6.19) corresponds exactly to the RTS Smoother [176]. Analogously to the EnKF, the transformation for updating the kth marginal state can be written as a function of the cross-covariance $\boldsymbol{\Sigma}_{\mathbf{X}_k,\mathbf{X}_{k+1}}$ between neighborhood states $(\mathbf{X}_k, \mathbf{X}_{k+1})$, the marginal inverse covariance $\boldsymbol{\Sigma}_{\mathbf{X}_{k+1}}^{-1}$, and the difference between a sample of the smoothed state $\widetilde{\boldsymbol{x}}_{k+1}$ and the non-conditioned state $\boldsymbol{x}_{k+1}$.*

Schrödinger problem. A one-step approach to sample from the filtering distribution is to solve the Schrödinger problem. Given samples from the filtering distribution at time t, this problem seeks a Markov transition kernel $q_{\mathbf{X}_{t+1}|\mathbf{X}_t}$ that links the filtering distributions at times t and $t+1$ as follows:

$$\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}(\boldsymbol{x}_{t+1}) = \int q_{\mathbf{X}_{t+1}|\mathbf{X}_t}(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}(\boldsymbol{x}_t)\mathrm{d}\boldsymbol{x}_t. \tag{6.20}$$

Finding the kernel $q_{\mathbf{X}_{t+1}|\mathbf{X}_t}$ allows for sampling the posterior distribution without having to first generate samples from the forecast distribution $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t}$ as in Section 6.3 or the smoothing distribution $\pi_{\mathbf{X}_t|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}$ above. Figure 6-18 graphically compares three scenarios for sampling from $\pi_{\mathbf{X}_{t+1}|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{t+1}}$ based on filtering, smoothing, and solving the Schrödinger problem.

In a recent survey of data assimilation methods, Reich proposed an approach to

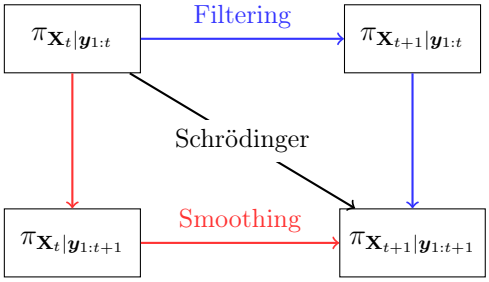


Figure 6-18: Illustration of three scenarios for sampling from the posterior distribution in a single step of a sequential Bayesian inference problem. The stochastic maps filter first forecasts the state and then pushes-forward samples to the filtering distribution. Alternatively, the stochastic map smoothing algorithm first samples from the $\mathbf{X}_t$ marginal of the smoothing distribution and then uses the dynamical model to sample from the filtering distribution. The Schrödinger problem produces samples in a single step using the transition kernel in (6.20).

solve the Schrödinger problem [180]. To find a nonparametric estimator for $q_{\mathbf{X}_{t+1}|\mathbf{X}_t}$, Reich uses the iterative Sinkhorn algorithm to twist the kernel for the forward dynamics to satisfy the condition in (6.20). Future work may explore how to build transformations based on KR rearrangements that directly couple the filtering distributions at times t and $t+1$.

Continuous-time filtering. From a more theoretical perspective, it will be interesting to explore the stochastic map algorithm in a continuous-time filtering setting. The hidden states and observations in the continuous-time model are described by the stochastic differential equations

$$d\mathbf{X}_t = f(\mathbf{X}_t)dt + d\mathbf{B}_t \tag{6.21}$$

$$d\mathbf{Y}_t = h(\mathbf{X}_t)dt + d\mathbf{W}_t, \tag{6.22}$$

where $d\mathbf{B}_t$ and $d\mathbf{W}_t$ denote two mutually independent Wiener processes taking values in $\mathbb{R}^d$ and $\mathbb{R}^m$, respectively. In this setting, the goal is to estimate the filtering distribution of $\mathbf{X}_t$ given the time history of observations $(\boldsymbol{y}_s)_{s\leq t}$

In principle, the solution of the continuous-time filtering problem is provided by the Kushner-Stratonovich (KS) and Zakai equations [120]. For the linear-Gaussian

case (i.e., linear f and h), the filtering distributions are also Gaussian and their evolution is described by the finite-dimensional Kalman-Bucy filter. For general dynamical and observations models, the KS and Zakai equations are stochastic PDEs and can not be solved in closed-form. Analogously to the setting of discrete-time observations, the filtering distributions in this setting are commonly approximated by interacting particle systems. Two SDEs for evolving these particles are the ensemble Kalman-Bucy filter, which is based on a Gaussian approximation of the forecast distribution, and the feedback particle filter (FPF) [235], which computes a nonlinear control term to move each particle. In the FPF, the control term contains a nonlinear Kalman gain $K_t(\boldsymbol{x})$ that can be computed by solving a Poisson equation. In [212], the authors prove that if the gain is computed exactly, the FPF is a consistent algorithm for sampling from the filtering distribution in the large-sample limit. Recently, [212] showed that the nonlinear Kalman gain can also be written as the derivative of a transport map. In particular, the nonlinear Kalman gain is given by $K_t = \frac{dS_\epsilon}{d\epsilon}|_{\epsilon=0}$ where S_ϵ is an ϵ-parametrized family of transport maps that push forward the filtering density $\pi_t \coloneqq \pi_{\mathbf{X}_t|(\boldsymbol{y}_s)_{s\leq t}}$ at time t, to the perturbed densities $\pi_{t,\epsilon}(\boldsymbol{x}) = \pi_t(\boldsymbol{x})(1+\epsilon h(\boldsymbol{x})-\mathbb{E}_{\pi_t}[h])$ for a scalar observation operator h. It will be interesting to use this link to connect the stochastic map (SM) filter with the feedback particle filter, and to understand the conditions for consistency of the SM filter in the limit of increasing ensemble size and complexity of the transport maps.

Other sources of low-dimensional structure. It will be interesting to identify and develop novel estimators for high-dimensional transport maps in the context of the stochastic map filter. While Section 6.4 shows how to exploit sparsity in the map that arises from (approximate) conditional independence, this structure may not be present in other data assimilation problems. For instance, the state may not be discretized on a regular spatial grid and the observation operator may be non-local, i.e., each observation is a function of all state components. In these cases, other structural properties of data assimilation problems may be used to regularize the estimation of the transport maps with limited ensemble sizes. These include

multi-scale decompositions of the state [230], and low-dimensional subspaces from the dynamical model such as the unstable subspace [34]. The next chapter will present an approach to identify subspaces of the state $\mathbf{X}$ and observation $\mathbf{Y}$ such that the prior-to-posterior update $T_{\boldsymbol{y}^*}$ only depends non-trivially on a low-dimensional projection of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Chapter 7

Parameter and data dimension reduction for Bayesian inference

7.1 Introduction

In several applications throughout science and engineering, the parameters $\mathbf{X}$ and/or observations $\mathbf{Y}$ in a Bayesian inference problem are high-dimensional vectors, possibly arising from the discretizations of infinite-dimensional signals. The goal of computational Bayesian inference is to characterize the conditional distribution for $\mathbf{X}|\mathbf{Y}$ from their high-dimensional joint distribution. While many sampling-based algorithms have been developed for Bayesian inference, their computational costs typically scale unfavorably with the growing dimensions of $\mathbf{X}$ and $\mathbf{Y}$ [182].

One method that has received increasing attention for reducing the cost of inference procedures is to approximate the posterior distribution as an update of the prior in a low-dimensional subspace of the parameters. Recent work has presented various approaches for identifying this subspace. [205] approximated the posterior covariance in linear-Gaussian inverse problems as a low-rank update of the prior covariance. This idea was extended to non-Gaussian Bayesian inference problems in [50, 239, 238] by using gradients of the likelihood function or the forward model to identify directions in the parameter space where the observations are most informative. The Bayesian update, in the form of Markov chain Monte Carlo (MCMC) sampling,

can then be restricted to this subspace. This procedure has shown immense benefit for scaling MCMC algorithms to sample from complex posterior distributions with high-dimensional parameters [49].

Reducing the parameter dimension is not sufficient, however, for inference algorithms such as approximation Bayesian computation (ABC) and the measure transport approaches presented in Chapter 5. As a concrete example, let $S^{\mathcal{X}}(\boldsymbol{y}, \cdot)\colon \mathbb{R}^d \to \mathbb{R}^d$ be a transport map that pushes forward the conditional density $\pi_{\mathbf{X}|\boldsymbol{y}}$ to a standard Gaussian reference density $\eta_{\mathbf{Z}_2}$ of dimension $\mathbb{R}^d$ for any $\boldsymbol{y} \in \mathbb{R}^m$. This map can be used to cheaply simulate from the posterior density $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y}^*)$ for an observation $\boldsymbol{y}^* \in \mathbb{R}^m$ by generating samples $\boldsymbol{Z}_2^i \sim \eta_{\mathbf{Z}_2}$ and inverting the map $S^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{X}^i) = \boldsymbol{Z}_2^i$ for each sample $\boldsymbol{X}^i$. Furthermore, we can construct low-variance Monte Carlo estimators for expectations of the posterior distribution by generating *many* i.i.d. samples once we have this map. Thus, the computational cost of inference is transferred to a problem of learning the map $S^{\mathcal{X}}$ as a function of $d + m$ inputs. For high-dimensional parameters and observations, this function approximation problem will in general suffer from the curse of dimensionality. To alleviate the dependence-dimension, Chapters 3 and 6 exploit sparse structure in triangular transports. For many physical problems, however, the map may not be sparse with respect to its canonical inputs. In these settings, a natural strategy is to reduce the dimensions of *both* the parameters and observations in order to learn these maps reliably given limited samples from the joint distribution.

In this chapter our goal is to concurrently identify low-dimensional projections of the parameters and observations in high-dimensional Bayesian inference problems. To do so, we look for a decomposition of the parameters into parts that are informed and uninformed by the observations along with a decomposition of the observations into parts that are informative and non-informative of the parameters. We then approximate the posterior by characterizing the conditional distribution for the informed parameters $\mathbf{X}_r \in \mathbb{R}^r$ for $r \ll d$ given only the informative observations $\mathbf{Y}_s \in \mathbb{R}^s$ for $s \ll m$. In the context of measure transport with triangular maps, these decomposi-

tions allow us to replace $S^{\mathcal{X}}$ with the approximate map

$$S^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}) = \begin{bmatrix} S_r^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}_r) \\ S_\perp^{\mathcal{X}}(\boldsymbol{y}, \boldsymbol{x}_r, \boldsymbol{x}_\perp) \end{bmatrix} \approx \begin{bmatrix} S_r^{\mathcal{X}}(\boldsymbol{y}_s, \boldsymbol{x}_r) \\ S_\perp^{\mathcal{X}}(\boldsymbol{x}_r, \boldsymbol{x}_\perp) \end{bmatrix}, \tag{7.1}$$

where $S_r \colon \mathbb{R}^{s+r} \to \mathbb{R}^r$ and $S_\perp \colon \mathbb{R}^d \to \mathbb{R}^{d-r}$ are functions of fewer variables than the original map components. In this chapter, we formulate a general information-theoretic approach for measuring the information lost in the posterior from using the map in (7.1). Furthermore, we define upper bounds that can be tractably minimized to find the variable decompositions. We note that in this work we will restrict ourselves to decompositions defined by linear subspaces of $\mathbf{X}$ and $\mathbf{Y}$.

The remainder of this chapter is organized as follows. In Section 7.2 we show the relation between the posterior approximation error and gradient information of the log-likelihood function and we propose several methods to identify optimal variable decompositions. In Section 7.3 we interpret the approximation error using conditional mutual information. In Section 7.4 we specialize our results to Gaussian likelihood functions with nonlinear forward models and discuss connections with related work in the linear-Gaussian case. In Section 7.5 we compare our approach to other classic dimension reduction strategies. Section 7.6 provides several inference algorithms that can leverage this joint dimension reduction of the parameters and observations. Lastly, Section 7.7 presents numerical experiments for forward models defined by an elliptic PDE, a high-dimensional imaging problem, and a stochastic differential equation.

7.2 Reducing parameter and observation dimensions

We begin by proposing a joint dimension reduction of the parameters and observations which relies on the detection of conditional independence between blocks of variables. Given two unitary matrices $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{m\times m}$, which we partition as column blocks $U = [U_r, U_\perp]$ and $V = [V_s, V_\perp]$ with $U_r \in \mathbb{R}^{d\times r}$ and $V_s \in \mathbb{R}^{m\times s}$, we decompose

$\mathbf{X}$ as

$$\mathbf{X} = U_r \mathbf{X}_r + U_\perp \mathbf{X}_\perp \qquad \text{where} \begin{cases} \mathbf{X}_r = U_r^T \mathbf{X} \\ \mathbf{X}_\perp = U_\perp^T \mathbf{X} \end{cases}, \tag{7.2}$$

and $\mathbf{Y}$ as

$$\mathbf{Y} = V_s \mathbf{Y}_s \ + V_\perp \mathbf{Y}_\perp \qquad \text{where} \begin{cases} \mathbf{Y}_s = V_s^T \mathbf{Y} \\ \mathbf{Y}_\perp = V_\perp^T \mathbf{Y} \end{cases}. \tag{7.3}$$

In this decomposition, $\mathbf{X}_\perp$ are interpreted as the un-informed parameters if they are independent of the data $\mathbf{Y}$ after conditioning on $\mathbf{X}_r$, that is $\mathbf{X}_\perp \perp\!\!\!\perp \mathbf{Y}|\mathbf{X}_r$. In the same way, $\mathbf{Y}_\perp$ are interpreted as the un-informative observations if they are independent of the informed parameters $\mathbf{X}_r$ after conditioning on $\mathbf{Y}_s$, that is $\mathbf{X}_r \perp\!\!\!\perp \mathbf{Y}_\perp|\mathbf{Y}_s$. Under these two conditional independence properties, the joint probability density function factorizes as $\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) = \pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\boldsymbol{x}_\perp|\boldsymbol{x}_r)\pi_{\mathbf{X}_r,\mathbf{Y}_s}(\boldsymbol{x}_r,\boldsymbol{y}_s)\pi_{\mathbf{Y}_\perp|\mathbf{Y}_s}(\boldsymbol{y}_\perp|\boldsymbol{y}_s)$, so that the posterior satisfies $\pi_{\mathbf{X}|\mathbf{Y}} = \pi^*_{\mathbf{X}|\mathbf{Y}}$ with

$$\pi^*_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) := \pi_{\mathbf{X}_r|\mathbf{Y}_s}(\boldsymbol{x}_r|\boldsymbol{y}_s)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\boldsymbol{x}_\perp|\boldsymbol{x}_r). \tag{7.4}$$

Put in words, the inference problem $\mathbf{X}|\mathbf{Y}$ can be transformed into a lower-dimensional inference problem $\mathbf{X}_r|\mathbf{Y}_s$. In practice, however, the conditional independence criterion $\mathbf{X}_\perp \perp\!\!\!\perp \mathbf{Y}|\mathbf{X}_r$ and $\mathbf{X}_r \perp\!\!\!\perp \mathbf{Y}_\perp|\mathbf{Y}_s$ might not be exactly satisfied and so $\pi_{\mathbf{X}|\mathbf{Y}} \neq \pi^*_{\mathbf{X}|\mathbf{Y}}$ in general. In this case, our goal is to identify the unitary matrices U and V and to select the smallest possible (in a sense to be clarified later on) effective dimensions $r \leq d$ and $s \leq m$ so that the Kullback-Leibler (KL) divergence from $\pi^*_{\mathbf{X}|\mathbf{Y}}$ to $\pi_{\mathbf{X}|\mathbf{Y}}$ is controlled in expectation over the data. That is,

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] \leq \epsilon, \tag{7.5}$$

for some prescribed tolerance $\epsilon > 0$. Let us mention that [52] also uses the expected KL divergence to reduce the dimension of $\mathbf{X}$, but not the one of $\mathbf{Y}$. The following proposition shows that given the decompositions of the parameters and observations in (7.2) and (7.3), the posterior approximation $\pi^*_{\mathbf{X}|\mathbf{Y}}$ in (7.4) is optimal for the expected KL divergence.

Proposition 14. *Let* $(\mathbf{X}, \mathbf{Y}) \sim \pi_{\mathbf{X},\mathbf{Y}}$ *be decomposed as in* (7.2) *and* (7.3). *Then the posterior approximation* $\pi^*_{\mathbf{X}|\mathbf{Y}}$ *defined in* (7.4) *satisfies*

$$\mathbb{E}\left[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] \leq \mathbb{E}\left[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right], \tag{7.6}$$

for any posterior approximation of the form $\widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) = f_1(\boldsymbol{x}_r, \boldsymbol{y}_s) f_2(\boldsymbol{x}_\perp, \boldsymbol{x}_r)$.

Proof. See Appendix E. □

Thus, once the matrices U, V are identified and the effective dimensions r, s determined, the optimal posterior approximation (7.4) is given by

$$\pi^*_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) \propto \pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)\pi_{\mathbf{X}}(\boldsymbol{x}),$$

where the marginal likelihood $\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)$ is accessible by marginalizing the likelihood function $\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})$ over $\boldsymbol{y}_\perp$ and $\boldsymbol{x}_\perp$ using the prior weight, i.e.,

$$\begin{aligned}\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r) &= \int_{\mathbb{R}^{m-s}} \int_{\mathbb{R}^{d-r}} \pi_{\mathbf{Y}|\mathbf{X}}(V_s\boldsymbol{y}_s + V_\perp\boldsymbol{y}_\perp|U_r\boldsymbol{x}_r + U_\perp\boldsymbol{x}_\perp)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\boldsymbol{x}_\perp|\boldsymbol{x}_r)\mathrm{d}\boldsymbol{x}_\perp\mathrm{d}\boldsymbol{y}_\perp \\ &= \frac{1}{\pi_{\mathbf{X}_r}(\boldsymbol{x}_r)} \int_{\mathbb{R}^{m-s}} \int_{\mathbb{R}^{d-r}} \pi_{\mathbf{X},\mathbf{Y}}(U_r\boldsymbol{x}_r + U_\perp\mathbf{X}_\perp, V_s\boldsymbol{y}_s + V_\perp\boldsymbol{y}_\perp)\mathrm{d}\boldsymbol{x}_\perp\mathrm{d}\boldsymbol{y}_\perp.\end{aligned} \tag{7.7}$$

Let us remark that with $s = m$ (*i.e.,* without reducing the observations), the approximate likelihood $\pi_{\mathbf{Y}_s|\mathbf{X}_r} = \pi_{\mathbf{Y}|\mathbf{X}_r}$ is the same as the one used in [239, 238] for reducing the dimension of the parameters.

The remainder of this section is organized as follows. In subsection 7.2.1 we provide a tractable upper bound for the posterior approximation in (7.4) that depends on the decompositions of $\mathbf{X}$ and $\mathbf{Y}$. Then, subsection 7.2.2 provides two methods for identifying the low-dimensional decompositions that minimize this upper bound, and subsection 7.2.3 presents a procedure for selecting the reduced dimensions r, s that satisfy the constraint in (7.5).

7.2.1 Gradient-based bound on expected KL

In this section we present our main result which consists of a gradient-based bound on the expected KL divergence in (7.5). This bound will guide the construction of the matrices U and V. We begin by presenting an assumption that is needed for our main result. In the following, $\|\cdot\|$ denotes the canonical norm of the Euclidean space.

Definition 2 (Logarithmic Sobolev inequality)**.** *A random variable* $\mathbf{Z}$ *with density* $\pi_{\mathbf{Z}}$ *on* $\mathbb{R}^p$ *satisfies the* logarithmic Sobolev inequality *if there exists a constant* $C < \infty$ *such that*

$$\int h(\boldsymbol{z}) \log\left(\frac{h(\boldsymbol{z})}{\int h \mathrm{d}\pi_{\mathbf{Z}}}\right) \pi_{\mathbf{Z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \leq \frac{C}{2} \int \|\nabla h(\boldsymbol{z})\|^2 h(\boldsymbol{z})\, \pi_{\mathbf{Z}}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}, \tag{7.8}$$

holds for any smooth function $h : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$*. The smallest constant* $C = C(\pi_{\mathbf{Z}})$ *such that* (7.8) *holds is called the* logarithmic Sobolev constant *of* $\mathbf{Z}$*.*

Definition 3 (Subspace logarithmic Sobolev inequality)**.** *A random variable* $\mathbf{Z}$ *with density* $\pi_{\mathbf{Z}}$ *on* $\mathbb{R}^p$ *satisfies the* subspace logarithmic Sobolev inequality *if there exists a constant* $\overline{C} < \infty$ *such that for any unitary matrix* $W \in \mathbb{R}^{p\times p}$ *and for any block decomposition* $W = [W_t, W_\perp]$ *with* $W_t \in \mathbb{R}^{p\times t}, t \leq p$*, and for any* $\boldsymbol{z}_\perp \in \mathbb{R}^{p-t}$*, the conditional random vector* $\mathbf{Z}_r|\mathbf{Z}_\perp = \boldsymbol{z}_\perp$ *with* $\mathbf{Z}_r = W_\perp^T\mathbf{Z}$ *and* $\mathbf{Z}_\perp = W_\perp^T\mathbf{Z}$ *satisfies the logarithmic Sobolev inequality with*

$$C(\pi_{\mathbf{Z}_r|\mathbf{Z}_\perp=\boldsymbol{z}_\perp}) \leq \overline{C}. \tag{7.9}$$

The smallest constant $\overline{C} = \overline{C}(\pi_{\mathbf{Z}})$ *such that* (7.9) *holds is called the* subspace logarithmic Sobolev constant *of* $\mathbf{Z}$*.*

Theorem 7.2.1. *Let* $(\mathbf{X}, \mathbf{Y})$ *be a random vector in* $\mathbb{R}^d \times \mathbb{R}^m$ *which satisfies the subspace logarithmic Sobolev inequality with constant* $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$*. Then for any unitary matrices* $U = [U_r, U_\perp] \in \mathbb{R}^{d\times d}$ *and* $V = [V_s, V_\perp]$ *we have*

$$\mathbb{E}\left[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot, Y)||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot, Y))\right] \leq \overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2 \left(\mathrm{Trace}(U_\perp^T H_{\mathbf{X}} U_\perp) + \mathrm{Trace}(V_\perp^T H_{\mathbf{Y}} V_\perp)\right). \tag{7.10}$$

Here, $\pi^*_{\mathbf{X}|\mathbf{Y}}$ is as in (7.4) and the matrices $H_{\mathbf{X}} \in \mathbb{R}^{d\times d}$ and $H_{\mathbf{Y}} \in \mathbb{R}^{m\times m}$ are given by

$$H_{\mathbf{X}} = \int \big(\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\big)^T \big(\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\big)\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \tag{7.11}$$

$$H_{\mathbf{Y}} = \int \big(\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\big) \big(\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\big)^T \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}, \tag{7.12}$$

where the matrix $\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(y|x) \in \mathbb{R}^{m\times d}$ is defined by

$$\Big(\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\Big)_{i,j} = \partial_{x_j}\partial_{y_i} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}).$$

Proof. See Section 7.3. □

In the rest of this chapter, we exploit the bound in (7.10) to find structured unitary matrices U, V which minimize the right-hand side of (7.10). Due to their central role, the matrices $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ are called the *diagnostic matrices*.

Before going further, let us comment on the assumption $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}}) < \infty$. As shown in [239], a sufficient condition for a density $\pi_{\mathbf{Z}}$ to satisfy the conditional logarithmic Sobolev inequality is that it has a convex support and it can be written as $\pi_{\mathbf{Z}}(\boldsymbol{z}) = \exp(-V(\boldsymbol{z}) - \Psi(\boldsymbol{z}))$ where V is a smooth convex function such that $\nabla^2 V(\boldsymbol{z}) \succeq \rho I$ for some $\rho > 0$, and where Ψ is a function with bounded oscillation with $\kappa = \sup \Psi - \inf \Psi < \infty$. Then, from the Bakry-Émery theorem [12] and the Holley-Stroock perturbation lemma [88], we obtain $\overline{C}(\pi_{\mathbf{Z}}) \leq \exp(\kappa)/\rho$. As shown in the following example, this condition is (trivially) satisfied when the joint density $\pi_{\mathbf{X},\mathbf{Y}}$ is Gaussian. We refer the reader to [239] for examples of (joint) densities which satisfy the subspace log-Sobolev inequality. In the general case, however, the constant $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$ won't be available in practice. In this case, we will still exploit the bound (7.10) without having access to $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$.

Example 6 (Gaussian joint density)**.** *Let $\pi_{\mathbf{X},\mathbf{Y}}$ be the joint density*

$$\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{y} - G\boldsymbol{x}\|^2 - \frac{1}{2}\|\boldsymbol{x}\|^2\right), \tag{7.13}$$

where $G \in \mathbb{R}^{m\times d}$. This corresponds to a Bayesian inverse problem with a standard

normal prior and a linear forward model $\boldsymbol{x} \mapsto G\boldsymbol{x}$ endowed with additive standard normal observational noise. Given that $\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \propto \exp(-V(\boldsymbol{x},\boldsymbol{y}))$ with the quadratic potential $V(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\binom{\boldsymbol{x}}{\boldsymbol{y}}^T \mathrm{Cov}\binom{\mathbf{X}}{\mathbf{Y}}^{-1}\binom{\boldsymbol{x}}{\boldsymbol{y}}$ where

$$\mathrm{Cov}\binom{\mathbf{X}}{\mathbf{Y}} = \begin{bmatrix} I_d + G^T G & -G^T \\ -G & I_m \end{bmatrix}^{-1} = \begin{bmatrix} I_d & G^T \\ G & GG^T + I_m \end{bmatrix},$$

we deduce (see [239]) that $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$ is bounded by $\lambda_{\max}(\mathrm{Cov}\binom{\mathbf{X}}{\mathbf{Y}})$, the largest eigenvalue of the joint covariance matrix $\mathrm{Cov}\binom{\mathbf{X}}{\mathbf{Y}}$. As shown in Appendix F, $\lambda_{\max}(\mathrm{Cov}\binom{\mathbf{X}}{\mathbf{Y}})$ can be analytically computed so that we obtain

$$\overline{C}(\pi_{\mathbf{X},\mathbf{Y}}) \leq \frac{1}{2}\left(2 + \sigma_{\max}(G)^2 + \sigma_{\max}(G)\sqrt{\sigma_{\max}(G)^2 + 4}\right), \tag{7.14}$$

where $\sigma_{\max}(G)$ the maximum singular value of G. Furthermore, the diagnostic matrices $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ in (7.11) and (7.12) are given by $H_{\mathbf{X}} = G^T G$ and $H_{\mathbf{Y}} = GG^T$.

Example 7 (Gaussian Likelihood, Gaussian prior)**.** *Continuing the previous example, we now assume that the forward model $G\colon \mathbb{R}^d \to \mathbb{R}^m$ is nonlinear so that the joint density is given by*

$$\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{y} - G(\boldsymbol{x})\|^2 - \frac{1}{2}\|\boldsymbol{x}\|^2\right).$$

Denoting the Jacobian of the forward model by $\nabla G(\boldsymbol{x}) \in \mathbb{R}^{m\times d}$, we can write

$$-\nabla^2 \log \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) = \begin{bmatrix} I + \nabla G(\boldsymbol{x})^T \nabla G(\boldsymbol{x}) & -\nabla G(\boldsymbol{x})^T \\ -\nabla G(\boldsymbol{x}) & I \end{bmatrix} - \begin{bmatrix} A(\boldsymbol{x},\boldsymbol{y}) & 0 \\ 0 & 0 \end{bmatrix},$$

where the matrix $A(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^{d\times d}$ is given by $(A(\boldsymbol{x},\boldsymbol{y}))_{i,j} = \sum_{k=1}^n \partial^2_{i,j} G_k(\boldsymbol{x})(\boldsymbol{y} - G(\boldsymbol{x}))_k$. As in the previous example, we have

$$\lambda_{\min}\left(\begin{bmatrix} I + \nabla G(\boldsymbol{x})^T \nabla G(\boldsymbol{x}) & -\nabla G(\boldsymbol{x})^T \\ -\nabla G(\boldsymbol{x}) & I \end{bmatrix}\right) = \lambda(\boldsymbol{x})^{-1},$$

where $\lambda(\boldsymbol{x}) = \frac{1}{2}(2 + \sigma_{\max}(\nabla G(\boldsymbol{x}))^2 + \sigma_{\max}(\nabla G(\boldsymbol{x}))\sqrt{\sigma_{\max}(\nabla G(\boldsymbol{x}))^2 + 4})$ *so that*

$$-\nabla^2 \log \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x}, \boldsymbol{y}) \succeq (\lambda(\boldsymbol{x})^{-1} - \lambda_{\max}(A(\boldsymbol{x}, \boldsymbol{y}))) \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Therefore, if there exists a constant $C < \infty$ *such that* $\lambda(\boldsymbol{x})^{-1} - \lambda_{\max}(A(\boldsymbol{x}, \boldsymbol{y})) \geq 1/C$ *uniformly over* $\boldsymbol{x}, \boldsymbol{y}$*, then* $\pi_{\mathbf{X},\mathbf{Y}}$ *satisfies the subspace log-Sobolev inequality with* $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}}) \leq C$*. Furthermore, since* $\nabla_{\mathbf{X}} \nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}) = \nabla G(\boldsymbol{x}) \in \mathbb{R}^{m \times d}$*, the diagnostic matrices are given by*

$$H_{\mathbf{X}} = \int \nabla G(\boldsymbol{x})^T \nabla G(\boldsymbol{x}) \pi_{\mathbf{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$
$$H_{\mathbf{Y}} = \int \nabla G(\boldsymbol{x}) \nabla G(\boldsymbol{x})^T \pi_{\mathbf{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

7.2.2 Constructing U, V by minimizing the upper bound

In this section we propose two different approaches to build the unitary matrices $U = [U_r, U_\perp]$ and $V = [V_s, V_\perp]$. We assume the reduced dimensions r, s are prescribed; the discussion on how to select r, s is postponed to Section 7.2.3. The first approach, referred to as *optimal rotations*, consists in minimizing the upper bound (7.10) by solving

$$\min_{U_\perp, V_\perp} \mathrm{Trace}(U_\perp^T H_{\mathbf{X}} U_\perp) + \mathrm{Trace}(V_\perp^T H_{\mathbf{Y}} V_\perp), \tag{7.15}$$

subject to $U_\perp^T U_\perp = \mathrm{I}_{d-r}$ and $V_\perp^T V_\perp = \mathrm{I}_{m-s}$. The second approach, referred to as *optimal permutations*, consists in solving (7.15) with the additional constrain that U, V are permutation matrices. That way, $\mathbf{X}_r = U_r^T \mathbf{X}$ and $\mathbf{Y}_s = V_s^T \mathbf{Y}$ contain a subset of coordinates of $\mathbf{X}$ and $\mathbf{Y}$ and hence the dimension reduction corresponds to a coordinate selection. As compared to the former approach, the latter preserves the interpretability of the components.

In both approaches, the optimal solutions U, V are independent of the reduced dimensions r, s. More specifically, there exists matrices U, V (independent of r, s) such that the solution to (7.15) can be extracted from the last columns of U, V for

any reduced dimensions r, s.

Optimal rotations

Let us recall Corollary 4.3.39 in [89] for the variational characterization of eigenvalues of Hermitian matrices.

Proposition 15. *Let $H \in \mathbb{R}^{p\times p}$ be a symmetric positive definite matrix with eigenpairs $(\lambda_i, w_i) \in \mathbb{R}_{>0} \times \mathbb{R}^p$, meaning $Hw_i = \lambda_i w_i$, where $\lambda_i \geq \lambda_{i+1}$ and $\|w_i\|_2 = 1$ for all i. Then, for any $t < p$ we have*

$$\min_{\substack{W_\perp \in \mathbb{R}^{p\times t} \\ W_\perp^T W_\perp = I_t}} \operatorname{Trace}(W_\perp^T H W_\perp) = \sum_{i=t+1}^{p} \lambda_i, \tag{7.16}$$

where the solution is given by $W_\perp = [w_{t+1}, \ldots, w_p]$.

Let $(\lambda_i(H_{\mathbf{X}}), u_i)$ and $(\lambda_i(H_{\mathbf{Y}}), v_i)$ denote the i-th largest eigenpairs of $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$, respectively. Then, Proposition 15 ensures that for any r, s

$$U_\perp = [u_{r+1}, \ldots, u_d]$$
$$V_\perp = [v_{s+1}, \ldots, v_m],$$

is the optimal solution to (7.15). This choice yields the optimal bound

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] \leq \overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2 \left(\sum_{i=r+1}^{d} \lambda_i(H_{\mathbf{X}}) + \sum_{i=s+1}^{m} \lambda_i(H_{\mathbf{Y}})\right). \tag{7.17}$$

The eigenvectors u_i, v_i can be precomputed without knowing r and s. The advantage of this is that to increase r or s, one only needs to compute the additional eigenvectors.

Remark. *In practice, it is sufficient to compute the matrices $U_r = [u_1, \ldots, u_r]$ and $V_s = [v_1, \ldots, v_s]$ to reduce the parameter and observation dimensions; see Section 7.6. The (possibly much larger) matrices $U_\perp$ and $V_\perp$ are never assembled in practice. Furthermore, the trailing eigenvalue sums on the right-hand side of (7.17) can be computed from the traces of the diagnostic matrices and their leading eigenvalues.*

Optimal permutations

We now constrain U, V to be permutation matrices so that $U\mathbf{X} = (\mathbf{X}_{\sigma_\mathbf{X}(1)}, \ldots, \mathbf{X}_{\sigma_\mathbf{X}(d)})$ and $V\mathbf{Y} = (\mathbf{Y}_{\sigma_\mathbf{Y}(1)}, \ldots, \mathbf{Y}_{\sigma_\mathbf{Y}(m)})$ where $\sigma_\mathbf{X}$ and $\sigma_\mathbf{Y}$ are permutations of $\{1, \ldots, d\}$ and $\{1, \ldots, m\}$, respectively. Then, (7.15) becomes

$$\min_{\sigma_\mathbf{X}, \sigma_\mathbf{Y}} \left(\sum_{i=r+1}^{d} (H_\mathbf{X})_{\sigma_\mathbf{X}(i), \sigma_\mathbf{X}(i)} + \sum_{i=s+1}^{m} (H_\mathbf{Y})_{\sigma_\mathbf{Y}(i), \sigma_\mathbf{Y}(i)} \right). \tag{7.18}$$

The optimal permutations $\sigma_\mathbf{X}$ and $\sigma_\mathbf{Y}$ are the ones which sort the diagonal terms of $H_\mathbf{X}$ and $H_\mathbf{Y}$ in decreasing order. That is,

$$\begin{aligned} (H_\mathbf{X})_{\sigma_\mathbf{X}(i), \sigma_\mathbf{X}(i)} &\geq (H_\mathbf{X})_{\sigma_\mathbf{X}(i+1), \sigma_\mathbf{X}(i+1)} \\ (H_\mathbf{Y})_{\sigma_\mathbf{Y}(i), \sigma_\mathbf{Y}(i)} &\geq (H_\mathbf{Y})_{\sigma_\mathbf{Y}(i+1), \sigma_\mathbf{Y}(i+1)}. \end{aligned}$$

This choice yields the upper bound

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}) || \pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})) \right] \leq \overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2 \left(\sum_{i=r+1}^{d} (H_\mathbf{X})_{\sigma_\mathbf{X}(i), \sigma_\mathbf{X}(i)} + \sum_{i=s+1}^{m} (H_\mathbf{Y})_{\sigma_\mathbf{Y}(i), \sigma_\mathbf{Y}(i)} \right). \tag{7.19}$$

Let us remark that because permutation matrices are unitary matrices, the bound in (7.19) is larger than or equal to the optimal bound in (7.17). Thus, the optimal permutation approach might be less efficient as compared to the optimal rotation approach. The trade-off is that it preserves the interpretability of the reduced components.

7.2.3 Selecting the reduced dimensions

We now discuss the problem of selecting the reduced dimensions. We propose to select r and s by minimizing the computational cost for exploring the reduced posterior $\pi^*_{\mathbf{X}|\mathbf{Y}}$ under the constraint the $\pi^*_{\mathbf{X}|\mathbf{Y}}$ is sufficiently accurate.

Let $c(r, s) \geq 0$ be a cost function that reflects the computational complexity of solving the reduced Bayesian inference problem with the posterior density $\pi^*_{\mathbf{X}|\mathbf{Y}}$

in (7.4). The choice of $c(r,s)$ strongly depends on the inference method (e.g., MCMC, ABC, transport maps, etc.). For instance, we may have $c(r,s) = \alpha_{\mathbf{X}} r + \alpha_{\mathbf{Y}} s$ or $c(r,s) = \alpha_{\mathbf{X}} r^2 + \alpha_{\mathbf{Y}} s^2$ for some weights $\alpha_{\mathbf{X}}, \alpha_{\mathbf{Y}} \geq 0$ which prescribe the relative complexity of reducing the parameters or the observations. Given a prescribed tolerance ϵ, the ideal way to select r, s is to solve

$$\min_{r,s} c(r,s) \qquad \text{s.t.} \qquad \mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})\right] \leq \epsilon. \tag{7.20}$$

In practice, the expected KL divergence is not accessible for general inverse problems. Instead, we select r, s by solving

$$\min_{r,s} c(r,s) \qquad \text{s.t.} \qquad B(r,s) \leq \epsilon', \tag{7.21}$$

where $B(r,s)$ is defined by either

$$B(r,s) = \sum_{i=r+1}^{d} \lambda_i(H_{\mathbf{X}}) + \sum_{i=s+1}^{m} \lambda_i(H_{\mathbf{Y}}),$$

$$\text{or} \qquad B(r,s) = \sum_{i=r+1}^{d} (H_{\mathbf{X}})_{\sigma_{\mathbf{X}}(i),\sigma_{\mathbf{X}}(i)} + \sum_{i=s+1}^{m} (H_{\mathbf{Y}})_{\sigma_{\mathbf{Y}}(i),\sigma_{\mathbf{Y}}(i)},$$

depending on whether one uses the optimal rotations (see Section 7.2.2) or the optimal permutations (see Section 7.2.2) to build U, V. Given that

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})\right] \leq \overline{C}(\pi_{\mathbf{X},\mathbf{Y}}) B(r,s),$$

the solution to (7.21) with $\epsilon' = \epsilon/\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$ provides a feasible approximate solution to (7.20). In the case where the log-Sobolev constant is not known, we propose to select r, s by solving (7.21) with $\epsilon' = \epsilon$. While the resulting solution may not satisfy $\mathbb{E}[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})] \leq \epsilon$, it still provides a good heuristic for selecting the reduced dimensions, as illustrated in Section 7.7.

Remark. *When $c(r,s) = \alpha_{\mathbf{X}} r + \alpha_{\mathbf{Y}} s$, the optimization problem in (7.21) can be formulated as a $0-1$ knapsack problem, which is known to be NP-complete [108].*

Given that (r, s) is only two-dimensional, we can often enumerate all combinations of the reduced dimensions to find the optimal solutions. Alternatively, an approximate solution that doesn't require enumeration is to split the constraint and to select r, s individually based on a weighted error tolerance for the parameters and observations. For example, we can identify r, s by finding the smallest integers that meet the constraints

$$\sum_{i=1}^{r} \lambda_i(H_{\mathbf{X}}) \leq \frac{\alpha_{\mathbf{X}}}{\alpha_{\mathbf{X}} + \alpha_{\mathbf{Y}}}\epsilon, \qquad \sum_{i=1}^{s} \lambda_i(H_{\mathbf{Y}}) \leq \frac{\alpha_{\mathbf{Y}}}{\alpha_{\mathbf{X}} + \alpha_{\mathbf{Y}}}\epsilon.$$

The setting $\alpha_{\mathbf{X}} = \alpha_{\mathbf{Y}}$ corresponds to choosing the reduced dimensions such that the two errors from reducing the parameters and the observations are balanced.

7.3 Information theory and proof of Theorem 7.2.1

In this section we relate the posterior approximation error to information-theoretic terms that quantify the conditional independence between random variables. We then show how to bound these quantities to derive the upper bound in Theorem 7.2.1.

We begin by defining mutual information and conditional mutual information, that are two well known measures for the strength of independence and conditional independence between random variables, respectively.

Definition 4. *Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables with joint density $\pi_{\mathbf{X},\mathbf{Y}}$. The mutual information between $\mathbf{X}$ and $\mathbf{Y}$ is given by*

$$I(\mathbf{X}; \mathbf{Y}) := \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x}, \boldsymbol{y}) \log\left(\frac{\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x}, \boldsymbol{y})}{\pi_{\mathbf{X}}(\boldsymbol{x})\pi_{\mathbf{Y}}(\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

where $\pi_{\mathbf{Y}}(\boldsymbol{y}) = \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{x}$ and $\pi_{\mathbf{X}}(\boldsymbol{x}) = \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{y}$.

The mutual information is equivalently expressed as the KL divergence from the product of the marginal densities to the joint probability density function, i.e., $I(\mathbf{X}; \mathbf{Y}) = D_{\mathrm{KL}}(\pi_{\mathbf{X},\mathbf{Y}} || \pi_{\mathbf{X}}\pi_{\mathbf{Y}})$. The mutual information measures the dependence of $\mathbf{X}$ and $\mathbf{Y}$. In particular, $I(\mathbf{X}; \mathbf{Y}) = 0$ yields $\pi_{\mathbf{X},\mathbf{Y}} = \pi_{\mathbf{X}}\pi_{\mathbf{Y}}$, meaning that $\mathbf{X}$ and $\mathbf{Y}$ are independent.

Definition 5. *The conditional mutual information between random variables* $\mathbf{X}$ *and* $\mathbf{Y}$ *given a* third *random variable* $\mathbf{Z}$ *with joint density* $\pi_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ *is given by*

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) \coloneqq \int \pi_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) \log\left(\frac{\pi_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})}{\pi_{\mathbf{X}|\mathbf{Z}}(\boldsymbol{x}|\boldsymbol{z})\pi_{\mathbf{Y}|\mathbf{Z}}(\boldsymbol{y}|\boldsymbol{z})}\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}\mathrm{d}\boldsymbol{z},$$

where $\pi_{\mathbf{X}|\mathbf{Z}}(\boldsymbol{x}|\boldsymbol{z}) = \int \pi_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})\mathrm{d}\boldsymbol{y}$ *and* $\pi_{\mathbf{Y}|\mathbf{Z}}(\boldsymbol{y}|\boldsymbol{z}) = \int \pi_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\boldsymbol{x},\boldsymbol{y}|\boldsymbol{z})\mathrm{d}\boldsymbol{x}$.

Analogously, the conditional mutual information $I(\mathbf{X};\mathbf{Y}|\mathbf{Z})$ is defined as the KL divergence from the product $\pi_{\mathbf{Y}|\mathbf{Z}}\pi_{\mathbf{X}|\mathbf{Z}}$ to the conditional density $\pi_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}$ in expectation over $\mathbf{Z}$, i.e., $I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = \mathbb{E}[D_{\mathrm{KL}}(\pi_{\mathbf{X},\mathbf{Y}|\mathbf{Z}}(\cdot,\cdot|\mathbf{Z})||\pi_{\mathbf{X}|\mathbf{Z}}(\cdot|\mathbf{Z})\pi_{\mathbf{Y}|\mathbf{Z}}(\cdot|\mathbf{Z}))]$. The conditional mutual information serves as a measure of conditional independence between random variables, i.e., $I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) = 0$ if and only if $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$. We refer the reader to Chapters 4 and 5 for more comprehensive discussion on estimators for the conditional mutual information and its connection to another measure of conditional independence.

The following proposition shows that the expected KL divergence from the optimal posterior approximation to the true posterior distribution is related to a difference of conditional mutual information terms.

Proposition 16. *Let* $\pi_{\mathbf{X}|\mathbf{Y}}$ *be the posterior density for* $\mathbf{X}|\mathbf{Y}$ *and* $\pi^*_{\mathbf{X}|\mathbf{Y}}$ *be the optimal posterior approximation in* (7.4) *with* r*-dimensional informed parameters and* s*-dimensional informative observations. Then, we have*

$$\mathbb{E}\left[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] = I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X}_r;\mathbf{Y}_s) \tag{7.22}$$

$$= I(\mathbf{X}_\perp;\mathbf{Y}|\mathbf{X}_r) + I(\mathbf{X};\mathbf{Y}_\perp|\mathbf{Y}_s) - I(\mathbf{X}_\perp;\mathbf{Y}_\perp|\mathbf{X}_r,\mathbf{Y}_s). \tag{7.23}$$

Proof. See Appendix E. □

Remark. *The right-hand sides of* (7.22) *and of* (7.23) *simplify if we only consider reducing either the dimension of the parameters or the observations alone. For in-*

stance, if the observations are not reduced, i.e., $\mathbf{Y}_s = \mathbf{Y}$*, we have*

$$\mathbb{E}[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))] = I(\mathbf{X}_\perp; \mathbf{Y}|\mathbf{X}_r).$$

Analogously, if the parameters are not reduced, i.e., $\mathbf{X}_r = \mathbf{X}$*, we have*

$$\mathbb{E}[D_{KL}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))] = I(\mathbf{X}; \mathbf{Y}_\perp|\mathbf{Y}_s).$$

Remark. *An important property of mutual information is that it is invariant to invertible marginal transformations of the variables. For instance, by applying the linear transformations* $\overline{\mathbf{X}} = A\mathbf{X}$ *and* $\overline{\mathbf{Y}} = B\mathbf{Y}$ *for some invertible matrices* $A \in \mathbb{R}^{d\times d}$ *and* $B \in \mathbb{R}^{m\times m}$*, we have* $I(\mathbf{X};\mathbf{Y}) = I(\overline{\mathbf{X}};\overline{\mathbf{Y}})$ *and* $I(\mathbf{X}_r;\mathbf{Y}_r) = I(\overline{\mathbf{X}}_r;\overline{\mathbf{Y}}_s)$*, but also* $I(\mathbf{X}_\perp;\mathbf{Y}|\mathbf{X}_r) = I(\overline{\mathbf{X}}_\perp;\overline{\mathbf{Y}}|\overline{\mathbf{X}}_r)$ *and* $I(\mathbf{X};\mathbf{Y}_\perp|\mathbf{Y}_s) = I(\overline{\mathbf{X}};\overline{\mathbf{Y}}_\perp|\overline{\mathbf{Y}}_s)$*.*

While the (conditional) mutual information is tractable to compute for Gaussians and certain classes of parametric distributions, it does not admit a closed-form expression for arbitrary non-Gaussian distributions. For a density that satisfies the subspace log-Sobolev inequality in (7.9), the following proposition provides an upper bound for the conditional mutual information based on integrated mixed partial derivatives of the log-density.

Proposition 17. *Let* $\pi_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ *be a joint density of random variables* $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ *that satisfies the subspace logarithmic Sobolev inequality with constant* $\overline{C}(\pi_{\mathbf{X},\mathbf{Y},\mathbf{Z}})$*. Then, the conditional mutual information is upper bounded by*

$$I(\mathbf{X};\mathbf{Y}|\mathbf{Z}) \leq \overline{C}(\pi_{\mathbf{X},\mathbf{Y},\mathbf{Z}})^2 \mathbb{E}\|\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{X},\mathbf{Y},\mathbf{Z}}(\mathbf{X},\mathbf{Y},\mathbf{Z})\|_F^2, \tag{7.24}$$

where $\|\cdot\|_F$ *denotes the Frobenius norm.*

Proof. The proof follows from Theorem 4.2.1 in Chapter 4. It is given in Appendix C. □

Collecting the results in Propositions 16 and 17, we now give the proof of Theorem 7.2.1.

Proof of Theorem 7.2.1. Given that the conditional mutual information is positive, Proposition 16 permits us to write

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] \leq I(\mathbf{X}_\perp;\mathbf{Y}|\mathbf{X}_r) + I(\mathbf{X};\mathbf{Y}_\perp|\mathbf{Y}_s).$$

By Proposition 17, each term in the right-hand side above is upper bounded by the expectation of mixed partial derivatives of the log-density $\pi_{\mathbf{X},\mathbf{Y}}$ as

$$\begin{aligned}
&\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] \\
&\quad\leq \overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2\left(\mathbb{E}\|\nabla_{\mathbf{X}_\perp}\nabla_{\mathbf{Y}}\log\pi_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y})\|_F^2 + \mathbb{E}\|\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}_\perp}\log\pi_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y})\|_F^2\right) \\
&\quad= \overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2\left(\mathbb{E}\|U_\perp\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\log\pi_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y})\|_F^2 + \mathbb{E}\|\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\log\pi_{\mathbf{X},\mathbf{Y}}(\mathbf{X},\mathbf{Y})V_\perp\|_F^2\right) \\
&\quad= \overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2\left(\mathbb{E}\|U_\perp\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\log\pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})\|_F^2 + \mathbb{E}\|\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}}\log\pi_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y}|\mathbf{X})V_\perp\|_F^2\right).
\end{aligned}$$

Expanding the Frobenius norm using the trace, we arrive at equation (7.10). □

7.4 Gaussian likelihood models

In this section we specialize the bound in Theorem 7.2.1 to forward models with additive Gaussian observational noise. Let $\mathbf{Y} = G(\mathbf{X}) + \boldsymbol{\mathcal{E}}$, where $G\colon \mathbb{R}^d \to \mathbb{R}^m$ is a (nonlinear) forward model and $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, \Gamma_{\mathrm{obs}})$ is a Gaussian observational error which is independent of $\mathbf{X}$. This process corresponds to a Gaussian likelihood $\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}) \propto \exp(-\frac{1}{2}\|\boldsymbol{y} - G(\boldsymbol{x})\|^2_{\Gamma^{-1}_{\mathrm{obs}}})$ and a joint density of the form

$$\pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\|\boldsymbol{y} - G(\boldsymbol{x})\|^2_{\Gamma^{-1}_{\mathrm{obs}}}\right)\pi_{\mathbf{X}}(\boldsymbol{x}), \tag{7.25}$$

where $\pi_{\mathbf{X}}$ is any prior density. Without further assumptions, the subspace log-Sobolev constant $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$ remains unknown.

7.4.1 Whitening

Next, we propose a change of variable for $\mathbf{X}$ and $\mathbf{Y}$ which can be interpreted as a preconditioning for the dimension reduction procedure. Notice that with the change of variables $\overline{\mathbf{X}} = A\mathbf{X}$ and $\overline{\mathbf{Y}} = B\mathbf{Y}$, the left-hand side of (7.10) remains unchanged (see Section 7.3) while the right-hand side is modified in several ways through the subspace log-Sobolev constant and the diagnostic matrices. Finding the optimal change of variables which minimizes the left-hand side of (7.10) is a difficult task, mostly because the subspace log-Sobolev constant $\overline{C}(\pi_{\mathbf{X}_A,\mathbf{Y}_B})$ is not readily available. Instead, we propose a heuristic which consists in whitening the parameters and observations as follows

$$\overline{\mathbf{X}} = \Gamma_{\text{pr}}^{-1/2}\mathbf{X} \quad \text{and} \quad \overline{\mathbf{Y}} = \Gamma_{\text{obs}}^{-1/2}\mathbf{Y}, \tag{7.26}$$

where $\Gamma_{\text{pr}} = \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T$ is the prior covariance, assuming it exists. Then, we reduce the dimensions of $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ using the corresponding diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ which, using (7.25) and (7.26), are given by

$$H_{\overline{\mathbf{X}}} = \Gamma_{\text{pr}}^{1/2}\left(\int \nabla G(\boldsymbol{x})^T\Gamma_{\text{obs}}^{-1}\nabla G(\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\Gamma_{\text{pr}}^{1/2} \tag{7.27}$$

$$H_{\overline{\mathbf{Y}}} = \Gamma_{\text{obs}}^{-1/2}\left(\int \nabla G(\boldsymbol{x})\Gamma_{\text{pr}}\nabla G(\boldsymbol{x})^T\pi_{\mathbf{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\Gamma_{\text{obs}}^{-1/2}. \tag{7.28}$$

Denoting the matrices containing the first eigenvectors of $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ as $\overline{U}_r = [\overline{u}_1, \dots, \overline{u}_r]$ and $\overline{V}_s = [\overline{v}_1, \dots, \overline{v}_s]$, respectively, the reduced parameters and observations are $\mathbf{X}_r = U_r^T\overline{\mathbf{X}}$ and $\mathbf{Y}_s = V_s^T\overline{\mathbf{Y}}$ which, using (7.26), are given by

$$\mathbf{X}_r = U_r^T\mathbf{X} \qquad \text{where} \qquad U_r = \Gamma_{\text{pr}}^{-1/2}\overline{U}_r \tag{7.29}$$

$$\mathbf{Y}_s = V_s^T\mathbf{Y} \qquad \text{where} \qquad V_s = \Gamma_{\text{obs}}^{-1/2}\overline{V}_s. \tag{7.30}$$

With the above definitions, the matrices U_r and V_s no longer have orthogonal columns in the Euclidean sense, but they satisfy $U_r^T\Gamma_{\text{pr}}U_r = \mathrm{I}_r$ and $V_s^T\Gamma_{\text{obs}}V_s = \mathrm{I}_s$.

Remark (Generalized eigenvalue problems)**.** *Let $u_i = \Gamma_{pr}^{-1/2}\overline{u}_i$ and $v_i = \Gamma_{obs}^{-1/2}\overline{v}_i$ denote the i-th columns of U_r and V_s, as defined in (7.29) and (7.30). These column*

vectors can be seen as the eigenvectors of the generalized eigenvalue problems

$$\mathcal{H}_{\overline{\mathbf{X}}} w_i = \lambda_i(\mathcal{H}_{\overline{\mathbf{X}}})\Gamma_{pr}^{-1} w_i, \qquad u_i = \Gamma_{pr}^{-1} w_i, \tag{7.31}$$

$$\mathcal{H}_{\overline{\mathbf{Y}}} v_i = \lambda_i(\mathcal{H}_{\overline{\mathbf{Y}}})\Gamma_{obs} v_i, \tag{7.32}$$

where

$$\mathcal{H}_{\overline{\mathbf{X}}} = \int \nabla G(\boldsymbol{x})^T \Gamma_{obs}^{-1} \nabla G(\boldsymbol{x}) \pi_{\mathbf{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \qquad \textit{and} \qquad \mathcal{H}_{\overline{\mathbf{Y}}} = \int \nabla G(\boldsymbol{x}) \Gamma_{pr} \nabla G(\boldsymbol{x})^T \pi_{\mathbf{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

We note that $\mathcal{H}_{\overline{\mathbf{X}}}$ is the same diagnostic matrix as the one introduced in [52, Section 4]. The diagnostic matrix $\mathcal{H}_{\overline{\mathbf{X}}}$ is also similar to the one proposed in [50] for finding the likelihood informed subspace, with the exception that is integrates with respect to the prior instead of the posterior density.

Remark. *Rather than the decompositions in (7.2) and (7.3), the proposed change of variables in (7.26) yields a decomposition of* $\mathbf{X}$ *and* $\mathbf{Y}$ *of the form*

$$\mathbf{X} = \Gamma_{pr}^{1/2}\left(\overline{U}_r \mathbf{X}_r + \overline{U}_\perp \mathbf{X}_\perp\right) \qquad \textit{where} \begin{cases} \mathbf{X}_r = \overline{U}_r^T \Gamma_{pr}^{-1/2} \mathbf{X} \\ \mathbf{X}_\perp = \overline{U}_\perp^T \Gamma_{pr}^{-1/2} \mathbf{X} \end{cases},$$

$$\mathbf{Y} = \Gamma_{obs}^{1/2}\left(\overline{V}_s \mathbf{Y}_s + \overline{V}_\perp \mathbf{Y}_\perp\right) \qquad \textit{where} \begin{cases} \mathbf{Y}_s = \overline{V}_s^T \Gamma_{obs}^{-1/2} \mathbf{Y} \\ \mathbf{Y}_\perp = \overline{V}_\perp^T \Gamma_{obs}^{-1/2} \mathbf{Y} \end{cases}.$$

7.4.2 Linear-Gaussian setting

We now further assume that the forward model $\boldsymbol{x} \mapsto G\boldsymbol{x}$ is linear where $G \in \mathbb{R}^{m\times d}$ is a matrix. In this case, the diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ are given by

$$H_{\overline{\mathbf{X}}} = (\Gamma_{\text{pr}}^{1/2} G^T \Gamma_{\text{obs}}^{-1/2})(\Gamma_{\text{obs}}^{-1/2} G \Gamma_{\text{pr}}^{1/2})$$
$$H_{\overline{\mathbf{Y}}} = (\Gamma_{\text{obs}}^{-1/2} G \Gamma_{\text{pr}}^{1/2})(\Gamma_{\text{pr}}^{1/2} G^T \Gamma_{\text{obs}}^{-1/2}).$$

The eigendecompositions of $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ are obtained by computing the singular value decomposition (SVD) of the so-called whitened forward model

$$\Gamma_{\text{pr}}^{1/2} G^T \Gamma_{\text{obs}}^{-1/2} = \sum_{i=1}^{\min\{d,m\}} \sigma_i \overline{u}_i \overline{v}_i^T. \tag{7.33}$$

In particular, the generalized eigenvalues are the same, i.e., $\lambda_i(H_{\overline{\mathbf{X}}}) = \lambda_i(H_{\overline{\mathbf{Y}}}) = \sigma_i^2$.

The eigendecompositions of $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ (or equivalently the generalized eigendecompositions (7.31) and (7.32) of $\mathcal{H}_{\overline{\mathbf{X}}}$ and $\mathcal{H}_{\overline{\mathbf{Y}}}$ as in Remark 7.4.1) have already been used to reduce the dimensions of parameters and observations in linear-Gaussian inverse problems. [205] used (7.31) to approximate the posterior covariance as a low-rank update of the prior covariance. Furthermore, Algorithm 1 in [205] solved both eigenvalue problems (7.31) and (7.32) to derive an approximation to the posterior mean (that minimizes a weighted Bayes risk with squared error loss) as a linear projection of the observation $\mathbf{Y}$. [76] also showed the equivalence between the solution to (7.32) and finding the vectors that solve $\max_{V_s} I(V_s^T \mathbf{Y}, \mathbf{X})$, which Proposition 16 shows is equivalent to minimizing the expected KL divergence for the posterior with reduced observations. To minimize the expected KL divergence for linear inverse problems, the authors used Riemannian optimization algorithms to find the column vectors V_s in a Grassmannian manifold. This was extended to non-linear forward models by using a Laplace approximation of the posterior distribution. Lastly, [98] derives mutual information bounds for coordinate selection of observations in linear-Gaussian problems. These bounds are used to develop greedy algorithms with guarantees for solving optimization problems with cardinality constraints.

7.4.3 Gap in the linear-Gaussian setting

We now analyze the gap between the upper bound in (7.10) and the posterior approximation error for a linear-Gaussian likelihood model with a Gaussian prior.

We begin by computing the expected KL divergence and its upper bound in (7.10). We denote the i-th largest singular value of the whitened forward model $\Gamma_{\text{pr}}^{1/2} G^T \Gamma_{\text{obs}}^{-1/2}$

in (7.33) by σ_i. Using the closed-form expression for the conditional mutual information of Gaussian variables (see Appendix F), we have

$$\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})\right] \overset{(7.22)}{=} I(\mathbf{X},\mathbf{Y}) - I(\mathbf{X}_r,\mathbf{Y}_s) = \frac{1}{2}\sum_{i>\min\{r,s\}}^{p} \log(1+\sigma_i^2). \tag{7.34}$$

In comparison, the upper bound in (7.10) evaluated at the optimal rotations $U_\perp$ and $V_\perp$ is given by

$$\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2 \left(\sum_{i>r}^{p}\sigma_i^2 + \sum_{i>s}^{p}\sigma_i^2\right), \tag{7.35}$$

where the subspace log-Sobolev constant $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$ can be bounded in terms of $\sigma_1 = \sigma_{\max}(G)$; see Example 6. Using a first-order Taylor series expansion of $\log(1+\sigma^2)$ as $\sigma \to 0$, the ratio between (7.35) and (7.34) satisfies

$$\frac{\mathbb{E}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})\right]}{\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2\left(\sum_{i>r}^{p}\sigma_i^2 + \sum_{i>s}^{p}\sigma_i^2\right)} = \frac{1}{2\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2}\frac{\sum_{i>\min\{r,s\}}^{p}\sigma_i^2 + \mathcal{O}(\sigma_i^4)}{\left(\sum_{i>r}^{p}\sigma_i^2 + \sum_{i>s}^{p}\sigma_i^2\right)} \tag{7.36}$$

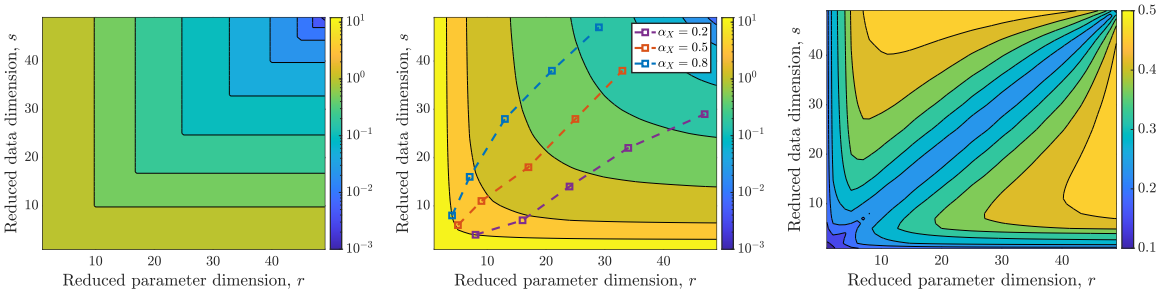
$$\overset{r=s}{=} \frac{1}{4\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2}(1+\mathcal{O}(\sigma_r^2)). \tag{7.37}$$

In the limit of $\sigma_r \to 0$, the above ratio converges to the constant $1/(4\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2)$. Thus, the expected KL and its bound are going to zero at the same rate. Let us remark that if either $r = m$ or $s = d$, i.e., when only the parameters or the observations are reduced, but not both, the ratio in (7.36) goes to $1/(2\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2)$.

As a numerical experiment, we consider the forward operator $G = I_d$ and set $m = d = 50$. Following Example 1 in [205], we construct the prior covariance as $\Gamma_{\mathrm{pr}} = WDW^T$ by randomly sampling a unitary matrix W for its singular vectors from the QR factorization of a matrix with standard Gaussian entries and defining a diagonal matrix D with decaying entries $D_{ii} = \lambda_0/i + \tau$ for its spectrum using $\lambda_0 = 1$ and $\tau = 10^{-6}$. We follow the same procedure to randomly sample the observation noise covariance Γ_{obs} with $\lambda_0 = 500$.

Figure 7-1a plots the expected KL divergence for each pair of reduced dimensions (r, s). Figure 7-1b plots the upper bound in (7.10), up to the log-Sobolev constant,

that we can tractably minimize to identify the subspaces for the parameters and observations. Figure 7-1c plots the ratio between the upper bound and the posterior approximation error. We note that these quantities are computed using the analytic expressions for the mutual information of Gaussian vectors, as above. For each tolerance ϵ, we observe the Pareto front of reduced dimensions that yield the same approximation error. The dashed lines in Figure 7-1b indicate the optimal reduced dimensions that solve (7.21) for five different tolerances and the linear cost function $c(r, s) = \alpha_{\mathbf{X}} r + \alpha_{\mathbf{Y}} s$ with different weights $\alpha_{\mathbf{X}}, \alpha_{\mathbf{Y}} \in \{0.2, 0.5, 0.8\}$ which trade-off the complexity of keeping the parameters versus the observations.



(a) (b) (c)

Figure 7-1: (a) The expected KL divergence for the posterior approximation error with reduced dimension r, s; (b) The tractable upper bound for the conditional mutual information terms; (c) The gap in the upper bound, up to the constant $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})^2$, which approaches 0.25 for $r = s$ and 0.5 for $r = d$ or $s = m$.

7.5 Comparisons to PCA and CCA

Two popular methods for linear dimension reduction are principal component analysis (PCA) and canonical correlation analysis (CCA).

PCA consists in reducing the dimension of a mean-zero random vector $\mathbf{X}$ by minimizing the L^2 error $\mathbb{E}[\|\mathbf{X} - U_r U_r^T \mathbf{X}\|^2]$ over the matrix $U_r \in \mathbb{R}^{d \times r}$ with orthogonal columns [90, 103]. The solution is $U_r = [u_1^{\text{PCA}}, \ldots, u_r^{\text{PCA}}]$ where u_i^{PCA} are the leading eigenvectors of the covariance matrix $\text{Cov}(\mathbf{X})$. That is,

$$\text{Cov}(\mathbf{X}) u_i^{\text{PCA}} = \lambda_i(\text{Cov}(\mathbf{X})) u_i^{\text{PCA}}. \tag{7.38}$$

The same procedure can be applied for reducing the dimension of $\mathbf{Y}$, which yields $V = [v_1, \ldots, v_s]$ where v_i^{PCA} are the leading eigenvectors of the covariance matrix $\text{Cov}(\mathbf{Y})$. That is,

$$\text{Cov}(\mathbf{Y})v_i^{\text{PCA}} = \lambda_i(\text{Cov}(\mathbf{Y}))v_i^{\text{PCA}}. \tag{7.39}$$

There are two main drawbacks of using this dimension reduction method for Bayesian inference problems. The first is that PCA is an *unsupervised* method. That is, the directions identified by PCA are meant to reconstruct $\mathbf{X}$ and $\mathbf{Y}$ marginally, but they do not account for the dependence between $\mathbf{X}$ and $\mathbf{Y}$. Second, an accurate low-dimensional PCA approximation depends on the fast decay of the eigenvalues of the covariances $\text{Cov}(\mathbf{X})$ and $\text{Cov}(\mathbf{Y})$. In many inference problems, however, we can have low-dimensional structure without necessarily having sharp decay in the spectrums; cf. Example 6 where $\text{Cov}(\mathbf{X}) = \text{I}_d$ and $\text{Cov}(\mathbf{Y}) = \text{I}_m + GG^T$.

Alternatively, CCA seeks linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ that are maximally correlated [91, 85]. That is, CCA solves

$$(U_r^{\text{CCA}}, V_r^{\text{CCA}}) = \underset{\substack{U_r^T\text{Cov}(\mathbf{X})U_r=\text{I}_r \\ V_r^T\text{Cov}(\mathbf{Y})V_r=\text{I}_r}}{\arg\max} \text{Trace}(U_r^T\text{Cov}(\mathbf{X},\mathbf{Y})V_r), \tag{7.40}$$

where $\text{Cov}(\mathbf{X},\mathbf{Y})$ is the cross-covariance of $\mathbf{X}$ and $\mathbf{Y}$ for $r \leq \min\{d, m\}$. The vectors $(U_r^{\text{CCA}})^T\mathbf{X}$ and $(V_r^{\text{CCA}})^T\mathbf{Y}$ are called the *pairs of canonical variables*. It can be shown that U_r, V_r are found by solving the generalized eigenvalue problems

$$\text{Cov}(\mathbf{X},\mathbf{Y})\text{Cov}(\mathbf{Y})^{-1}\text{Cov}(\mathbf{Y},\mathbf{X})u_i^{\text{CCA}} = \rho_i\text{Cov}(\mathbf{X})u_i^{\text{CCA}} \tag{7.41}$$

$$\text{Cov}(\mathbf{Y},\mathbf{X})\text{Cov}(\mathbf{X})^{-1}\text{Cov}(\mathbf{X},\mathbf{Y})v_i^{\text{CCA}} = \rho_i\text{Cov}(\mathbf{Y})v_i^{\text{CCA}}, \tag{7.42}$$

where the eigenvectors are ordered based on a descending order for the eigenvalues $\rho_i \in [-1, 1]$.

The next proposition shows that, for linear-Gaussian likelihood models, our dimension reduction approach using whitening (see Section 7.4.1) is the same as CCA. The proof of this result is provided in Appendix E.

Proposition 18. *Let* $\mathbf{Y} = G\mathbf{X} + \boldsymbol{\mathcal{E}}$ *where* $G \in \mathbb{R}^{m\times d}$*, and* $\boldsymbol{\mathcal{E}} \perp\!\!\!\perp \mathbf{X}$ *with* $\mathbb{E}[\boldsymbol{\mathcal{E}}] = 0$ *and* $Cov(\boldsymbol{\mathcal{E}}) = \Gamma_{obs}$*. Then, the solution to* (7.40) *is given by*

$$U_r^{CCA} = \Gamma_{pr}^{-1/2}\overline{U}_r, \qquad \text{and} \qquad V_r^{CCA} = \Gamma_{obs}^{-1/2}\overline{V}_r,$$

where $\overline{U}_r$ *and* $\overline{V}_r$ *are the matrices containing the first* r *eigenvectors of the diagnostic matrices* $H_{\overline{\mathbf{X}}}, H_{\overline{\mathbf{Y}}}$ *defined in* (7.28) *and* (7.27)*, respectively.*

The method proposed in the present chapter can be seen as a generalization of CCA to nonlinear observation models. To find the relevant subspaces for the parameters and observations, we use gradient information of the forward model. We show in Section 7.7 that our approach produces rotations with more accurate posterior approximations than CCA for the same reduced dimensions. Lastly, we note that CCA is constrained to identifying subspaces for $\mathbf{X}$ and $\mathbf{Y}$ with the same dimension, i.e., $r = s$. In contrast, we can trade-off the two dimensions r, s while meeting a desired error tolerance; see Section 7.2.3.

7.6 Algorithms

In this section we present numerical algorithms to identify and exploit the low-dimensional subspaces for the informed parameters and informative observations when solving a Bayesian inference problem. To do so, we assume that we can evaluate mixed partial derivatives of the likelihood function and we can sample from the joint density of the parameters and observations. A sample $(\boldsymbol{X}^i, \boldsymbol{Y}^i)$ from the joint density $\pi_{\mathbf{X},\mathbf{Y}}$ can be easily obtained by sampling $\boldsymbol{X}^i \sim \pi_{\mathbf{X}}$ and sampling $\boldsymbol{Y}^i \sim \pi_{\mathbf{Y}|\mathbf{X}}(\cdot|\boldsymbol{X}^i)$.

Algorithm 8 presents the complete approach for identifying the subspaces given the reduced dimensions r, s. In the following subsections we show how to use the decomposition of the parameters and observations for (approximate) posterior sampling using two classes of inference algorithms. In 7.6.1 we assume we have access to evaluations of the likelihood function and the prior density, and in 7.6.2 we consider inference methods that only require samples from the joint density $\pi_{\mathbf{X},\mathbf{Y}}$.

Algorithm 8: Identify decomposition of parameter and observation spaces

Input : Prior density $\pi_{\mathbf{X}}$, Likelihood $\pi_{\mathbf{Y}|\mathbf{X}}$, Sample size n, Reduced dimensions r, s

Output: Orthonormal vectors U_r, V_s

1 Draw n i.i.d. samples $\{(\boldsymbol{X}^i, \boldsymbol{Y}^i)\}_{i=1}^n \sim \pi_{\mathbf{X},\mathbf{Y}} = \pi_{\mathbf{Y}|\mathbf{X}}\pi_{\mathbf{X}}$

2 Compute $\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{Y}^i|\boldsymbol{X}^i)$ for $i = 1, \ldots, n$

3 Assemble matrices $H_{\mathbf{X}}, H_{\mathbf{Y}}$ using Monte Carlo estimates

4 **Optimal rotations**: Solve eigenvalue problems $H_{\mathbf{X}}u_i = \lambda_i(H_{\mathbf{X}})u_i$, $H_{\mathbf{Y}}v_i = \lambda_i(H_{\mathbf{Y}})v_i$ for the eigenvectors corresponding to the r, s leading eigenvalues, **or**

5 **Optimal permutations**: Identify indices of largest diagonal entries in $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ and set $u_1, \ldots, u_r$ and $v_1, \ldots, v_s$ to those canonical unit vectors

6 Assemble $U_r = [u_1, \ldots, u_r]$, $V_s = [v_1, \ldots, v_s]$

7.6.1 Inference methods requiring likelihood

Markov chain Monte Carlo (MCMC) algorithms repeatedly evaluate the posterior density (up to a normalizing constant) to accept or reject candidate posterior samples. After identifying the matrices U, V using Algorithm 8, we define a Monte Carlo estimator for the reduced likelihood function in (7.7) in order to sample from the posterior density $\pi_{\mathbf{X}_r|\mathbf{Y}_s}$. Recall that the reduced likelihood is given by

$$\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r) = \int \pi_{\mathbf{Y}_s|\mathbf{X}}(\boldsymbol{y}_s|U_r\boldsymbol{x}_r + U_\perp\boldsymbol{x}_\perp)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\boldsymbol{x}_\perp|\boldsymbol{x}_r)\mathrm{d}\boldsymbol{x}_\perp, \tag{7.43}$$

where the marginalized likelihood $\pi_{\mathbf{Y}_s|\mathbf{X}}$ is given by

$$\pi_{\mathbf{Y}_s|\mathbf{X}}(\boldsymbol{y}_s|\boldsymbol{x}) = \int \pi_{\mathbf{Y}|\mathbf{X}}(V_s\boldsymbol{y}_s + V_\perp\boldsymbol{y}_\perp|\boldsymbol{x})\mathrm{d}\boldsymbol{y}_\perp. \tag{7.44}$$

The next example shows that when the likelihood is Gaussian, one can analytically compute the above integral in (7.44) so that the marginalized likelihood $\pi_{\mathbf{Y}_s|\mathbf{X}}$ is accessible in closed form.

Example 8. *For the (whitened) Gaussian-likelihood model in Section 7.4.1, we have*

the rotated observation model

$$\mathbf{Y}_s = V_s^T \Gamma_{obs}^{-1/2} \mathbf{Y} = V_s^T \Gamma_{obs}^{-1/2} G(\boldsymbol{x}) + V_s^T \Gamma_{obs}^{-1/2} \boldsymbol{\mathcal{E}}$$
$$\mathbf{Y}_\perp = V_\perp^T \Gamma_{obs}^{-1/2} \mathbf{Y} = V_\perp^T \Gamma_{obs}^{-1/2} G(\boldsymbol{x}) + V_\perp^T \Gamma_{obs}^{-1/2} \boldsymbol{\mathcal{E}}.$$

Given that the observational noise components $V_s^T \Gamma_{obs}^{-1/2} \boldsymbol{\mathcal{E}}$ *and* $V_\perp^T \Gamma_{obs}^{-1/2} \boldsymbol{\mathcal{E}}$ *are independent and have identity covariance, the marginalized likelihood is Gaussian with the form*

$$\pi_{\mathbf{Y}_s|\mathbf{X}}(\boldsymbol{y}_s|\boldsymbol{x}) = (2\pi s)^{-1/2} \exp\left(-\frac{1}{2} \|\boldsymbol{y}_s - V_s^T \Gamma_{obs}^{-1/2} G(\boldsymbol{x})\|_2^2 \right).$$

While the integral in (7.44) can be computed analytically, in general there is no closed form expression for the integral in (7.43). Thus, the reduced likelihood $\pi_{\mathbf{Y}_s|\mathbf{X}_r}$ needs to be estimated numerically. We consider here the Monte-Carlo estimator

$$\widehat{\pi}_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r) = \frac{1}{m} \sum_{i=1}^{m} \pi_{\mathbf{Y}_s|\mathbf{X}}(\boldsymbol{y}_s|U_r \boldsymbol{x}_r + U_\perp \boldsymbol{X}_\perp^i), \quad \boldsymbol{X}_\perp^i \sim \pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\cdot|\boldsymbol{x}_r). \tag{7.45}$$

We refer to [52] for an intensive discussion on different sampling strategies and on the impact of the sample size m versus the truncated dimension r. As shown in [239, 51], the variance of the above estimator is low when the error bound $\sum_{i=r+1}^{d} \lambda_i(H_{\mathbf{X}})$ is small. In practice, it is sufficient to use few samples (e.g., $m = 1$) or even deterministic approximations (e.g., by setting $\boldsymbol{X}_\perp^i$ to the conditional prior mean). More interestingly, using pseudo-marginal arguments [8], it is shown in [52] that redrawing fresh samples $\boldsymbol{X}_\perp^i$ at each MCMC iterations permits sampling from the *exact* reduced posterior.

7.6.2 Inference methods requiring joint samples

We now show how to sample from the posterior distribution given a collection of samples from the joint distribution of parameters and observations. We assume we have an algorithm that generates (approximate) conditional samples $\boldsymbol{X}_c^i \sim \pi_{\mathbf{X}|\mathbf{Y}}$ given

joint samples $(\boldsymbol{X}^i, \boldsymbol{Y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$. For example, this algorithm may be the measure transport approach presented in Chapter 5. This approach estimates an invertible map $S^{\mathcal{X}}\colon \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ that pushes forward all conditional densities $\pi_{\mathbf{X}|\mathbf{Y}}$ to a Gaussian reference for all realizations of $\mathbf{Y}$. Instead of using this algorithm directly with the high-dimensional variables $\mathbf{X}$ and $\mathbf{Y}$, we apply the following steps:

1. Identify the decompositions of $\mathbf{X}$ and $\mathbf{Y}$ using Algorithm 8
2. Project the parameters and observation samples: $\boldsymbol{X}_r^i = U_r^T \boldsymbol{X}^i$ and $\mathrm{Y}_s^i = V_s^T \boldsymbol{Y}^i$
3. Generate samples $\boldsymbol{X}_{c,r}^i \sim \pi_{\mathbf{X}_r|\boldsymbol{y}_s}$ from joint samples $(\boldsymbol{X}_r^i, \boldsymbol{Y}_r^i) \sim \pi_{\mathbf{X}_r,\mathbf{Y}_s}$
4. Reconstruct parameter samples in the original high-dimensional space: $\boldsymbol{X}_c^i = U_r \boldsymbol{X}_{c,r}^i + U_\perp \boldsymbol{X}_\perp^i$ where $\boldsymbol{X}_\perp^i \sim \pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\cdot|\boldsymbol{X}_{c,r}^i)$.

Let us remark that we can also sample $\boldsymbol{X}_\perp^i$ from the prior distribution for $\mathbf{X}_\perp$ conditioned on $\mathbf{X}_r = \boldsymbol{X}_r^i$ in step 4 above. While $\boldsymbol{X}_r^i$ is not conditioned on the observations, this procedure remains consistent for posterior sampling if the parameter satisfies the conditional independence property $\mathbf{X}_\perp \perp\!\!\!\perp \mathbf{Y}|\mathbf{X}_r$.

7.7 Numerical experiments

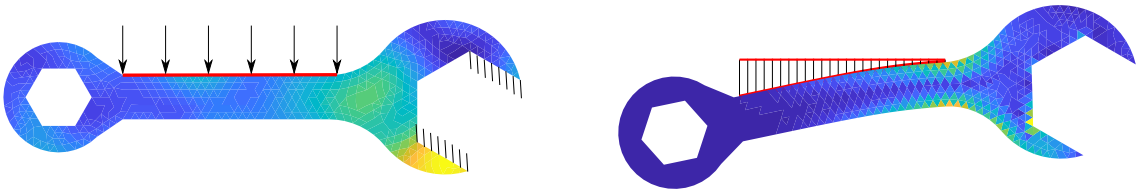
7.7.1 Linear elasticity

The first problem we consider is to infer the Young modulus field of a wrench body with domain $\mathcal{D} \subset \mathbb{R}^2$ given measurements of the displacement along its frontier [116, 199]. Let $u\colon \mathcal{D} \to \mathbb{R}^2$ represent the displacement field subject to an external force f applied on a subset of $\partial\mathcal{D}$. The displacement u satisfies the coupled elliptic PDE $\mathrm{div}(K : \epsilon(u)) = 0$ where $\epsilon(u) = \frac{1}{2}(\nabla u + \nabla u^T)$ is the strain field, and K is the Hooke tensor such that

$$K : \epsilon(u) \coloneqq \frac{E}{1+\nu}\epsilon(u) + \frac{\nu E}{1-\nu^2}\,\mathrm{Trace}(\epsilon(u))\mathrm{I}_2, \tag{7.46}$$

where $\nu = 0.3$ is the Poisson's ratio and $E\colon \mathcal{D} \to \mathbb{R}_{>0}$ is the Young's modulus. The displacement field is also subject to a Dirichlet boundary condition on the right-hand

side of the wrench. We model the Young's modulus field with a log-normal prior, i.e., $\mathbf{X} = \log E \sim \mathcal{N}(0, C)$ where $C(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2)$ is the Gaussian covariance function on $\mathcal{D} \times \mathcal{D}$ with length-scale one. For each realization of $\mathbf{X} = \log E$, we discretize the domain using a finite element mesh and solve the PDE for the solution u on each element on the domain. Figure 7-2 displays the input and a realization of the solution.



(a) Realization of the parameter (Young modulus field)

(b) Corresponding observation (vertical displacement along the red line)

Figure 7-2: Setting of the linear elasticity problem. (a) The log of the Young Modulus parameter field $\log(E)$. The black arrow represents the force applied to the body and the dashed lines denote the imposed boundary condition. (b) The von-Mises stress of the displacement field u. The observations used for inference are the displacements along the red line.

To determine the subspaces for the informed parameters and informative observations, we compute the gradients of the forward model at 500 samples from the joint density $\pi_{\mathbf{X},\mathbf{Y}}$. We use these gradients to define the matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ in (7.27) and (7.28), using the recommended whitening transformation. Figure 7-3 plots the sum of the trailing eigenvalues of $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ for the parameters and observations, respectively, which are labeled as CMI in the plot. These two sums correspond to the two terms in the upper bound for the expected KL divergence in (7.10). The fast decay in the eigenvalue sums indicates that linear-based dimension reduction can be used to accurately approximate the posterior distribution. We also evaluate the upper bound (up to the unknown conditional log-Sobolev constant) at the parameter and observation modes computed using either CCA or PCA. The approximation errors for both of these these alternative techniques decay at slower rates than the proposed dimension reduction approach in this work. In this example, the number of computable CCA or PCA modes is limited by the numerical rank of the covariance

matrices for $\mathbf{X}$ and $\mathbf{Y}$.

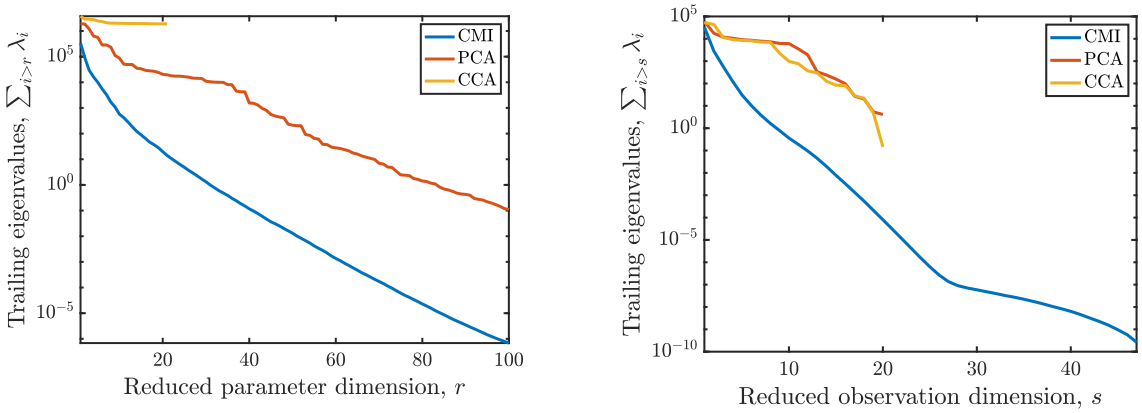
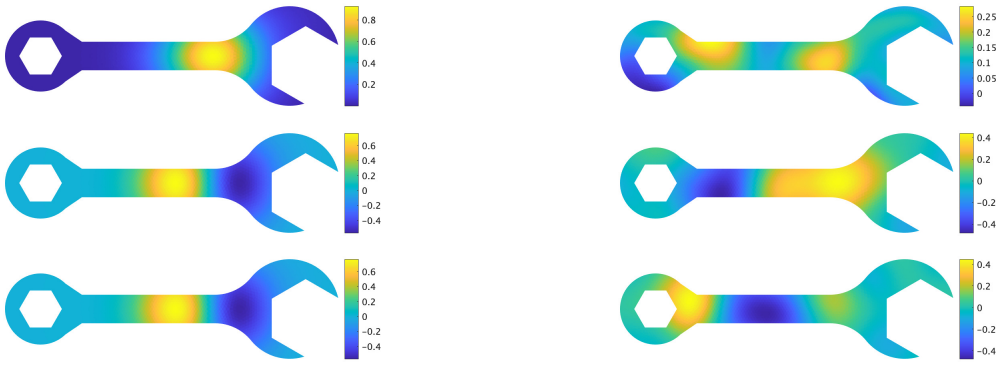


Figure 7-3: Upper bound for the expected KL divergence using three dimension reduction strategies with increasing reduced dimensions for (a) the parameters and (b) the observations in the linear elasticity inverse problem.

Figure 7-4a plots the first three modes $\Gamma_{\text{pr}}^{1/2}\overline{u}_i$ for the parameter space, where $\overline{u}_i$ is an eigenvector of the diagnostic matrix $H_{\overline{\mathbf{X}}}$. We observe that the informed part of the parameters is isolated near the wrench's axis of rotation, where there is typically higher stress. In comparison, Figure 7-4b plots the first three modes obtained using CCA, which display more global dependence on the displacement. Analogously, Figure 7-5 plots the first five modes $\Gamma_{\text{obs}}^{1/2}\overline{v}_i$ for the observation space, where $\overline{v}_i$ is an eigenvector of $H_{\overline{\mathbf{Y}}}$. The first mode has a stronger dependence on the displacement at the left-most part of the wrench's frontier, which is also the point of highest vertical displacement. In comparison, we observe that the first five modes obtained using CCA are more oscillatory, and hence depend on higher-frequency components of the displacement field.

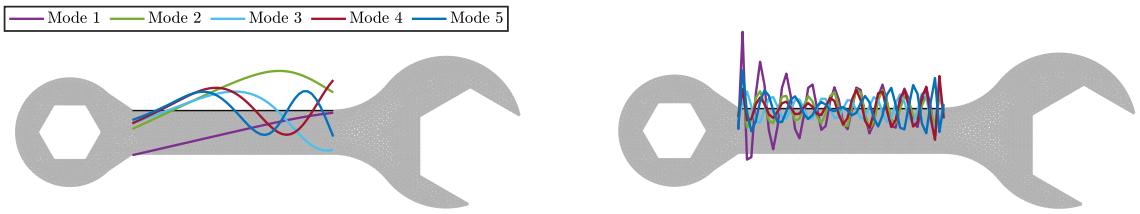
7.7.2 High-dimensional image observations

Next, we consider an inference problem with a non-Gaussian likelihood model that describes the location of a feature in a high-dimensional image. The feature is represented by two parameters $(x_1^*, x_2^*) \in [-16, 16]^2$ that define its horizontal and vertical position in the image. The forward model $G\colon \mathbb{R}^3 \to \mathbb{R}^{32\times 32}$ maps these location pa-



(a) (b)

Figure 7-4: (a) The first three modes (i.e., eigenvectors) from $H_{\overline{\mathbf{X}}}$ for the informed parameters in the linear elasticity inverse problem, and (b) the first three CCA modes



(a) (b)

Figure 7-5: (a) The first fives modes (i.e., eigenvectors) from $H_{\overline{\mathbf{Y}}}$ for the informative observations in the linear elasticity inverse problem, and (b) the first five CCA modes.

rameters and a contrast parameter $\gamma \in \mathbb{R}_+$ to a 32×32 pixel image. The forward model generates the image by computing the following probability $p_{\boldsymbol{x}} \in [0, 1]$ for each coordinate (x_1, x_2) in the image:

$$r_{\boldsymbol{x}} = (x_1 - x_1^*)^2 + (x_2 - x_2^*)^2$$
$$p_{\boldsymbol{x}} = 0.9 - 0.8 \exp(-0.5(r_{\boldsymbol{x}}/\sigma^2)^\gamma),$$

where $\sigma^2 \in \mathbb{R}_+$ determines the width of the feature. The intensity for each pixel $I_{\boldsymbol{x}} \in [0, 1]$ is then sampled independently from the continuous Bernoulli distribution with shape parameter $p_{\boldsymbol{x}}$, which has the density

$$\pi_{I|P}(I_{\boldsymbol{x}}|p_{\boldsymbol{x}}) \propto p_{\boldsymbol{x}}^{I_{\boldsymbol{x}}}(1 - p_{\boldsymbol{x}})^{1-I_{\boldsymbol{x}}}. \tag{7.47}$$

This problem was considered in [135] in the context of likelihood-free inference (LFI) using a discrete Bernoulli distribution for each component of the intensities $I_{\boldsymbol{x}}$. We use the relaxation in (7.47) to have continuous values for $I_{\boldsymbol{x}}$.

In our experiment we fix $\sigma = 3$, sample the contrast $\gamma \sim U[0.25, 5]$, and sample (x_1^*, x_2^*) from a uniform prior density $\pi_{\mathbf{X}}$ supported over the domain $[-16, 16]^2$. Figure 7-6 displays three candidate realizations of the grayscale images with different contrasts. Our goal here is to reduce the dimension of the observations without projecting the parameters $\mathbf{X} = (x_1^*, x_2^*, \gamma) \in \mathbb{R}^3$, given that they are already low-dimensional. This can be interpreted as defining summary statistics for the observations that are linear projections of $\mathbf{Y} = \text{vec}(I_{\boldsymbol{x}}) \in \mathbb{R}^{1024}$, such that $\pi_{\mathbf{X}|\mathbf{Y}} \approx \pi_{\mathbf{X}|\mathbf{Y}_s}$. Automatic methods for defining summary statistics are commonly used in many LFI procedures such as approximate Bayesian computation [66].

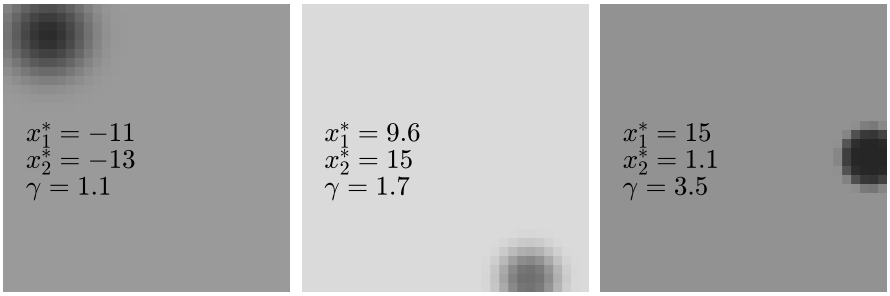


Figure 7-6: Three sample images from the model for image intensities.

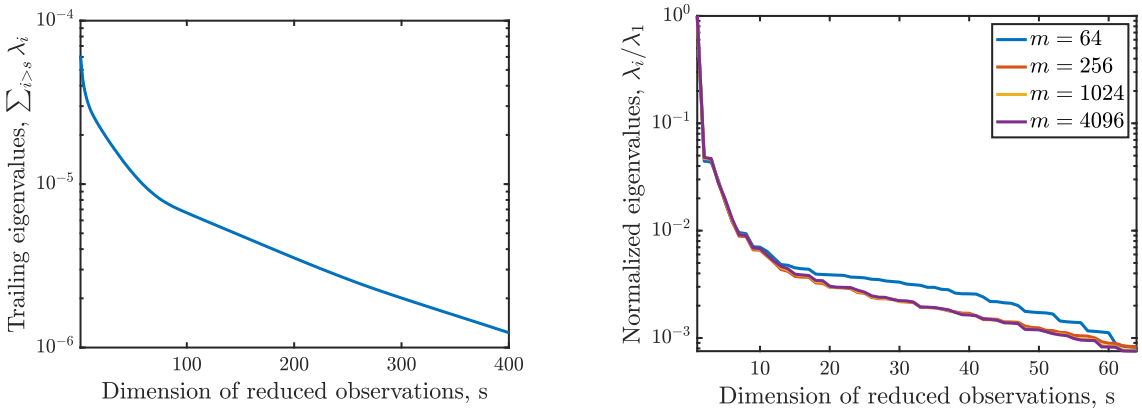


Figure 7-7: (a) The sum of the trailing eigenvalues of $H_{\mathbf{Y}}$ for projecting the observations. (b) The first 64 eigenvalues with increasing image resolution.

To reduce the dimension of the observations, we evaluate the mixed partial derivatives of the log-likelihood $\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} \log \pi_{\mathbf{Y}|\mathbf{X}} \in \mathbb{R}^{3\times 1024}$ at $n = 10^5$ samples $(\mathbf{X}, \mathbf{Y}) \sim \pi_{\mathbf{X},\mathbf{Y}}$ and assemble a Monte Carlo estimate for the matrix $H_{\mathbf{Y}}$ in (7.12). Figure 7-7a displays the sum of the trailing eigenvalues for $H_{\mathbf{Y}}$, which controls the posterior approximation error in expected KL divergence, up to the log-Sobolev constant for the joint density $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$. We observe fast decay in this sum for the initial eigenvalues, indicating that low-dimensional projections of the image are sufficient to approximately infer the parameters. We note that this decay is also unaffected by the grid resolution; see the right plot in Figure 7-7.

Figure 7-8 displays the ten eigenvectors of the diagnostic matrix corresponding to its ten leading eigenvalues. We observe low-frequency oscillations in the first eigenvectors. These modes provide sufficient information from the image to approximately determine the location of the circular feature, while higher-order eigenvectors distinguish finer features of the image in order to more accurately infer the feature's position. Let us recall that CCA is limited to computing at most $\min\{d, p\} = 3$ modes for this example.

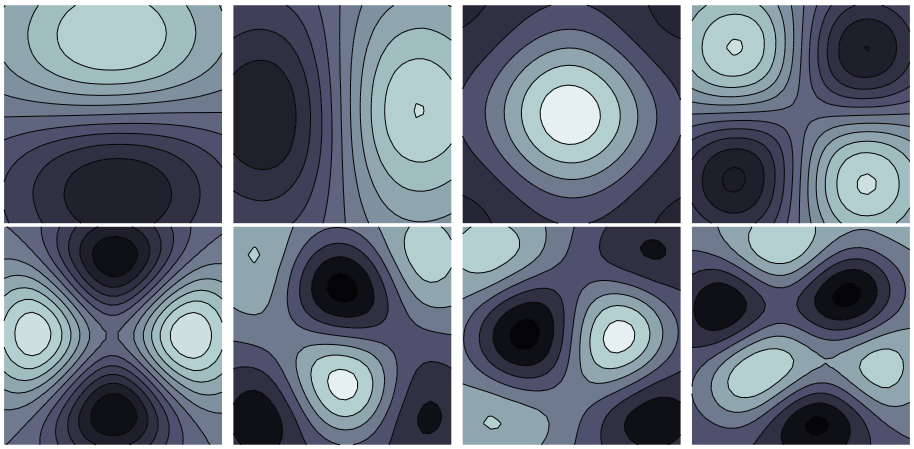
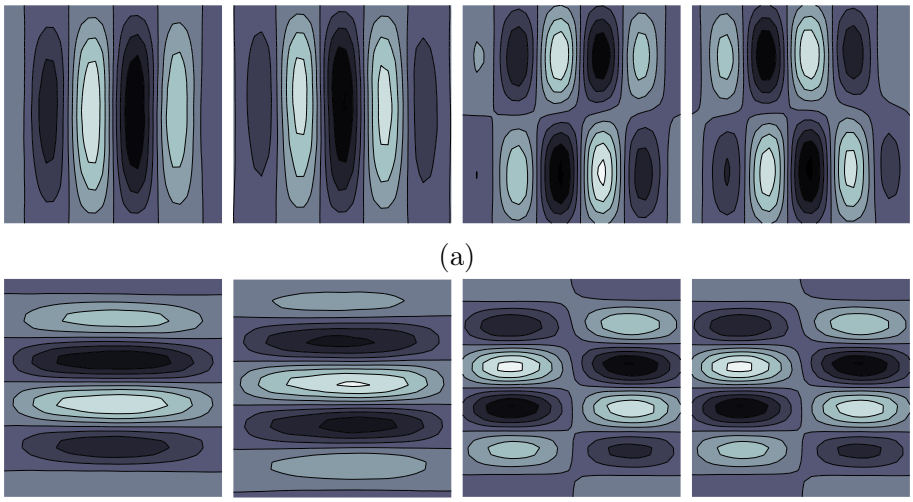


Figure 7-8: The first eight leading non-constant eigenvectors of the data informed matrix $H_{\mathbf{Y}}$

The gradients of the log-likelihood are also useful to determine *goal-oriented* subspaces for the observations for inferring a subset of the parameters. For the two location parameters x_1^*, and x_2^*, we compute the diagnostic matrix $H_{\mathbf{Y}}$ using the mixed partial derivatives of the log-likelihood function with respect to $\mathbf{Y}$ and one of

the location parameters. Figure 7-9 plots the first four eigenvectors of each matrix. As intuitively expected, the eigenvectors extract horizontal and vertical variations in the image to infer the x_1^* and x_2^* parameters, respectively.



(a)

(b)

Figure 7-9: The first four leading eigenvectors of the data informed matrix $H_{\mathbf{Y}}$ for informing the parameter (a) x_1^*, and (b) x_2^*.

7.7.3 Conditioned diffusion

In this section we infer the driving force applied to a particle given observations of its path. The particle's path is described by a function $u\colon [0,1] \to \mathbb{R}$. The path undergoes dynamics given by the stochastic differential equation

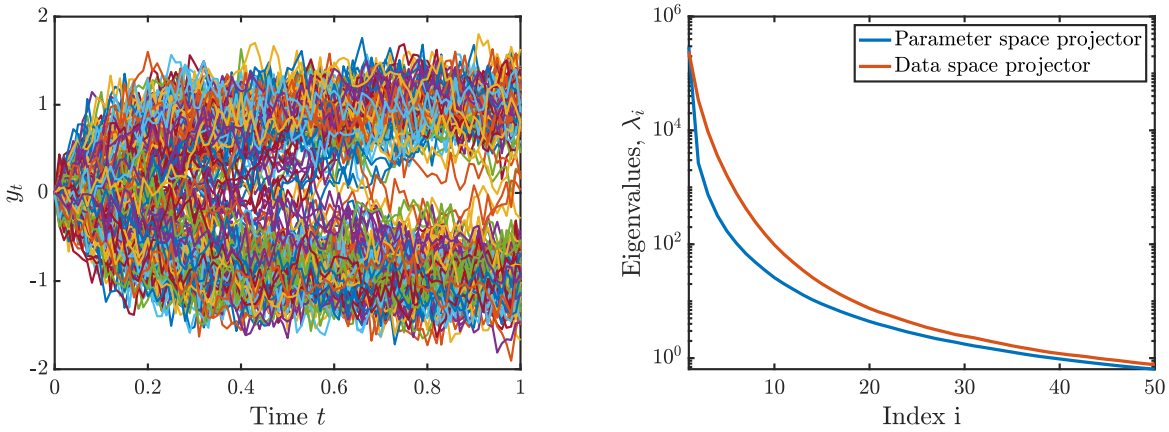
$$\mathrm{d}u_t = f(u_t)\mathrm{d}t + \mathrm{d}X_t, \quad u_0 = 0 \tag{7.48}$$

where $f\colon \mathbb{R} \to \mathbb{R}$ is the nonlinear drift function $f(u) = \beta u(1-u^2)/(1+u^2)$ for $\beta > 0$ and $\mathrm{d}X_t$ is an increment of the Brownian motion $X \sim \mathcal{N}(0, C)$ with covariance function $C(t,t') = \min(t,t')$. In our setup, we set $\beta = 1$ and discretize the ODE using a Euler-Maruyama scheme with time-step $\Delta t = 10^{-2}$ so that $d = 100$. At M equispaced times $t_1, \ldots, t_m$ in the interval $[0,1]$, we collect noisy observations of the state

$$Y_{t_i} = u_{t_i} + \mathcal{E}_i, \tag{7.49}$$

where $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_m)$. Joining the forward and observation models, the map from parameters to observations can be written in vector form as $\mathbf{Y} = MG(\mathbf{X}) + \boldsymbol{\mathcal{E}}$ where M is a selection operator that extracts the state at the observed times and G is the nonlinear operator that maps the noise $\mathbf{X} \in \mathbb{R}^d$ to the path $u \in \mathbb{R}^d$. In our study we set $\sigma = 0.1$, and $M = \mathrm{I}_d$, i.e., we observe all components of the discretized path vector.

To identify the subspaces for the parameters and observations we compute Monte Carlo estimators of the diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$ in (7.27) and (7.28) using $n = 10^6$ prior samples $\boldsymbol{X}^i \sim \pi_{\mathbf{X}}$ and evaluations of the forward model $u^i = G(\boldsymbol{X}^i)$ for $i = 1, \ldots, n$. Figure 7-10a plots 200 samples of the observations $\boldsymbol{Y}^i$ and Figure 7-10b plots the leading eigenvalues of the two diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$. Let us remark that unlike the linear-Gaussian setting in Section 7.4.2, the eigenvalues of the two diagnostic matrices with a nonlinear forward model are different for $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$.

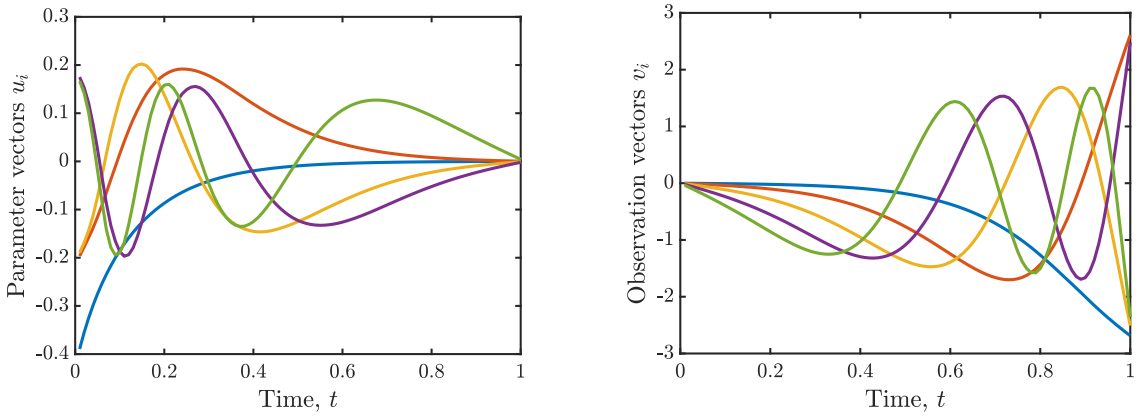


(a) (b)

Figure 7-10: (a) Samples of the observations $\boldsymbol{Y}$ from the conditioned diffusion model. (b) The leading 50 eigenvalues of the diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$.

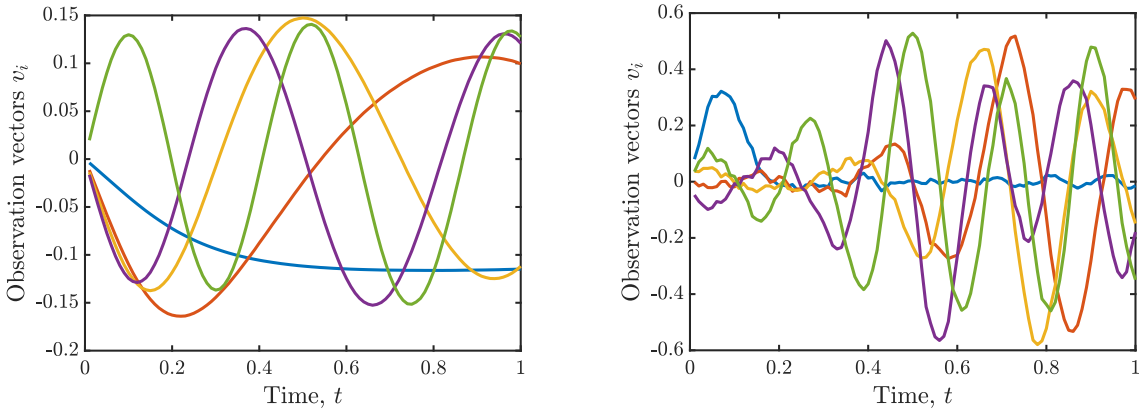
Figure 7-11 plots the eigenvectors for the parameters and observations corresponding to the leading 5 eigenvalues of $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$, respectively. The parameter-space eigenvectors capture the bulk behavior of sample path realizations, while the observation-space eigenvectors show more variation near $T = 1$. An intuitive explanation for this behavior is that conditioning on the observations at $T = 1$ will provide more information about the particle's final location and the driving force needed to

get there. For contrast, we plot the eigenvectors arising from PCA and CCA with 10^6 samples from the joint distribution of parameters and observations in Figures 7-12a and 7-12b, respectively. We observe that PCA modes are more "global" than the ones obtained from the CMI bound proposed in this chapter, and the CCA modes are more irregular and oscillatory.



(a) (b)

Figure 7-11: (a) Parameter space and (b) observation space eigenvectors of the diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$.



(a) (b)

Figure 7-12: Observation space eigenvectors from (a) PCA and (b) CCA.

To evaluate the true approximation errors resulting from the parameter and observation-space projectors, we compute the conditional mutual information (CMI) for each projection. To estimate the CMI $I(\mathbf{X}_\perp; \mathbf{Y}|\mathbf{X}_r)$, we generate n samples

$(\boldsymbol{X}^i, \boldsymbol{Y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$ and construct the Monte Carlo estimator

$$\widehat{I}(\mathbf{X}_\perp; \mathbf{Y}|\mathbf{X}_r) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{Y}^i|\boldsymbol{X}^i)}{\pi_{\mathbf{Y}|\mathbf{X}_r}(\boldsymbol{Y}^i|\boldsymbol{X}_r^i)}, \tag{7.50}$$

where $\pi_{\mathbf{Y}|\mathbf{X}_r}(\boldsymbol{y}|\boldsymbol{x}_r) = \int \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|U_r\boldsymbol{x}_r + U_\perp\boldsymbol{x}_\perp)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\boldsymbol{x}_\perp|\boldsymbol{x}_r)\mathrm{d}\boldsymbol{x}_\perp$. Analogously to the estimator in (7.45) for the reduced likelihood $\pi_{\mathbf{Y}_s|\mathbf{X}_r}$, we estimate $\pi_{\mathbf{Y}|\mathbf{X}_r}$ using the following Monte Carlo estimator given m samples from the conditional prior

$$\widehat{\pi}_{\mathbf{Y}|\mathbf{X}_r}(\boldsymbol{y}|\boldsymbol{x}_r) = \frac{1}{m}\sum_{j=1}^{m} \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|U_r\boldsymbol{x}_r + U_\perp\boldsymbol{X}_\perp^j), \quad \boldsymbol{X}_\perp^j \sim \pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\cdot|\boldsymbol{x}_r). \tag{7.51}$$

To see the impact of m on estimating the conditional mutual information, Figure 7-13a plots the estimator with $m \in \{10, 100, 1000\}$ as well as a single sample, i.e., $m = 1$, at the prior mean $\int \boldsymbol{x}_\perp \mathrm{d}\pi_{\mathbf{X}_\perp|\mathbf{X}_r}$ with $n = 10^4$ samples in (7.50). We observe a convergence of the CMI estimators with increasing m. Furthermore, the CMI closely matches the trend for the upper bound (up to the conditional log-Sobolev constant), which indicates that for this example the bound can be used as a good error indicator even without knowing $\overline{C}(\pi_{\mathbf{X},\mathbf{Y}})$.

To estimate the conditional mutual information $I(\mathbf{Y}_\perp; \mathbf{X}|\mathbf{Y}_s) = I(\mathbf{X}; \mathbf{Y}) - I(\mathbf{X}; \mathbf{Y}_s)$, we generate n samples $(\boldsymbol{X}^i, \boldsymbol{Y}^i) \sim \pi_{\mathbf{X},\mathbf{Y}}$ and construct the Monte Carlo estimator

$$\widehat{I}(\mathbf{X}_\perp; \mathbf{Y}|\mathbf{X}_r) := \frac{1}{n}\sum_{i=1}^{n} \log \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{Y}^i|\boldsymbol{X}^i)}{\pi_{\mathbf{Y}}(\boldsymbol{Y}^i)} - \log \frac{\pi_{\mathbf{Y}_s|\mathbf{X}}(\boldsymbol{Y}_s^i|\boldsymbol{X}^i)}{\pi_{\mathbf{Y}_s}(\boldsymbol{Y}_s^i)}, \tag{7.52}$$

where the marginal likelihoods $\pi_{\mathbf{Y}}$ and $\pi_{\mathbf{Y}_s}$ are estimated using m prior samples for each sample $\boldsymbol{Y}^i$, e.g., $\pi_{\mathbf{Y}}(\boldsymbol{Y}^i) = \frac{1}{m}\sum_{j=1}^{m} \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{Y}^i|\boldsymbol{X}^j)$ for $\boldsymbol{X}^j \sim \pi_{\mathbf{X}}$. To compute the likelihood for the reduced observation $\mathbf{Y}_s$, we analytically marginalize the Gaussian likelihood by projecting the mean and covariance of the observational noise using the formula in Example 8. This avoids an additional numerical integration. Figure 7-13b plots the estimated conditional mutual information for the projected observations with increasing reduced dimension along with the upper bound in (7.17) up to the conditional log-Sobolev constant. We observe that the estimators converge with in-

creasing sample size m. Furthermore, the upper bound closely matches the trend for the true approximation error, especially for larger s.

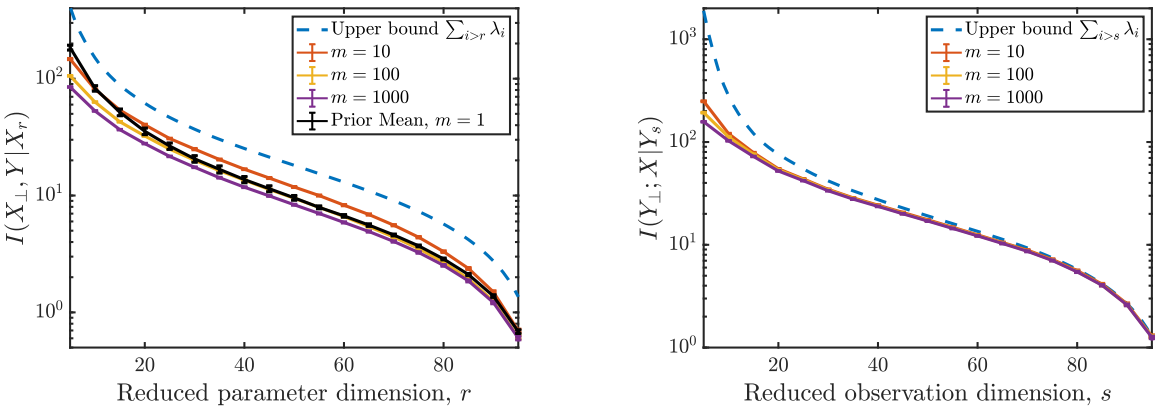
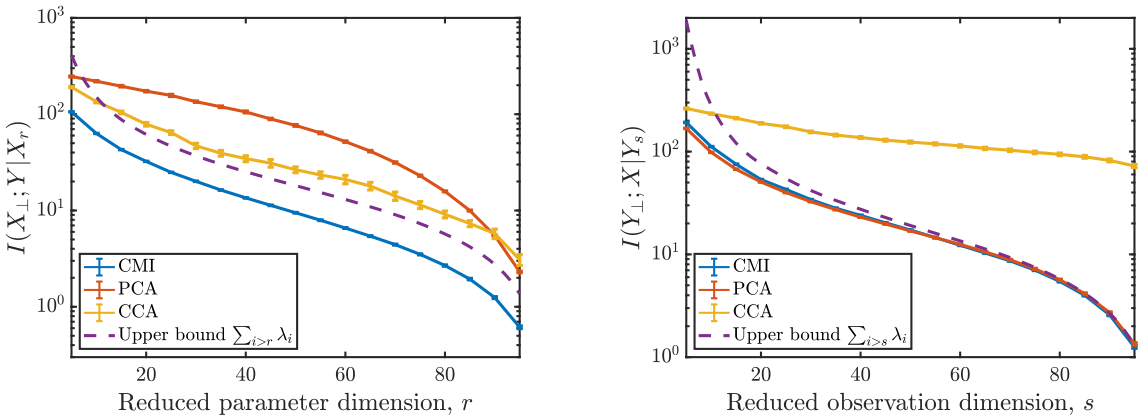


Figure 7-13: Convergence of the estimators for the conditional mutual information with projected (a) parameters and (b) observations in the conditioned diffusion model.

We also compare the optimal rotations identified from the upper bound in (7.10) to the subspaces resulting using PCA and CCA. Figures 7-14a and 7-14b plot the conditional mutual information for the posterior approximation error of each dimension reduction technique with increasing dimensions for the reduced parameters and observations, respectively. The conditional mutual information is computed using the Monte Carlo estimators in (7.50) and (7.52) with $n = 10^4$ and $m = 100$. In this non-Gaussian example, the CMI bounds present the lowest error for the projection of the parameters. We note that PCA performs very similarly to the CMI bound for reducing the observations in this example, despite having very different modes; see Figures 7-11b and 7-12. Together with the parameters, the rotations identified from the CMI bound provide the lowest overall posterior approximation error.

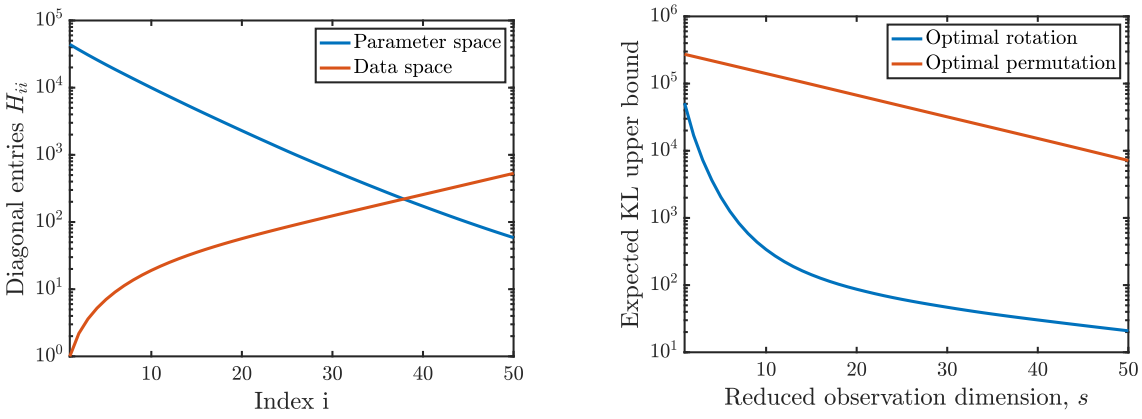
Lastly, we consider the problem of selecting coordinates of the parameters and observations that minimize the upper bound in (7.19). We recall that this corresponds to finding a decreasing ordering for the diagonal entries of the diagnostic matrices $H_{\overline{\mathbf{X}}}$ and $H_{\overline{\mathbf{Y}}}$. Figure 7-15a plots the values for the diagonal entries in their canonical ordering with increasing time t. We observe that the forcing at the initial time, i.e., $t = 0$, is the most informed parameter, while the most informative observation is the



(a) (b)

Figure 7-14: Comparison of three dimension reduction strategies for reducing the dimension of (a) the parameters and (b) the observations in the conditioned diffusion model

position at the final time, i.e., $t = 1$. To compare the effect of coordinate selection to the optimal rotations found above on the posterior approximation error, Figure 7-15b plots the upper bounds for the expected KL divergence with projected observations (up to the conditional log-Sobolev constant). For this example, we observe that the optimal rotations yield an improvement of at least two orders of magnitude and converges at a faster rate, particularly for lower-dimensional observations $\mathbf{Y}_s$.



(a) (b)

Figure 7-15: (a) Diagonal entries of the diagonal matrices $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$. (b) Comparison of the upper bound in (7.10) for optimal rotations and coordinate selection of the observations in the conditioned diffusion model

7.7.4 Sequential Bayesian inference

Decompositions of the parameters and observations are also relevant for the stochastic map filtering algorithm introduced in Chapter 6. Let $T_{\boldsymbol{y}^*} = S(\boldsymbol{y}^*, \cdot)^{-1} \circ S(\boldsymbol{y}, \boldsymbol{x})$ be a composed map that pushes forward joint samples $(\mathbf{X}, \mathbf{Y}) \sim \pi_{\mathbf{X},\mathbf{Y}}$ to samples from the conditional density $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y}^*)$. If the observations are only informative of the parameters $\mathbf{X}_\perp$, we can constrain the action of $T_{\boldsymbol{y}^*}$ to depart from the identity only along the subspace of the reduced parameters spanned by U_r, i.e., $T_{\boldsymbol{y}^*}$ does not update the uninformed parameters. In this case, the prior-to-posterior transformation is given by

$$T_{\boldsymbol{y}^*}(\boldsymbol{y}, \boldsymbol{x}) = \begin{bmatrix} U_r & U_\perp \end{bmatrix} \begin{bmatrix} T'_{\boldsymbol{y}_s^*}(V_s^T \boldsymbol{y}, U_r^T \boldsymbol{x}) \\ \boldsymbol{x}_\perp \end{bmatrix}, \tag{7.53}$$

where $T'_{\boldsymbol{y}_s^*} \colon \mathbb{R}^{s+r} \to \mathbb{R}^r$ is a low-dimensional map that depends on the reduced parameters and observations and $\boldsymbol{y}_s^* = V_s^T \boldsymbol{y}^*$.

For the linear-Gaussian likelihood model in Section 7.4.2, the optimal rotations for $\mathbf{X}$ and $\mathbf{Y}$ yield a forward model that is diagonal in the rotated coordinate system, i.e., $\mathbf{Y}_s = V_s^T \mathbf{Y} = \Sigma U_r^T \mathbf{X} + V^T \boldsymbol{\mathcal{E}}$ where $\Sigma \in \mathbb{R}^{r \times s}$ is a diagonal matrix with zeros in rows and columns greater than $\min\{r, s\}$, $U_r^T \mathbf{X} \sim \mathcal{N}(0, \mathrm{I}_r)$ and $V_s^T \boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, \mathrm{I}_s)$. From the Bayesian update for independent Gaussian random variables, the low-dimensional map in this case is given by $T'_{\boldsymbol{y}_s^*}(\boldsymbol{y}_s, \boldsymbol{x}_r) = \boldsymbol{x}_r - \Sigma^T(\Sigma\Sigma^T + \mathrm{I}_s)^{-1}(\boldsymbol{y}_s^* - \boldsymbol{y}_s)$. The map in (7.53) written with respect to the original variables is then given by $T_{\boldsymbol{y}^*}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} - K_{r,s}(\boldsymbol{y} - \boldsymbol{y}^*)$ where

$$K_{r,s} \coloneqq U_r \Sigma^T (\Sigma\Sigma^T + \mathrm{I})^{-1} V_s^T, \tag{7.54}$$

is a low-rank Kalman gain and has dimensions $d \times m$. Let us remark that the expression in (7.54) reveals the order of the inference procedure. From right to left, the observation $\boldsymbol{y}$ is rotated and projected using V_s^T, assimilated in a subspace of the rotated parameter, and then lifted to the original parameter space using U_r. We refer the reader to [123, 171] where the low-dimensional map in (7.53) based on the dimension reduction technique proposed in this chapter was used to improve the tracking

performance of ensemble filters for fluid dynamics applications.

7.8 Discussion and extensions

This chapter proposed a gradient-based dimension reduction method for the d parameters and m observations in non-Gaussian Bayesian inference problems. We reduced the dimensions by identifying linear subspaces for the informed part of the parameters $\mathbf{X}_r \in \mathbb{R}^r$ with $r \ll d$ and the informative part of the observations $\mathbf{Y}_s \in \mathbb{R}^s$ with $s \ll m$. These subspaces yield posterior approximations that depart from the prior distribution only along a low-dimensional subspace conditioned only on the informative observations. The subspaces are defined as the minimizers of a tractable quadratic upper bound for the expected KL divergence from the approximation to the true posterior distribution. For linear-Gaussian likelihood models with $r = s$, these reduce to subspaces identified by canonical correlation analysis (CCA). For several inference problems with non-linear forward models, we showed that this technique provides interpretable projections of the parameters and observations, and they yield more accurate posterior approximations than both CCA and principal component analysis. We outline some directions for future work below.

Gradient-free methods. The optimal rotations of the variables are defined from eigendirections of two diagnostic matrices $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ containing the mixed partial derivatives of the log-likelihood function. For applications with large-scale forward models, however, these derivatives may be unavailable or computationally expensive to evaluate. It will be interesting to develop estimators for $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ based only on differences of forward model evaluations. Furthermore, analyzing the sample complexity of these estimators will be useful to determine the number of samples required to achieve a desired approximation to the subspaces of interest.

Nonlinear dimension reduction. For nonlinear forward models, linear dimension reduction may require many modes to achieve small posterior approximation error. In these cases, one may instead seek few nonlinear functions of the parameters that are

informed by nonlinear features of the observations. See [22] for an approach to identify these nonlinear features in the context of surrogate modeling. It will be interesting to extend the guarantees on posterior approximation error to permit nonlinear functions of the parameters and observations, and to compare the resulting features to those identified by nonlinear supervised dimension reduction methods such as [7, 145].

Chapter 8

Conclusions

This thesis addresses key problems in probabilistic modeling and Bayesian inference by using transportation of measures to represent probability distributions. Measure transport seeks an invertible transformation, known as a transport map, that couples a random variable with a complex target distribution to one that is tractable (e.g., a standard Gaussian). Having access to this map enables sampling from or evaluating the density of the target distribution. A core challenge, however, is that transport maps are rarely known analytically. Instead, these maps must be estimated from a limited collection of samples. This thesis develops tailored transport map estimators for solving various modeling and inference problems.

One complex target distribution that appears throughout science and engineering is the posterior for the states or parameters of a statistical model. In addition to being high-dimensional in many applications, the unnormalized posterior density is often computationally prohibitive to evaluate, i.e., the likelihood function or the prior density are intractable. As an example, Chapter 5 considers a polymer science model where the likelihood function for the observations is defined by an integral over an infinite-dimensional latent variable. Similarly, Chapter 6 presents nonlinear filtering problems where the prior densities for the states of a dynamical system are analytically unavailable. In these cases, likelihood-free inference algorithms based on measure transport offer a principled approach to estimate or sample from posterior densities given only joint samples of parameters and observations. In this thesis, we build to-

wards solving high-dimensional and non-Gaussian Bayesian inference problems using transportation of measures. In the process, we present several contributions to topics in probabilistic modeling and Bayesian inference.

A core component of our transport-based algorithms is the approximation of monotone triangular transport maps. In Chapter 3 we propose a general representation for monotone functions using smooth bijective operators. This representation results in a smooth unconstrained optimization problem for the map that is proven to have no spurious local minima, a key contrast to other commonly used map parameterizations. To approximate the map, we introduce the Adaptive Transport Map (`ATM`) algorithm that automatically adapts the map complexity to a collection of samples from a target density. We show that this algorithm reduces the bias and variance of map estimators, and it discovers sparsity in the map's variable dependence that arises from conditional independencies in each marginal conditional. In practice, the `ATM` algorithm performs similarly to methods that have knowledge of the true map sparsity, which helps to scale the algorithm for high-dimensional distributions with this structure. In Chapter 4 we also use the sparse structure of triangular transports to consistently learn all pairwise conditional independence properties of a (possibly) non-Gaussian distribution. To do so, we introduce a score based on Hessian information of the joint log-density and propose a threshold estimator that consistently recovers the Markov structure with increasing sample sizes. We show how our algorithm, named Sparsity Identification in Non-Gaussian Distributions, iteratively leverages sparsity in an estimated transport map to learn the structure of undirected probabilistic graphical models.

In the context of Bayesian inference, we propose novel methodologies for likelihood-free and sequential inference problems with high-dimensional parameters and observations. First, in Chapter 5 we introduce a prior-to-posterior transformation for conditional sampling based on the composition of triangular maps. This map pushes forward joint samples of parameters and observations to posterior samples. Given any estimate for the triangular map, we show theoretically and numerically that posterior approximations produced by the composed map have lower bias and variance than us-

ing popular non-composed maps. Second, we generalize triangular to block-triangular maps that reduce the sensitivity of triangular maps to variable ordering. We show that block-triangularity reliably characterizes the posterior distributions for random fields in PDEs and for image in-painting problems. Furthermore, we show that monotone block-triangular maps minimize a particular transportation cost, thereby providing a new computational method for finding optimal transport maps. In Chapter 6 we use the composed prior-to-posterior maps to propose a nonlinear ensemble filtering algorithm called the Stochastic Map (SM) filter. We show that the SM filter reduces to the celebrated ensemble Kalman filter (EnKF) when the triangular maps are affine functions, while nonlinear maps reduce the bias of the EnKF for capturing non-Gaussian filtering distributions. Specifically, the SM filter with sparse transport maps achieves state-of-the-art performance for tracking the states of the chaotic Lorenz-96 system. Lastly, to scale these transport-based methods to high-dimensional inference problems, Chapter 7 proposes a technique to jointly reduce the dimensions of parameters and observations with error guarantees on the posterior approximation. In combination with the SM filter, this technique was shown in [171] to yield stable filtering with small ensemble sizes and reconstruct turbulent flow fields with higher fidelity than traditional filters, such as the EnKF.

In this thesis we show how structured nonlinear transport maps yield improved inference results in geophysical and aerodynamics problems. To broaden the impact of these modeling and inference algorithms to other scientific domains, it is important to investigate alternative sources of low-dimensional structure in these maps. For instance, not all posterior distributions will satisfy (approximate) conditional independencies and yield maps with sparse variable dependence. In other dynamical systems the map may only be sparse under a variable transformation, e.g., images and multi-scale random fields are often sparsely represented after a wavelet transformation. Another source of structure occurs when the target density departs from the (standard Gaussian) reference only along a low-dimensional subspace. In a variational inference context, [28] identified this subspace from gradients of the target density (as in Chapter 7) and used it to estimate triangular maps that depart from

the identity transformation only in few coordinates. While gradients of the target are unavailable in density estimation and likelihood-free settings, future work may use recent score estimation methods [202] to approximate these gradients from samples alone. Ultimately, it will be interesting to provide diagnostics for different types of low-dimensional structure and introduce transport map estimators that can automatically exploit this structure with low computational cost.

Another important direction is to improve transport-based methods to sample from constrained input spaces. For many physical problems, the states or parameters in modeling and inference problems are known to satisfy physical laws and constraints, e.g., the inference of flow-fields that are divergence-free or rotation matrices that live in a Lie group. While transport-based methods can be applied in generic settings without knowledge of these constraints, samples generated from estimated maps may violate these properties. To generate more physically-consistent samples, it will be interesting to develop methods that directly encode constraints into the map. This will improve the generalization performance of transport maps for posterior sampling at previously unseen observations, especially in settings with limited training data. Together, these directions will bring transport-based modeling and inference closer into practice.

Appendix A

Proofs for Chapter 3

A.1 Proof of Inequality (3.2)

Recall that the KR rearrangement S_{KR} is a transport map that satisfies $S_{\mathrm{KR}}^{\sharp}\eta = \pi$, where η is the PDF of the standard Gaussian measure on $\mathbb{R}^d$ and π is the target PDF. Corollary 3.10 in [27] states that for any PDF ν on $\mathbb{R}^d$ of the form $\nu(\boldsymbol{x}) := f(\boldsymbol{x})\eta(\boldsymbol{x})$ with $f \log f \in L^1_\eta$, the inequality

$$\int \|\boldsymbol{x} - T(\boldsymbol{x})\|^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq 2 \int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{A.1}$$

holds, where T is the KR rearrangement such that $T_\sharp \eta = \nu$. Let S be an increasing lower triangular map as in (2.3) and let $\nu = S_\sharp \pi$. Thus we have $T = S \circ S_{\mathrm{KR}}^{-1}$ and so the left-hand side of (A.1) becomes

$$\int \|\boldsymbol{x} - T(\boldsymbol{x})\|^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int \|\boldsymbol{x} - S \circ S_{\mathrm{KR}}^{-1}(\boldsymbol{x})\|^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int \|S_{\mathrm{KR}}(\boldsymbol{x}) - S(\boldsymbol{x})\|^2 \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

and the right-hand

$$2 \int f(\boldsymbol{x}) \log f(\boldsymbol{x}) \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 2\mathcal{D}_{\mathrm{KL}}(\nu || \eta) = 2\mathcal{D}_{\mathrm{KL}}(\pi || S^{\sharp}\eta),$$

which yields (3.2)

A.2 Convexity of map optimization problem

Lemma A.2.1. *The optimization problem* $\min_{\{s:\partial_k s>0\}} \mathcal{J}_k(s)$ *is strictly convex.*

Proof. Let $s_1, s_2 : \mathbb{R}^k \to \mathbb{R}$ be two strictly increasing functions with respect to x_k, i.e., $\partial_k s_1(\boldsymbol{x}_{\leq k}) > 0$ and $\partial_k s_2(\boldsymbol{x}_{\leq k}) > 0$. For any $0 < t < 1$, the function $s_t = ts_1 + (1-t)s_2$ is also strictly increasing functions with respect to x_k. Finally, because both $\xi \mapsto \frac{1}{2}\xi^2$ and $\xi \mapsto -\log(\xi)$ are strictly convex functions, we have

$$\begin{aligned}\mathcal{J}_k(s_t) &\overset{(3.4)}{=} \int \left(\frac{1}{2}s_t(\boldsymbol{x}_{\leq k})^2 - \log \partial_k s_t(\boldsymbol{x}_{\leq k})\right)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &< \int \left(t\frac{1}{2}s_1(\boldsymbol{x}_{\leq k})^2 + (1-t)\frac{1}{2}s_1(\boldsymbol{x}_{\leq k})^2\right) - \Big(t\log \partial_k s_1(\boldsymbol{x}_{\leq k}) + (1-t)\log \partial_k s_2(\boldsymbol{x}_{\leq k})\Big)\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= t\mathcal{J}_k(s_1) + (1-t)\mathcal{J}_k(s_2),\end{aligned}$$

which shows that $\mathcal{J}_k$ is strictly convex. □

A.3 Proof of Theorem 3.3.1

To prove Theorem 3.3.1 we need the following lemma.

Lemma A.3.1. *Let*

$$H^1([0,1]) = \left\{ f : [0,1] \to \mathbb{R} \text{ such that } \|f\|^2_{H^1([0,1])} := \int_0^1 f(t)^2 + f'(t)^2 \,\mathrm{d}t \right\}.$$

Then

$$|f(0)| \leq \sqrt{2}\|f\|_{H^1([0,1])}, \tag{A.2}$$

holds for any $f \in H^1([0,1])$.

Proof. Because $\mathcal{C}^\infty([0,1])$ is dense in $H^1([0,1])$, it suffices to show (A.2) for any $f \in \mathcal{C}^\infty([0,1])$. By the mean value theorem, there exists $0 \leq z \leq 1$ such that

$$f(z) = \frac{1}{1-0}\int_0^1 f(t)\mathrm{d}t.$$

Thus we can write

$$\begin{aligned}|f(0)|^2 &\leq 2|f(z)-f(0)|^2 + 2|f(z)|^2 \\ &= 2\left|\int_0^z f'(t)\mathrm{d}t\right|^2 + 2\left|\int_0^1 f(t)\mathrm{d}t\right|^2 \\ &\leq 2\int_0^1 |f'(t)|^2\,\mathrm{d}t + 2\int_0^1 |f(t)|^2\,\mathrm{d}t.\end{aligned}$$

This concludes the proof. □

We now prove Theorem 3.3.1.

Proof. For any $f \in V_k$, Lemma A.3.1 permits to write

$$\begin{aligned}\int |f(\boldsymbol{x}_{<k},0)|^2\eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x}_{<k} &\overset{(A.2)}{\leq} \int\left(2\int_0^1 |f(\boldsymbol{x}_{<k},t)|^2 + |\partial_k f(\boldsymbol{x}_{<k},t)|^2\mathrm{d}t\right)\eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x}_{<k} \\ &\leq C_T\int\int_0^1 \Big(|f(\boldsymbol{x}_{<k},t)|^2 + |\partial_k f(\boldsymbol{x}_{<k},t)|^2\Big)\eta_{<k}(\boldsymbol{x}_{<k})\eta_1(t)\mathrm{d}t\mathrm{d}\boldsymbol{x}_{<k} \\ &\leq C_T\int\int_{-\infty}^{+\infty} \Big(|f(\boldsymbol{x}_{<k},t)|^2 + |\partial_k f(\boldsymbol{x}_{<k},t)|^2\Big)\eta_{\leq k}(\boldsymbol{x}_{<k},t)\mathrm{d}t\mathrm{d}\boldsymbol{x}_{<k} \\ &= C_T\|f\|_{V_k}^2,\end{aligned}$$

where $C_T = 2\sup_{0\leq t\leq 1}\eta_1(t)^{-1}$. □

A.4 Proof of Proposition 1

The proof relies on Theorem 3.3.1 and on the following generalized integral Hardy inequality, see [153].

Lemma A.4.1. *Let $\eta_{\leq k}$ be the standard Gaussian density on $\mathbb{R}^k$. Then there exists a constant C_H such that for any $v \in L^2_\eta(\mathbb{R}^k)$,*

$$\int\left(\int_0^{x_k} v(\boldsymbol{x}_{<k},t)\mathrm{d}t\right)^2\eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \leq C_H\int v(\boldsymbol{x})^2\eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{A.3}$$

Proof of Lemma A.4.1. Let us recall the integral Hardy inequality [153].

Theorem A.4.2 (from [153]). *For weight $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ and $u \in L^2_\rho(\mathbb{R})$, there exists a constant $C_H < \infty$ such that*

$$\int_0^{+\infty} \left(\int_0^x u(t)\mathrm{d}t \right)^2 \rho(x)\mathrm{d}x \leq C_H \int_0^{+\infty} u(x)^2 \rho(x)\mathrm{d}x \tag{A.4}$$

if and only if

$$\sup_{x>0} \left(\int_x^{+\infty} \rho(t)\mathrm{d}t \right)^{1/2} \left(\int_0^x \rho(t)^{-1}\mathrm{d}t \right)^{1/2} < +\infty. \tag{A.5}$$

We apply Theorem A.4.2 with the one-dimensional standard Gaussian $\rho = \eta_1$. In order to check the condition (A.5), we need to show that

$$D(x) \coloneqq \left(\int_x^{+\infty} \rho(t)\mathrm{d}t \right)^{1/2} \left(\int_0^x \rho(t)^{-1}\mathrm{d}t \right)^{1/2} = \left(\int_x^{+\infty} e^{-t^2/2} dt \right)^{1/2} \left(\int_0^x e^{t^2/2} dt \right)^{1/2},$$

is bounded. Since $x \mapsto D(x)$ is a continuous function with a finite limit as $x \to 0$, it is sufficient to show that $D(x)$ has a finite limit when $x \to \infty$. For $x > 1$, $\int_x^{+\infty} e^{-t^2/2} dt \leq e^{-x^2/2}$ and $D(x)^2 \leq e^{-x^2/2} \int_0^x e^{t^2/2} dt$. Furthermore, using integration-by-parts we have $\int_0^x e^{t^2/2} dt = \int_0^1 e^{t^2/2} dt + e^{x^2/2}/x - \sqrt{e} + \int_1^x e^{t^2/2}/t^2 dt$. As $x \to \infty$ the dominating term in the sum is $e^{x^2/2}/x$. Thus, $e^{-x^2/2} \int_0^x e^{t^2/2} dt$ behaves asymptotically as $\mathcal{O}(\frac{1}{x})$, so that $D(x) \to 0$ when $x \to \infty$. Thus, condition (A.5) is satisfied.

Thus, by the Hardy inequality in (A.4) for $u \in L^2_\eta(\mathbb{R})$ we have

$$\int_0^{+\infty} \left(\int_0^{x_k} u(t)\mathrm{d}t \right)^2 \eta(x_k)\mathrm{d}x_k \leq C_H \int_0^{+\infty} u(x_k)^2 \eta(x_k)\mathrm{d}x_k. \tag{A.6}$$

For the symmetric density $\eta(x_k) = \eta(-x_k)$ we also have

$$\int_{-\infty}^0 \left(\int_0^{x_k} u(t)\mathrm{d}t \right)^2 \eta(x_k)\mathrm{d}x_k \leq C_H \int_{-\infty}^0 u(x_k)^2 \eta(x_k)\mathrm{d}x_k. \tag{A.7}$$

Combining the results in (A.6) and (A.7) we have

$$\int_{-\infty}^{+\infty} \left(\int_0^{x_k} u(t)\mathrm{d}t \right)^2 \eta(x_k)\mathrm{d}x_k \leq C_H \int_{-\infty}^{+\infty} u(x_k)^2 \eta(x_k)\mathrm{d}x_k.$$

Setting $u(t) = v(\boldsymbol{x}_{<k}, t)$ and integrating both sides over $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$ with the standard

Gaussian weight function $\eta(\boldsymbol{x}_{<k})$ gives the result. □

We now prove Proposition 1.

Proof. By Proposition 3.3.1, Lemma A.4.1 and by the Lipschitz property of g, we can write

$$\begin{aligned}
\|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|^2_{L^2_{\eta\leq k}} &\leq 2\int \Big(f_1(\boldsymbol{x}_{<k},0) - f_2(\boldsymbol{x}_{<k},0)\Big)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad + 2\int \Big(\int_0^{x_k} g\big(\partial_k f_1(\boldsymbol{x}_{<k},t)\big) - g\big(\partial_k f_2(\boldsymbol{x}_{<k},t)\big)\mathrm{d}t\Big)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq 2C_T\|f_1 - f_2\|^2_{V_k} + 2C_H\|g(\partial_k f_1) - g(\partial_k f_2)\|^2_{L^2_{\eta\leq k}} \\
&\leq 2C_T\|f_1 - f_2\|^2_{V_k} + 2C_H L^2\|\partial_k f_1 - \partial_k f_2\|^2_{L^2_{\eta\leq k}} \\
&\leq 2(C_T + C_H L^2)\|f_1 - f_2\|^2_{V_k}, \qquad \text{(A.8)}
\end{aligned}$$

any $f_1, f_2 \in V_k$. Furthermore, using the Lipschitz property of g we have

$$\begin{aligned}
\|\partial_k\mathcal{R}_k(f_1) - \partial_k\mathcal{R}_k(f_2)\|^2_{L^2_{\eta\leq k}} &= \int \Big(g(\partial_k f_1(\boldsymbol{x}_{<k},t)) - g(\partial_k f_2(\boldsymbol{x}_{<k},t))\Big)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq L^2\int \Big(\partial_k f_1(\boldsymbol{x}_{<k},t) - \partial_k f_2(\boldsymbol{x}_{<k},t)\Big)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\leq L^2\|f_1 - f_2\|^2_{V_k}. \qquad \text{(A.9)}
\end{aligned}$$

Combining (A.8) with (A.9), we obtain (3.20) with $C = \sqrt{2(C_T + C_H L^2) + L^2}$.

It remains to show that $\|\mathcal{R}_k(f)\|_{V_k} < \infty$ for any $f \in V_k$. Letting with $f_1 = f$ and $f_2 = 0$ in (3.20), the triangle inequality yields

$$\|\mathcal{R}_k(f)\|_{V_k} \leq \|\mathcal{R}_k(0)\|_{V_k} + C\|f\|_{V_k}$$

Because $\mathcal{R}_k(0) : \boldsymbol{x} \mapsto g(0)x_k$ is the affine function, we have $\|\mathcal{R}_k(0)\|^2_{L^2_{\eta\leq k}} = g(0)^2\int x_k^2\eta(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and $\|\partial_k\mathcal{R}_k(0)\|^2_{L^2_{\eta\leq k}} = g(0)^2$ are finite and so is $\|\mathcal{R}_k(0)\|_{V_k}$. Thus, $\mathcal{R}_k(f) \in V_k$ for all $f \in V_k$. □

A.5 Proof of Proposition 2

Proof. For any $f \in V_k$ we have

$$\begin{aligned}
|\mathcal{L}_k(f)| &= \left| \int \left(\frac{1}{2}\mathcal{R}_k(f)^2 - \log(\partial_k \mathcal{R}_k(f)) \right) \mathrm{d}\pi \right| \\
&\overset{(3.21)}{\leq} \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\leq k}}} + C_\pi \int |\log(g(\partial_k f))| \, \mathrm{d}\eta_{\leq k} \\
&\leq \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\leq k}}} + C_\pi \int |g(0)| + |\log(g(\partial_k f)) - g(0)| \, \mathrm{d}\eta_{\leq k} \\
&\overset{(3.23)}{\leq} \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\leq k}}} + C_\pi |g(0)| + C_\pi L \int |\partial_k f - 0| \, \mathrm{d}\eta_{\leq k} \\
&\leq \frac{C_\pi}{2} \|\mathcal{R}_k(f)\|^2_{L^2_{\eta_{\leq k}}} + C_\pi |g(0)| + C_\pi L \|f\|^2_{V_k}.
\end{aligned}$$

Because Proposition 1 ensures $\mathcal{R}_k(f) \in V_k \subset L^2_{\eta_{\leq k}}$, we have that $\mathcal{L}_k(f)$ is finite for any $f \in V_k$. Now, for any $f_1, f_2 \in V_k$, we can write

$$\begin{aligned}
|\mathcal{L}_k(f_1) - \mathcal{L}_k(f_2)| &= \left| \int \left(\frac{1}{2}\mathcal{R}_k(f_1)^2 - \frac{1}{2}\mathcal{R}_k(f_2)^2 - \log(\partial_k \mathcal{R}_k(f_1)) + \log(\partial_k \mathcal{R}_k(f_2)) \right) \mathrm{d}\pi \right| \\
&\overset{(3.21)}{\leq} C_\pi \int \frac{1}{2} \Big| \mathcal{R}_k(f_1)^2 - \mathcal{R}_k(f_2)^2 \Big| + \Big| \log(g(\partial_k f_1)) - \log(g(\partial_k (f_2)) \Big| \mathrm{d}\eta \\
&\overset{(3.23)}{\leq} \frac{C_\pi}{2} \|\mathcal{R}_k(f_1) + \mathcal{R}_k(f_2)\|_{L^2_{\eta_{\leq k}}} \|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{L^2_{\eta_{\leq k}}} + C_\pi L \|\partial_k f_1 - \partial_k f_2\|_{L^2_{\eta_{\leq k}}} \\
&\overset{(3.20)}{\leq} C_\pi \frac{\|\mathcal{R}_k(f_1)\|_{L^2_{\eta_{\leq k}}} + \|\mathcal{R}_k(f_2)\|_{L^2_{\eta_{\leq k}}}}{2} C \|f_1 - f_2\|_{V_k} + C_\pi L \|f_1 - f_2\|_{V_k}.
\end{aligned}$$

This shows that $\mathcal{L}_k \colon V_k \to \mathbb{R}$ is continuous. To show that $\mathcal{L}_k$ is differentiable, we let $f, \varepsilon \in V_k$ so that

$$\begin{aligned}
\mathcal{L}_k(f + \varepsilon) &= \int \left(\frac{1}{2}\mathcal{R}_k(f + \varepsilon)^2 - \log(\partial_k \mathcal{R}_k(f + \varepsilon)) \right) \mathrm{d}\pi \\
&= \int \left(\frac{1}{2}\mathcal{R}_k(f)(\boldsymbol{x})^2 + \mathcal{R}_k(f)(\boldsymbol{x}) \varepsilon(\boldsymbol{x}_{<k}, 0) \right. \\
&\quad \left. + \mathcal{R}_k(f)(\boldsymbol{x}) \left(\int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\quad - \int \log \circ g(\partial_k f) + (\log \circ g)'(\partial_k f) \partial_k \varepsilon \mathrm{d}\pi + \mathcal{O}(\|\varepsilon\|^2_{V_k}) \\
&= \mathcal{L}_k(f) + \ell(\varepsilon) + \mathcal{O}(\|\varepsilon\|^2_{V_k})
\end{aligned}$$

where $\ell\colon V_k \to \mathbb{R}$ is the linear form defined by

$$\ell(\varepsilon) = \int \mathcal{R}_k(f)(\boldsymbol{x}) \left(\varepsilon(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) - (\log \circ g)'(\partial_k f(\boldsymbol{x})) \partial_k \varepsilon(\boldsymbol{x}) \ \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

If ℓ is continuous, meaning if there exists a constant C_ℓ such that $|\ell(\varepsilon)| \leq C_\ell \|\varepsilon\|_{V_k}$ for any $\varepsilon \in V_k$, then the Riesz representation theorem states that there exists a vector $\nabla \mathcal{J}_k(f) \in V_k$ such that $\ell(\varepsilon) = \langle \nabla \mathcal{J}_k(f), \varepsilon \rangle_{V_k}$. This proves $\mathcal{J}$ is differentiable everywhere.

To show that ℓ is continuous, we write

$$\begin{aligned}
|\ell(\varepsilon)| &\overset{(3.21)}{\leq} C_\pi \int \left| \mathcal{R}_k(f)(\boldsymbol{x}) \left(\varepsilon(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) \right| \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\quad + C_\pi \int \left| (\log \circ g)'(\partial_k f(\boldsymbol{x})) \partial_k \varepsilon(\boldsymbol{x}) \right| \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\overset{(3.23)}{\leq} C_\pi \|\mathcal{R}_k(f)\|_{L^2_{\eta_{\leq k}}} \sqrt{\int \left| \varepsilon(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right|^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} \\
&\quad + C_\pi L \|\partial_k \varepsilon\|_{L^2_{\eta_{\leq k}}} \\
&\overset{(3.22)}{\leq} C_\pi \|\mathcal{R}_k(f)\|_{L^2_{\eta_{\leq k}}} \sqrt{2C_T \|\varepsilon\|^2_{V_k} + 2C_H L^2 \int \left| \partial_k \varepsilon(\boldsymbol{x})) \right|^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}} + C_\pi L \|\partial_k \varepsilon\|_{L^2_{\eta_{\leq k}}} \\
&\leq C_\pi \Big(\|\mathcal{R}_k(f)\|_{L^2_{\eta_{\leq k}}} \sqrt{2C_T + 2C_H L^2} + L \Big) \|\varepsilon\|_{V_k}.
\end{aligned}$$

This concludes the proof. □

A.6 Proof of the local Lipschitz regularity (3.25)

Proposition 19. *In addition to the assumptions of Proposition 2, we further assume there exists a constant $L < \infty$ such that for all $\xi, \xi' \in \mathbb{R}$ we have*

$$|g'(\xi) - g'(\xi')| \leq L|\xi - \xi'| \tag{A.10}$$

$$|(\log \circ g)'(\xi) - (\log \circ g)'(\xi')| \leq L|\xi - \xi'|. \tag{A.11}$$

Then there exists $M < \infty$ such that

$$\|\nabla \mathcal{L}_k(f_1) - \nabla \mathcal{L}_k(f_2)\|_{V_k} \leq M(1 + \|\mathcal{R}_k(f_2)\|_{V_k})\|f_1 - f_2\|_{\overline{V}_k},$$

for any $f_1, f_2 \in \overline{V}_k$, where $\overline{V}_k = \{f \in V_k, \partial_k f \in L^\infty\}$ is the space endowed with the norm $\|f\|_{\overline{V}_k} = \|f\|_{V_k} + \|\partial_k f\|_{L^\infty}$.

Proof. Recall the definition (3.24) of $\nabla \mathcal{L}_k(f)$

$$\begin{aligned}\langle \nabla \mathcal{L}_k(f), \varepsilon \rangle_{V_k} = &\int \mathcal{R}_k(f)(\boldsymbol{x}) \left(\varepsilon(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g'\big(\partial_k f(\boldsymbol{x}_{<k}, t)\big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &- \int (\log \circ g)'(\partial_k f(\boldsymbol{x})) \partial_k \varepsilon(\boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned}$$

Then for any $f_1, f_2 \in \overline{V}_k$. we can write

$$\langle \nabla \mathcal{L}_k(f_1) - \nabla \mathcal{L}_k(f_2), \varepsilon \rangle_{V_k} = A + B + C + D,$$

where

$$\begin{aligned}A &= \int \Big(\mathcal{R}_k(f_1)(\boldsymbol{x}) - \mathcal{R}_k(f_2)(\boldsymbol{x}) \Big) \varepsilon(\boldsymbol{x}_{<k}, 0) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ B &= \int \Big(\mathcal{R}_k(f_1)(\boldsymbol{x}) - \mathcal{R}_k(f_2)(\boldsymbol{x}) \Big) \left(\int_0^{x_k} g'(\partial_k f_1(\boldsymbol{x}_{<k}, t)) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ C &= \int \mathcal{R}_k(f_2)(\boldsymbol{x}) \left(\int_0^{x_k} \Big(g'(\partial_k f_1(\boldsymbol{x}_{<k}, t)) - g'(\partial_k f_2(\boldsymbol{x}_{<k}, t)) \Big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ D &= \int \Big((\log \circ g)'(\partial_k f_1(\boldsymbol{x})) - (\log \circ g)'(\partial_k f_2(\boldsymbol{x})) \Big) \partial_k \varepsilon(\boldsymbol{x}) \pi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.\end{aligned}$$

For the first term A we write

$$\begin{aligned}|A| &\overset{(3.21)}{\leq} C_\pi \int \Big| \mathcal{R}_k(f_1)(\boldsymbol{x}) - \mathcal{R}_k(f_2)(\boldsymbol{x}) \Big| |\varepsilon(\boldsymbol{x}_{<k}, 0)| \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &\leq C_\pi \|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{V_k} \left(\int |\varepsilon(\boldsymbol{x}_{<k}, 0)|^2 \eta_{<k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\ &\overset{(3.17)}{\leq} C_\pi \sqrt{C_T} \|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{V_k} \|\varepsilon\|_{V_k} \\ &\overset{(3.20)}{\leq} C_\pi \sqrt{C_T} C \|f_1 - f_2\|_{V_k} \|\varepsilon\|_{V_k}.\end{aligned}$$

For the second term B we write

$$
\begin{aligned}
|B| &\overset{(3.21)}{\leq} C_\pi \|\mathcal{R}_k(f_1) - \mathcal{R}_k(f_2)\|_{V_k} \left(\int \left(\int_0^{x_k} g'(\partial_k f_1(\boldsymbol{x}_{<k}, t)) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\overset{(A.3)}{\underset{(3.20)}{\leq}} C_\pi \sqrt{C_H} C \|f_1 - f_2\|_{V_k} \Big(\int \Big(g'(\partial_k f_1(\boldsymbol{x}_{\leq k})) \partial_k \varepsilon(\boldsymbol{x}_{\leq k}) \Big)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Big)^{1/2} \\
&\overset{(3.22)}{\leq} C_\pi \sqrt{C_H} C L \|f_1 - f_2\|_{V_k} \Big(\int \Big(\partial_k \varepsilon(\boldsymbol{x}_{\leq k}) \Big)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Big)^{1/2} \\
&\leq C_\pi \sqrt{C_H} C L \|f_1 - f_2\|_{V_k} \|\varepsilon\|_{V_k}.
\end{aligned}
$$

For the third term C we write

$$
\begin{aligned}
|C| &\overset{(3.21)}{\leq} C_\pi \|\mathcal{R}_k(f_2)\|_{V_k} \left(\int \left(\int_0^{x_k} \Big(g'(\partial_k f_1(\boldsymbol{x}_{<k}, t)) - g'(\partial_k f_2(\boldsymbol{x}_{<k}, t)) \Big) \partial_k \varepsilon(\boldsymbol{x}_{<k}, t) \mathrm{d}t \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\overset{(A.3)}{\leq} C_\pi \sqrt{C_H} \|\mathcal{R}_k(f_2)\|_{V_k} \left(\int \left(\Big(g'(\partial_k f_1(\boldsymbol{x}_{\leq k})) - g'(\partial_k f_2(\boldsymbol{x}_{\leq k})) \Big) \partial_k \varepsilon(\boldsymbol{x}_{\leq k}) \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\leq C_\pi \sqrt{C_H} \|\mathcal{R}_k(f_2)\|_{V_k} \left(\operatorname{ess\,sup} \left| g' \circ \partial_k f_1 - g' \circ \partial_k f_2 \right| \right) \left(\int \left(\partial_k \varepsilon(\boldsymbol{x}_{\leq k}) \right)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \\
&\overset{(A.10)}{\leq} C_\pi \sqrt{C_H} L \|\mathcal{R}_k(f_2)\|_{V_k} \left(\operatorname{ess\,sup} \left| \partial_k f_1 - \partial_k f_2 \right| \right) \|\varepsilon\|_{V_k} \\
&\leq C_\pi \sqrt{C_H} L \|\mathcal{R}_k(f_2)\|_{V_k} \|f_1 - f_2\|_{\overline{V}_k} \|\varepsilon\|_{V_k}.
\end{aligned}
$$

For the last term D we write

$$
\begin{aligned}
|D| &\overset{(3.21)}{\leq} C_\pi \left(\int \Big((\log \circ g)'(\partial_k f_1(\boldsymbol{x})) - (\log \circ g)'(\partial_k f_2(\boldsymbol{x})) \Big)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \|\varepsilon\|_{V_k} \\
&\overset{(A.11)}{\leq} C_\pi L \left(\int \Big(\partial_k f_1(\boldsymbol{x}) - \partial_k f_2(\boldsymbol{x}) \Big)^2 \eta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right)^{1/2} \|\varepsilon\|_{V_k} \\
&\leq C_\pi L \|f_1 - f_2\|_{V_k} \|\varepsilon\|_{V_k}.
\end{aligned}
$$

Thus, because $\|f_1 - f_2\|_{V_k} \leq \|f_1 - f_2\|_{\overline{V}_k}$ we obtain

$$
\begin{aligned}
\frac{|\langle \nabla \mathcal{L}_k(f_1) - \nabla \mathcal{L}_k(f_2), \varepsilon \rangle_{V_k}|}{\|\varepsilon\|_{V_k}} &\leq C_\pi \Big(\sqrt{C_T} C + \sqrt{C_H} C L + \sqrt{C_H} L \|\mathcal{R}_k(f_2)\|_{V_k} + L \Big) \|f_1 - f_2\|_{\overline{V}_k} \\
&\leq M(1 + \|\mathcal{R}_k(f_2)\|_{V_k}) \|f_1 - f_2\|_{\overline{V}_k},
\end{aligned}
$$

where

$$M = C_\pi \max\{\sqrt{C_T}C + \sqrt{C_H}CL + L; \sqrt{C_H}L\}.$$

This concludes the proof. □

A.7 Proof of Proposition 3

Proof. Let $s_1, s_2 \in V_k$ be strictly increasing functions with respect to x_k that satisfy $\partial_k s_i(\boldsymbol{x}_{\leq k}) \geq c$ for $i = 1, 2$. By the Lipschitz property of g^{-1} on the domain $[c, \infty)$ with constant L_c, we can write

$$\begin{aligned}
\|\partial_k \mathcal{R}_k^{-1}(s_1) - \partial_k \mathcal{R}_k^{-1}(s_2)\|^2_{L^2_{\eta_{\leq k}}} &= \int (g^{-1}(\partial_k s_1(\boldsymbol{x}_{\leq k})) - g^{-1}(\partial_k s_2(\boldsymbol{x}_{\leq k})))^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\leq L_c^2 \int (\partial_k s_1(\boldsymbol{x}_{\leq k}) - \partial_k s_2(\boldsymbol{x}_{\leq k}))^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\leq L_c^2 \|s_1 - s_2\|^2_{V_k}. \qquad \text{(A.12)}
\end{aligned}$$

Applying Proposition 3.3.1 for $s_1, s_2 \in V_k$ and Lemma A.4.1 for $\partial_k \mathcal{R}_k^{-1}(s_i) = g^{-1}(\partial_k s_i) \in L^2_{\eta_{\leq k}}$ with $i = 1, 2$ we have

$$\begin{aligned}
\|\mathcal{R}_k^{-1}(s_1) - \mathcal{R}_k^{-1}(s_2)\|^2_{L^2_{\eta_{\leq k}}} &\leq 2 \int \left(s_1(\boldsymbol{x}_{<k}, 0) - s_2(\boldsymbol{x}_{<k}, 0)\right)^2 \eta_{\leq k}(\boldsymbol{x}) d\boldsymbol{x} \\
&\quad + 2 \int \left(\int_0^x g^{-1}(\partial_k s_1) - g^{-1}(\partial_k s_2) \mathrm{d}t \right)^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\leq 2C_T \|s_1 - s_2\|^2_{V_k} \\
&\quad + 2C_H \int (g^{-1}(\partial_k s_1(\boldsymbol{x}_{\leq k})) - g^{-1}(\partial_k s_2(\boldsymbol{x}_{\leq k})))^2 \eta_{\leq k}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&\leq (2C_T + 2C_H L_c^2) \|s_1 - s_2\|^2_{V_k}, \qquad \text{(A.13)} \\
& \qquad \text{(A.14)}
\end{aligned}$$

where the last inequality follows from the continuity of the partial derivative in (A.12). Combining (A.12) with (A.13), we obtain (3.27) with $C_c = (2C_T + 2C_H L_c^2)$.

It remains to show that $\|\mathcal{R}_k^{-1}(s)\|_{V_k} < \infty$ for any $s \in V_{k,g(M)}$. Letting $s_1 = s$ and

$s_2 = g(0)x_k$, the triangle inequality yields

$$\|\mathcal{R}_k^{-1}(s)\|_{V_k} \leq \|\mathcal{R}_k^{-1}(g(0)x_k)\|_{V_k} + C_M\|s - g(0)x_k\|_{V_k}.$$

The function $\mathcal{R}_k^{-1}(g(0)x_k)$ is zero. Therefore, $\|\mathcal{R}_k^{-1}(s)\|_{V_k} \leq C_M\|s - g(0)x_k\|_{V_k} \leq C_M(\|s\|_{V_k} + \|g(0)x_k\|_{V_k})$. For a linear function, $\|g(0)x_k\|_{V_k}^2 = \|g(0)x_k\|_{L^2_{\eta_{\leq k}}}^2 + \|g(0)\|_{L^2_{\eta_{\leq k}}}^2$ is finite, and so $\|\mathcal{R}_k^{-1}(s)\|_{V_k} < \infty$ for $s \in V_k$. Furthermore, if $\partial_k s \geq c > 0$, then $\partial_k \mathcal{R}_k^{-1}(s) = g^{-1}(\partial_k s) \geq g^{-1}(c) > -\infty$ and so $\operatorname{ess\,inf} \mathcal{R}_k^{-1}(s) > -\infty$. □

A.8 Proof of Proposition 4

Proof. To show that $\mathcal{R}_k(V_k) = \{\mathcal{R}(f) : f \in V_k\}$ is convex, let $f_1, f_2 \in V_k$ and $0 \leq \alpha \leq 1$. We need to show that there exists $f_\alpha \in V_k$ such that $\mathcal{R}(f_\alpha) = S_\alpha$ where

$$S_\alpha \coloneqq \alpha\mathcal{R}_k(f_1) + (1-\alpha)\mathcal{R}_k(f_2).$$

Let

$$f_\alpha(\boldsymbol{x}_{\leq k}) \coloneqq \mathcal{R}^{-1}(S_\alpha)(\boldsymbol{x}_{\leq k}) = S_\alpha(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t.$$

It remains to show that $f_\alpha \in V_k$, meaning that $f_\alpha \in L^2_{\eta_{\leq k}}$ and $\partial_k f_\alpha \in L^2_{\eta_{\leq k}}$. By convexity of $\xi \mapsto g^{-1}(\xi)^2$, we have

$$\begin{aligned}
\|\partial_k f_\alpha\|_{L^2_{\eta_{\leq k}}}^2 &= \int g^{-1}\big(\alpha\partial_k\mathcal{R}_k(f_1) + (1-\alpha)\partial_k\mathcal{R}_k(f_1)\big)^2\mathrm{d}\eta_{\leq k} \\
&= \int g^{-1}\big(\alpha g(\partial_k f_1) + (1-\alpha)g(\partial_k f_2)\big)^2\mathrm{d}\eta_{\leq k} \\
&\leq \int \alpha g^{-1}(g(\partial_k f_1))^2 + (1-\alpha)g^{-1}(g(\partial_k f_1))^2\mathrm{d}\eta_{\leq k} \\
&= \alpha\|\partial_k f_1\|_{L^2_{\eta_{\leq k}}}^2 + (1-\alpha)\|\partial_k f_2\|_{L^2_{\eta_{\leq k}}}^2. \qquad \text{(A.15)}
\end{aligned}$$

Thus $\partial_k f_\alpha \in L^2_{\eta_{\leq k}}$. Furthermore we have

$$\begin{aligned}\|f_\alpha\|^2_{L^2_{\eta_{\leq k}}} &= \int \left(S_\alpha(\boldsymbol{x}_{<k}, 0) + \int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t\right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &\leq 2\int S_\alpha(\boldsymbol{x}_{<k}, 0)^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x} + 2\int \left(\int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t\right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

To show that the above quantity is finite, Theorem 3.17 permits us to write

$$\begin{aligned}\int S_\alpha(\boldsymbol{x}_{<k}, 0)^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x} &= \int \Big(\alpha f_1(\boldsymbol{x}_{<k}, 0) + (1-\alpha) f_2(\boldsymbol{x}_{<k}, 0)\Big)^2 \eta_{<k}(\boldsymbol{x}_{<k})\mathrm{d}\boldsymbol{x} \\ &\leq C_T \|\alpha f_1 + (1-\alpha) f_2\|^2_{V_k},\end{aligned}$$

which is finite. Finally, because $g^{-1}(\partial_k S_\alpha) = \partial_k f_\alpha \in L^2_{\eta_{\leq k}}$ by (A.15), Lemma A.4.1 yields

$$\begin{aligned}\int \left(\int_0^{x_k} g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{<k}, t))\mathrm{d}t\right)^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &\leq C_H \int g^{-1}(\partial_k S_\alpha(\boldsymbol{x}_{\leq k}))^2 \eta_{\leq k}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= C_H \|\partial_k f_\alpha\|^2_{L^2_{\eta_{\leq k}}},\end{aligned}$$

which is finite. We deduce that $f_\alpha \in L^2_{\eta_{\leq k}}$ and therefore that $f_\alpha \in V_k$. □

A.9 Proof for the KR rearrangement

Proof. Let $S_{\mathrm{KR},k}$ be the kth component of the KR rearrangement, given by composing the inverse CDF of the standard Gaussian marginal $F_{\eta,k}(x_k)$ with the CDF of the target's kth marginal conditional $F_{\pi_k}(x_k|\boldsymbol{x}_{<k})$. That is,

$$S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = F^{-1}_{\eta_k} \circ F_{\pi_k}(x_k|\boldsymbol{x}_{<k}). \tag{A.16}$$

Differentiating the above formula yields

$$\partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = \frac{\pi_k(x_k|\boldsymbol{x}_{<k})}{\eta_k(S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}))} = \frac{\pi_k(F^{-1}_{\pi_k}(F_{\pi_k}(x_k|\boldsymbol{x}_{<k})|\boldsymbol{x}_{<k})|\boldsymbol{x}_{<k})}{\eta_k(F^{-1}_{\eta_k} \circ F_{\pi_k}(x_k|\boldsymbol{x}_{<k}))}, \tag{A.17}$$

where $F_{\pi_k}^{-1}(\cdot|\boldsymbol{x}_{<k})$ denotes the inverse of the map $x_k \mapsto F_{\pi_k}(x_k|\boldsymbol{x}_{<k})$ for each $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$.

The goal is to show $S_{\mathrm{KR},k} \in V_k$, that is $S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$ and $\partial_k S_{\mathrm{KR},k} \in L^2_{\eta_{\leq k}}$. To do so, it is sufficient to characterize the tail behavior of the map. Without loss of generality we consider the left tail. From condition (3.31), we have

$$F_{\eta_k}^{-1}(c_k F_{\eta_k}(x_k)) \leq S_{\mathrm{KR},k}(x_k|\boldsymbol{x}_{<k}) \leq F_{\eta_k}^{-1}(C_k F_{\eta_k}(x_k)).$$

for all $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$. From Theorems 1 and 2 in [38], there exists strictly positive constants $\alpha_i, \beta_i > 0$ for $i = 1, 2$ such that

$$\alpha_1 \exp(-\beta_1 x_k^2) \leq F_{\eta_k}(x_k) \leq \alpha_2 \exp(-\beta_2 x_k^2) \tag{A.18}$$

for $x_k < 0$, and so we also have

$$\sqrt{1/\beta_2 \log(\alpha_2/u)} \leq F_{\eta_k}^{-1}(u) \leq \sqrt{1/\beta_1 \log(\alpha_1/u)}. \tag{A.19}$$

for $u < F_{\eta_k}(0)$. Combining (A.18) and (A.19)

$$\sqrt{\frac{1}{\beta_2}\log\left(\frac{\alpha_2}{c_k\alpha_1}\right) + \frac{\beta_1}{\beta_2}x_k^2} \leq S_{\mathrm{KR},k}(x_k|\boldsymbol{x}_{<k}) \leq \sqrt{\frac{1}{\beta_1}\log\left(\frac{\alpha_1}{C_k\alpha_2}\right) + \frac{\beta_2}{\beta_1}x_k^2}$$

for all $\boldsymbol{x}_{<k} \in \mathbb{R}^{k-1}$ and $x_k < 0$. Thus, $S_{\mathrm{KR},k}(x_k|\boldsymbol{x}_{<k}) = \mathcal{O}(x_k)$ as $x_k \to -\infty$. Given that the squared moments of η_k are finite, we have $S_{\mathrm{KR},k} \in L^2_{\eta_k}$.

To bound the derivative $\partial_k S_{\mathrm{KR},k}$ in (A.17) we consider the following limit

$$\begin{aligned}
\lim_{x_k \to -\infty} \partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) &= \lim_{u \to 0^+} \frac{\pi_k(F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})|\boldsymbol{x}_{<k})}{\eta_k(F_{\eta_k}^{-1}(u))} \\
&= \lim_{u \to 0^+} \frac{(F_{\eta,k}^{-1})'(u)}{(F_{\pi_k}^{-1})'(u|\boldsymbol{x}_{<k})} \\
&= \lim_{u \to 0^+} \frac{F_{\eta_k}^{-1}(u)}{F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})},
\end{aligned} \tag{A.20}$$

where in the second equality we used the implicit function theorem and the third

equality follows from l'Hôpital's rule. To analyze the ratio $F_{\eta_k}^{-1}(u)/F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})$, we combine the lower bound in (3.31) and the bounds in (A.18) to get

$$\frac{F_{\eta_k}^{-1}(u)}{F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})} \leq \frac{F_{\eta_k}^{-1}(u)}{F_{\eta_k}^{-1}(u/c_k)} \leq \sqrt{\frac{\beta_2(\log\alpha_1 - \log(u))}{\beta_1(\log\alpha_2 - \log(u/c_k))}}.$$

Similarly, from the upper bound in (3.31) and the bounds in (A.18), we have

$$\frac{F_{\eta_k}^{-1}(u)}{F_{\pi_k}^{-1}(u|\boldsymbol{x}_{<k})} \geq \frac{F_{\eta_k}^{-1}(u)}{F_{\eta_k}^{-1}(u/C_k)} \geq \sqrt{\frac{\beta_1(\log\alpha_2 - \log(u))}{\beta_2(\log\alpha_1 - \log(u/C_k))}}.$$

Thus, $\partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) = \mathcal{O}(1)$ as $x_k \to -\infty$, and we have $\partial_k S_{\mathrm{KR},k} \in L^2_{\eta_k}$.

Lastly, taking the limit in (A.20) we have $\lim_{x_k\to-\infty} \partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) \geq \sqrt{\beta_1/\beta_2}$. For a target distribution π with full support, all marginal conditional densities satisfy $\pi_k(x_k|\boldsymbol{x}_{<k}) > 0$ for each $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. Given that the $\partial_k S_{\mathrm{KR},k}$ does not approach zero as $|x_k| \to \infty$, we can find a strictly positive constant $c_k > 0$ such that $\partial_k S_{\mathrm{KR},k}(\boldsymbol{x}_{\leq k}) \geq c_k$ for all $\boldsymbol{x}_{\leq k} \in \mathbb{R}^k$. This shows that $\operatorname{ess\,inf} \partial_k S_{\mathrm{KR},k} > 0$. □

Appendix B

Additional details for Chapter 3

B.1 Multi-index refinement for the wavelet basis

In this section we show how to greedily enrich the index set Λ for a one-dimensional wavelet basis parameterized by the tuple of indices (l, m) representing the level l and translation m of each wavelet $\psi_{(l,m)}$. To define the allowable indices, we construct a binary tree where each node is indexed by (l, m) and has two children with indices $(l+1, 2m)$ and $(l+1, 2m+1)$. The root of the tree has index $(0, 0)$ and corresponds to the mother wavelet ψ. Analogously to the downward closed property for polynomial indices, we only add nodes to the tree (i.e., indices in Λ) if its parents have already been added. Thus, given any multi-index set Λ_t, we define its reduced margin as

$$\Lambda_t^{\mathrm{RM}} = \left\{ \alpha = (l, m) \notin \Lambda_t \text{ such that } \begin{array}{ll} (l-1, m/2) \in \Lambda_t & \text{for odd } m \\ (l-1, (m-1)/2) \in \Lambda_t & \text{for even } m \end{array} \right\}.$$

Then, the `ATM` algorithm with a wavelet basis follows from Algorithm 1 with this construction for the reduced margin at each iteration.

B.2 Architecture details of alternative methods

In this section we present the details of the alternative methods to `ATM` that we consider in Section 3.5.

For each normalizing flow model, we use the recommended stochastic gradient descent optimizer with a learning rate of 10^{-3}. We partition 10% of the samples in each training set to be validation samples and use the remaining samples for training the model. We select the optimal hyper-parameters for each dataset by fitting the density with the training data and choosing the parameters that minimize the negative log-likelihood of the approximate density on the validation samples. We also use the validation samples to set the termination criteria during the optimization.

We follow the implementation of [186] to define the architectures of these models. The hyper-parameters we consider for the neural networks in the `MDN` and `NF` models are: 2 hidden layers, 32 hidden units in each layer, $\{5, 10, 20, 50, 100\}$ centers or flows, weight normalization, and a dropout probability of $\{0, 0.2\}$ for regularizing the neural networks during training. For `CKDE` and `NKDE` we select the bandwidth of the kernel estimators using 5-fold cross-validation.

Appendix C

Proofs for Chapter 4

Proof of Theorem 4.2.1. The proof contains two steps: we first show the result when $d = 2$, then extend it to $d > 2$.

Assume $d = 2$. The conditional mutual information $I(X_i; X_j|\boldsymbol{X}_{-ij})$ becomes the mutual information $I(X_i; X_j)$, given by

$$\begin{aligned} I(X_i; X_j) &= \int \pi(x_i, x_j) \log\left(\frac{\pi(x_i, x_j)}{\pi(x_i)\pi(x_j)}\right) \mathrm{d}x_i \mathrm{d}x_j \\ &= \int h(\boldsymbol{x}) \log\left(\frac{h(\boldsymbol{x})}{\int h(\boldsymbol{x}')\pi(x_i')\pi(x_j')\mathrm{d}x_i'\mathrm{d}x_j'}\right) \pi(x_i)\pi(x_j)\mathrm{d}x_i\mathrm{d}x_j, \end{aligned}$$

where $h(\boldsymbol{x}) = h(x_i, x_j) = \pi(x_i, x_j)/(\pi(x_i)\pi(x_j))$. Next we apply the logarithmic Sobolev inequality to bound $I(X_i; X_j)$. To do that, we need to show that the product density $(x_i, x_j) \mapsto \pi(x_i)\pi(x_j)$ satisfies the logarithmic Sobolev inequality. By Theorem 4.4. in [81], $(x_i, x_j) \mapsto \pi(x_i)\pi(x_j)$ has a logarithmic Sobolev constant $C(\pi(x_i)\pi(x_j))$ bounded by $\max\{C(\pi(x_i)); C(\pi(x_j))\}$. By Assumption (4.8) the joint density $\pi(x_i, x_j)$ has a logarithmic Sobolev inequality bounded by C_0 and therefore the marginals $\pi(x_i)$ and $\pi(x_j)$ satisfy $C(\pi(x_i)) \leq C_0$ and $C(\pi(x_j)) \leq C_0$ (this can be easily shown by restricting (4.8) to univariate functions $h : (x_i, x_j) \mapsto h(x_i)$ or $h : (x_i, x_j) \mapsto h(x_j)$). We deduce that the product of marginals satisfies the logarith-

mic Sobolev inequality with constant $C(\pi(x_i)\pi(x_j)) \leq C_0$. We can write

$$\begin{aligned} I(X_i;X_j) &\leq \frac{C_0}{2}\int \left\| \begin{matrix} \partial_i \log h(x_i,x_j) \\ \partial_j \log h(x_i,x_j) \end{matrix} \right\|_2^2 h(x_i,x_j)\pi(x_i)\pi(x_j)\mathrm{d}x_i\mathrm{d}x_j \\ &= \frac{C_0}{2}\int \left(\int \big(\partial_i \log \pi(x_i,x_j) - \partial_i \log \pi(x_i)\big)^2 \pi(x_j|x_i)\mathrm{d}x_j \right) \pi(x_i)\mathrm{d}x_i \\ &+ \frac{C_0}{2}\int \left(\int \big(\partial_j \log \pi(x_i,x_j) - \partial_j \log \pi(x_j)\big)^2 \pi(x_i|x_j)\mathrm{d}x_i \right) \pi(x_j)\mathrm{d}x_j. \quad \text{(C.1)} \end{aligned}$$

Next we apply the Poincaré inequality to bound the two integrands in the above expression. By assumption (4.8), the density $x_i \mapsto \pi(x_i|x_j)$ has a logarithmic Sobolev constant bounded by C_0 so that it satisfies the Poincaré inequality

$$\int \left(f(x_i) - \int f(x_i')\pi(x_i'|x_j)\mathrm{d}x_i' \right)^2 \pi(x_i|x_j)\mathrm{d}x_i \leq C_0 \int f'(x_i)^2 \pi(x_i|x_j)\mathrm{d}x_i, \quad \text{(C.2)}$$

for any continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$. The Poincaré inequality (C.2) is classically obtained from a logarithmic Sobolev inequality (4.7) by letting $h = 1 + \varepsilon f$ and by taking a Taylor expansion as $\varepsilon \to 0$. Notice that

$$\int \big(\partial_j \log \pi(x_i',x_j)\big)\pi(x_i'|x_j)\mathrm{d}x_i' = \int \frac{\partial_j \pi(x_i',x_j)}{\pi(x_i',x_j)} \frac{\pi(x_i',x_j)}{\pi(x_j)}\mathrm{d}x_i' = \frac{\partial_j \int \pi(x_i',x_j)\mathrm{d}x_i'}{\pi(x_j)} = \partial_j \log \pi(x_j),$$

so that the Poincaré inequality (C.2) with $f(x_i) = \partial_j \log \pi(x_i,x_j)$ yields

$$\int \big(\partial_j \log \pi(x_i,x_j) - \partial_j \log \pi(x_j)\big)^2 \pi(x_i|x_j)\mathrm{d}x_i \leq C_0 \int \big(\partial_i\partial_j \log \pi(x_i,x_j)\big)^2 \pi(x_i|x_j)\mathrm{d}x_i.$$

In the same way, by permuting i and j, we obtain

$$\int \big(\partial_i \log \pi(x_i,x_j) - \partial_i \log \pi(x_i)\big)^2 \pi(x_j|x_i)\mathrm{d}x_j \leq C_0 \int \big(\partial_i\partial_j \log \pi(x_i,x_j)\big)^2 \pi(x_j|x_i)\mathrm{d}x_j.$$

Using the above two inequalities in (C.1) yields

$$\begin{aligned} I(X_i; X_j) &\leq \frac{C_0}{2} \int \left(C_0 \int \left(\partial_i \partial_j \log \pi(x_i, x_j)\right)^2 \pi(x_j|x_i) \mathrm{d}x_j \right) \pi(x_i) \mathrm{d}x_i \\ &\quad + \frac{C_0}{2} \int \left(C_0 \int \left(\partial_i \partial_j \log \pi(x_i, x_j)\right)^2 \pi(x_i|x_j) \mathrm{d}x_i \right) \pi(x_j) \mathrm{d}x_j \\ &= C_0^2 \int \left(\partial_i \partial_j \log \pi(x_i, x_j)\right)^2 \pi(x_i, x_j) \mathrm{d}x_i \mathrm{d}x_j. \end{aligned}$$

This shows that $I(X_i; X_j) \leq C_0^2 \Omega_{i,j}$ when $d = 2$.

Assume now that $d > 2$. For any $\boldsymbol{x}_{-ij} \in \mathbb{R}^{d-2}$, replacing $\pi(x_i, x_j)$ by $\pi(x_i, x_j|\boldsymbol{x}_{-ij})$ in the previous analysis allows us to write

$$\begin{aligned} &\int \pi(x_i, x_j|\boldsymbol{x}_{-ij}) \log \left(\frac{\pi(x_i, x_j|\boldsymbol{x}_{-ij})}{\pi(x_i|\boldsymbol{x}_{-ij}) \pi(x_j|\boldsymbol{x}_{-ij})} \right) \mathrm{d}x_i \mathrm{d}x_j \\ &\quad \leq C_0^2 \int \left(\partial_i \partial_j \log \pi(x_i, x_j|\boldsymbol{x}_{-ij})\right)^2 \pi(x_i, x_j|\boldsymbol{x}_{-ij}) \mathrm{d}x_i \mathrm{d}x_j \\ &\quad = C_0^2 \int \left(\partial_i \partial_j \log \pi(\boldsymbol{x})\right)^2 \pi(x_i, x_j|\boldsymbol{x}_{-ij}) \mathrm{d}x_i \mathrm{d}x_j, \end{aligned}$$

where the last equality is obtained by $\log \pi(x_i, x_j|\boldsymbol{x}_{-ij}) = \log \pi(\boldsymbol{x}) - \log \pi(\boldsymbol{x}_{-ij})$. Multiplying by the marginal $\pi(\boldsymbol{x}_{-ij})$ and integrating over $\boldsymbol{x}_{-ij} \in \mathbb{R}^{d-2}$ we obtain (4.9), which concludes the proof. □

Proof of Proposition 9. Let $\boldsymbol{\nu}_\rho$ be a measure on $\mathbb{R}^d$ that is Markov with respect to a graph $\mathcal{G}$ and has a strictly positive density ρ. By the Hammersley-Clifford theorem, the density ρ factorizes as

$$\rho(\boldsymbol{z}) = \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \varphi_c(\boldsymbol{z}_c), \tag{C.3}$$

where φ_c are nonnegative potential functions, $\mathcal{Z}$ is a normalizing constant and $\mathcal{C}$ is the set of maximal cliques of $\mathcal{G}$ [119]. A clique is a fully connected subset of nodes and a maximal clique is a clique that is not a strict subset of another clique.

The pullback density of ρ through a differentiable diagonal transport map D is

given by

$$\begin{aligned} D^\sharp\rho(\boldsymbol{x}) &= \rho \circ D(\boldsymbol{x}) |\det \nabla D(\boldsymbol{x})| \\ &= \rho \circ D(\boldsymbol{x}) \prod_{i=1}^{d} \partial_i D_i(x_i) \\ &= \frac{1}{\mathcal{Z}} \prod_{c \in \mathcal{C}} \varphi_c(D_c(\boldsymbol{x}_c)) \prod_{i=1}^{d} \partial_i D_i(x_i), \end{aligned}$$

where D_c represents the subset of components of D corresponding to the nodes in clique c. Collecting the derivatives of the map components for the nodes in each clique, we have

$$\begin{aligned} D^\sharp\rho(\boldsymbol{x}) &= \frac{1}{\mathcal{Z}} \left(\varphi_{c_1}(D_{c_1}(\boldsymbol{x}_{c_1})) \prod_{i \in c_1} \partial_i D_i(x_i) \right) \left(\varphi_{c_2}(D_{c_2}(\boldsymbol{x}_{c_2})) \prod_{i \in c_2; i \notin c_1} \partial_i D_i(x_i) \right) \dots \\ &\qquad \left(\varphi_{c_M}(D_{c_M}(\boldsymbol{x}_{c_M})) \prod_{i \in c_M; i \notin c_1, \dots, c_{M-1}} \partial_i D_i(x_i) \right) \\ &\equiv \frac{1}{\mathcal{Z}} \psi_{c_1}(\boldsymbol{x}_{c_1}) \psi_{c_2}(\boldsymbol{x}_{c_2}) \cdot \dots \cdot \psi_{c_M}(\boldsymbol{x}_{c_M}), \end{aligned} \tag{C.4}$$

where $\psi_{c_j}(x_{c_j})$ defines new potential functions of the variables $\boldsymbol{X}_{c_j}$ in the maximal clique c_j, and $M = |\mathcal{C}|$ represents the cardinality of $\mathcal{C}$. From (C.4), the density of $D^\sharp\rho$ also factorizes according to $\mathcal{G}$. Thus, by Proposition 3.8 in Lauritzen [119], $D^\sharp\boldsymbol{\nu}_\rho$ is Markov with respect to $\mathcal{G}$. □

We note that the contrapositive of Proposition 9 immediately follows: If the minimal I-map of $S^\sharp\boldsymbol{\nu}_\rho$ is not equivalent to that of $\boldsymbol{\nu}_\rho$, then the map S is not diagonal.

Proof of Proposition 10. Let ρ be a strictly positive density on $\mathbb{R}^d$ and let D be a differentiable diagonal transport map. The log of the pullback density $\pi = D^\sharp\rho$ is

given by

$$\log \pi(\boldsymbol{x}) = \log D^\sharp \rho(\boldsymbol{x}) = \log \rho \circ D(\boldsymbol{x}) + \log |\det \nabla D(\boldsymbol{x})|$$
$$= \log \rho \circ D(\boldsymbol{x}) + \sum_{k=1}^{d} \log \partial_k D_k(x^k). \tag{C.5}$$

The partial derivatives of the log-density with respect to x_i, x_j are given by

$$\partial_i \partial_j \log \pi(\boldsymbol{x}) = \partial_i \partial_j \log \rho \circ D(\boldsymbol{x})\, \partial_i D_i(x_i) \partial_j D_j(x_j),$$

by using that each term in the log-determinant of the map's Jacobian only depends on a single variable. Thus, entry (i, j) of the conditional independence score for π is given by

$$\Omega_{ij} = \int |\partial_i \partial_j \log \rho \circ D(\boldsymbol{x})\, \partial_i D_i(x_i) \partial_j D_j(x_j)|^2 \, \pi(\boldsymbol{x}) d\boldsymbol{x}. \tag{C.6}$$

Using the measure transformation $\boldsymbol{\nu}_\pi = D^\sharp \boldsymbol{\nu}_\rho$, we rewrite the conditional independence score of π in (C.6) as the following expectation over ρ

$$\Omega_{ij} = \int \left|\partial_i \partial_j \log \rho(\boldsymbol{z})\, \partial_i D_i(D_i^{-1}(z_i)) \partial_j D_j(D_j^{-1}(z_j))\right|^2 \rho(\boldsymbol{z}) d\boldsymbol{z}. \tag{C.7}$$

Finally, by applying the inverse function theorem to each continuously differentiable map component to get $\partial_i D_i(D_i^{-1}(z_i)) = \partial_i D_i^{-1}(z_i)$, we arrive at the score in (4.31). □

Proof of Proposition 7. For the polynomial degree $\beta = 1$, the transport map S is an affine function $S(\boldsymbol{x}) = L(\boldsymbol{x} - \mathbf{c})$ with an invertible lower triangular matrix $L \in \mathbb{R}^{d \times d}$ and vector $\mathbf{c} \in \mathbb{R}^d$. For this class of transport maps, the minimization problem in (4.15) becomes

$$\arg\max_S \frac{1}{n} \sum_{l=1}^{n} \log S^\sharp \eta(\boldsymbol{X}^l) = \arg\max_S \frac{1}{n} \sum_{l=1}^{n} \log \eta \circ S(\boldsymbol{X}^l) + \log |\det \nabla S(\boldsymbol{X}^l)|$$
$$= \arg\max_{L, \mathbf{c}} \frac{1}{n} \sum_{l=1}^{n} \left(-\frac{1}{2} (\boldsymbol{X}^l - \mathbf{c})^T L^T L (\boldsymbol{X}^l - \mathbf{c}) + \log \det L \right). \tag{C.8}$$

For any invertible matrix $L^T L$, the optimal $\mathbf{c}$ is the empirical mean $\widehat{\mathbf{m}} := \frac{1}{n}\sum_{l=1}^{n} \boldsymbol{X}^l$. Substituting this value for $\mathbf{c}$ in (C.8), the objective for L is given by

$$\arg\min_{L} \left\{ \text{Trace}\left(L^T L \widehat{\Sigma}\right) - \log\det\left(L^T L\right) \right\}, \tag{C.9}$$

where $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}^l - \mathbf{c})(\boldsymbol{X}^l - \mathbf{c})^T$ denotes the empirical covariance matrix and Trace is the matrix trace operator. Setting the gradient of (C.9) with respect to $L^T L$ to zero yields the optimal L as the inverse of the Cholesky factor of $\widehat{\Sigma}$ (which exists for $n \geq d$). Thus, the pullback density $S^\sharp \eta$ for a standard Gaussian density η yields a multivariate Gaussian approximation to π with mean $\widehat{\mathbf{m}}$ and covariance matrix $\widehat{\Sigma}$. □

Proof of Proposition 8. The proof considers two types of errors: a false positive occurs when the true score Ω_{ij} is zero but the threshold estimate $\overline{\Omega}_{ij} = \widehat{\Omega}_{ij}\mathbb{1}(\widehat{\Omega}_{ij} > \tau_{ij})$ is nonzero; and a false negative occurs when the true score Ω_{ij} is nonzero but the threshold estimate $\overline{\Omega}_{ij}$ is zero.

For each pair of variables (i, j), let $g\colon \boldsymbol{\alpha} \mapsto \mathbb{E}_\pi |\partial_i \partial_j \log S^\sharp_{\boldsymbol{\alpha}} \eta(\boldsymbol{x})|^2$ be a continuous function of the coefficients $\boldsymbol{\alpha}$. Then, we have $\widehat{\Omega}_{ij} = g(\widehat{\boldsymbol{\alpha}})$ (as defined in (4.18)) and $\widehat{\upsilon}_{ij} = (\nabla_{\boldsymbol{\alpha}} g(\widehat{\boldsymbol{\alpha}})^T \Gamma(\widehat{\boldsymbol{\alpha}})^{-1} \nabla_{\boldsymbol{\alpha}} g(\widehat{\boldsymbol{\alpha}}))^{1/2}$ (defined below (4.23)). Assuming that g is twice differentiable, a Taylor expansion of the score estimator around $\boldsymbol{\alpha}^*$ yields

$$g(\widehat{\boldsymbol{\alpha}}) = g(\boldsymbol{\alpha}^*) + \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)^T (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) + \frac{1}{2}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)^T \nabla^2_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*) + o_p(\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*\|^2), \tag{C.10}$$

where the remainder is a term that tends to zero in probability.

For conditionally dependent variables, we have $g(\boldsymbol{\alpha}^*) = \Omega_{ij} \neq 0$ and, by assumption, $\nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*) \neq 0$. Therefore, we can truncate the expansion in (C.10) to first-order terms and from the delta method we have $\sqrt{n}(\widehat{\Omega}_{ij} - \Omega_{ij}) = \sqrt{n}(g(\widehat{\boldsymbol{\alpha}}) - g(\boldsymbol{\alpha}^*)) \xrightarrow{d} \mathcal{N}(0, \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)^T \Gamma(\boldsymbol{\alpha}^*)^{-1} \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*))$ as $n \to \infty$. From the continuous mapping theorem, $\widehat{\upsilon}_{ij}$ tends in probability to the constant $\upsilon_{ij} = (\nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)^T \Gamma(\boldsymbol{\alpha}^*)^{-1} \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*))^{1/2} \neq 0$ (nonzero for positive-definite Fisher information matrix $\Gamma(\boldsymbol{\alpha}^*)$ by the assumption $\nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*) \neq 0$) as $\widehat{\boldsymbol{\alpha}} \to \boldsymbol{\alpha}^*$. Then from Slutsky's theorem (Corollary 2.3.2 in [124]), the ratio $\sqrt{n}(\widehat{\Omega}_{ij}/\widehat{\upsilon}_{ij} - \Omega_{ij}/\upsilon_{ij}) \xrightarrow{d} R$ as $n \to \infty$ where $R \sim \mathcal{N}\left(0, \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)^T \Gamma(\boldsymbol{\alpha}^*)^{-1} \nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)/\upsilon_{ij}^2\right) =$

$\mathcal{N}(0,1)$. Thus, the probability of false negatives given by

$$\begin{aligned}\mathbb{P}\left(\overline{\Omega}_{ij}=0;\Omega_{ij}\neq 0\right) &= \mathbb{P}\left(\widehat{\Omega}_{ij}<\frac{f(n)\widehat{\upsilon}_{ij}}{\sqrt{n}};\,\Omega_{ij}\neq 0\right)\\ &= \mathbb{P}\left(\sqrt{n}\left(\frac{\widehat{\Omega}_{ij}}{\widehat{\upsilon}_{ij}}-\frac{\Omega_{ij}}{\upsilon_{ij}}\right)<f(n)-\frac{\sqrt{n}\Omega_{ij}}{\upsilon_{ij}};\,\Omega_{ij}\neq 0\right),\end{aligned}$$

asymptotically converges to

$$\mathbb{P}\left(R<f(n)-\frac{\sqrt{n}\Omega_{ij}}{\upsilon_{ij}}\right)\leq e^{-\frac{1}{2}\left(f(n)-\frac{\sqrt{n}\Omega_{ij}}{\upsilon_{ij}}\right)^2}, \tag{C.11}$$

where the last inequality provides an explicit bound and follows from the Gaussian tail bound $\mathbb{P}(R<r)\leq e^{-r^2/2}$ for $r\leq -1$. For $f(n)/\sqrt{n}\to 0$, the term $f(n)-\frac{\sqrt{n}\Omega_{ij}}{\upsilon_{ij}}$ tends to negative infinity as $n\to\infty$, and therefore, the probability of a false negative in (C.11) converges to zero for any f that grows more slowly than $\sqrt{n}$. Conversely, if $f(n)=c\sqrt{n}$ for some $c\geq \Omega_{ij}/\upsilon_{ij}$, the left-hand side in (C.11) does not go to zero.

For conditionally independent variables, we have $\partial_i\partial_j \log S^{\sharp}_{\boldsymbol{\alpha}^*}\eta(\boldsymbol{x})=0$ for all $\boldsymbol{x}\in\mathbb{R}^d$; thus

$$g(\boldsymbol{\alpha}^*)=\mathbb{E}|\partial_i\partial_j \log S^{\sharp}_{\boldsymbol{\alpha}^*}\eta(\boldsymbol{X})|^2=0$$

and

$$\nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)=2\mathbb{E}[\partial_i\partial_j \log S^{\sharp}_{\boldsymbol{\alpha}^*}\eta(\boldsymbol{X})\nabla_{\boldsymbol{\alpha}}\partial_i\partial_j \log S^{\sharp}_{\boldsymbol{\alpha}^*}\eta(\boldsymbol{X})]=0$$

. Under the assumption that $\nabla^2_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}^*)\neq 0$, we can truncate the expansion in (C.10) to second-order terms and approximate $\widehat{\Omega}_{ij}$ asymptotically using a quadratic form (i.e., a degree two polynomial). By Proposition 2.1 in [61], the Wald statistic $n\widehat{\Omega}^2_{ij}/\widehat{\upsilon}^2_{ij}=\frac{ng(\widehat{\boldsymbol{\alpha}})^2}{\nabla_{\boldsymbol{\alpha}} g(\widehat{\boldsymbol{\alpha}})^T\Gamma(\widehat{\boldsymbol{\alpha}})^{-1}\nabla_{\boldsymbol{\alpha}} g(\widehat{\boldsymbol{\alpha}})}\xrightarrow{d} W$, where the random variable W satisfies, by Proposition 3.4 in [61], $\mathbb{P}(W\geq x)\leq \mathbb{P}(\frac{1}{4}X\geq x)$ for $X\sim\chi^2_p$ and all $x>0$. Here, χ^2_p denotes a chi-squared variable with $p\geq 1$ degrees of freedom where p corresponds to the dimension of $\boldsymbol{\alpha}$. Then, the probability of false positives, which can be written as

$$\mathbb{P}(\overline{\Omega}_{ij}\neq 0;\Omega_{ij}=0)=\mathbb{P}\left(\widehat{\Omega}_{ij}>\frac{f(n)\widehat{\upsilon}_{ij}}{\sqrt{n}};\Omega_{ij}=0\right)=\mathbb{P}\left(\frac{n\widehat{\Omega}^2_{ij}}{\widehat{\upsilon}^2_{ij}}>f(n)^2;\Omega_{ij}=0\right)$$

since $\widehat{\Omega}_{ij}$ and $\widehat{v}_{ij}$ are non-negative, is asymptotically given by

$$\mathbb{P}\left(W > f(n)^2\right) \leq \mathbb{P}\left(X - p > 4p\left(\frac{f(n)^2}{p} - \frac{1}{4}\right)\right) \leq e^{-f(n)^2/p+1/4}, \tag{C.12}$$

where the last inequality uses the χ^2_p squared tail bound $\mathbb{P}(X - p \geq 4px) \leq \mathbb{P}(X - p \geq 2\sqrt{px} + 2x) \leq e^{-x}$ for $x \geq 1$ (Lemma 1 in [118]). For $f(n) \to \infty$, the right-hand side in (C.12) and hence the left-hand side converge to zero as $n \to \infty$.

To complete the proof we use a union bound over all pairs (i, j) to bound the probability of failing to recover the true edge set

$$\mathbb{P}(\widehat{E}_n \neq E) \leq \sum_{(i,j)\in E} \mathbb{P}(\overline{\Omega}_{ij} = 0; \Omega_{ij} \neq 0) + \sum_{(i,j)\notin E} \mathbb{P}(\overline{\Omega}_{ij} \neq 0; \Omega_{ij} = 0). \tag{C.13}$$

Using the asymptotic results in (C.11) and (C.12) to bound each term in (C.13) we have $\mathbb{P}(\widehat{E}_n \neq E) \to 0$ as $n \to \infty$. □

Appendix D

Proofs for Chapter 5

Proof of Proposition 11. First, we express the density for the output of the composed map π_c in (5.23) in terms of the approximate conditional density from using a single map π_s as

$$\begin{aligned}
\pi_c(\boldsymbol{x}) &= \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^\sharp \pi_{\mathbf{Z}_2}(\boldsymbol{x}) \\
&= \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^\sharp \left(\frac{\pi_{\mathbf{Z}_2}(\boldsymbol{x})}{\eta_{\mathbf{Z}_2}(\boldsymbol{x})} \eta_{\mathbf{Z}_2}(\boldsymbol{x}) \right) \\
&= \left. \frac{\pi_{\mathbf{Z}_2}(\boldsymbol{z})}{\eta_{\mathbf{Z}_2}(\boldsymbol{z})} \right|_{\boldsymbol{z}=\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{x})} \eta_{\mathbf{Z}_2}(\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{x})) |\nabla_{\boldsymbol{x}} \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{x})| \\
&= \left. \frac{\pi_{\mathbf{Z}_2}(\boldsymbol{z})}{\eta_{\mathbf{Z}_2}(\boldsymbol{z})} \right|_{\boldsymbol{z}=\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \boldsymbol{x})} \pi_s(\boldsymbol{x}).
\end{aligned} \tag{D.1}$$

For any realization of the observation $\boldsymbol{y}^*$, the KL divergence from the approximate posterior density π_c in (D.1) to the true posterior $\pi_{\mathbf{X}|\boldsymbol{y}^*}$ is given by

$$\begin{aligned}
\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\boldsymbol{y}^*} || \pi_c) &= \mathbb{E}_{\pi_{\mathbf{X}|\boldsymbol{y}^*}} \left[\log \pi_{\mathbf{X}|\boldsymbol{y}^*}(\mathbf{X}) - \log \pi_c(\mathbf{X}) \right] \\
&= \mathbb{E}_{\pi_{\mathbf{X}|\boldsymbol{y}^*}} \left[\log \pi_{\mathbf{X}|\boldsymbol{y}^*}(\mathbf{X}) - \log \pi_s(\mathbf{X}) - \log \left. \frac{\pi_{\mathbf{Z}_2}(\boldsymbol{z})}{\eta_{\mathbf{Z}_2}(\boldsymbol{z})} \right|_{\boldsymbol{z}=\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \mathbf{X})} \right] \\
&= \mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\boldsymbol{y}^*} || \pi_S) - \mathbb{E}_{\pi_{\mathbf{X}|\boldsymbol{y}^*}} \left[\log \left. \frac{\pi_{\mathbf{Z}_2}(\boldsymbol{z})}{\eta_{\mathbf{Z}_2}(\boldsymbol{z})} \right|_{\boldsymbol{z}=\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \mathbf{X})} \right].
\end{aligned} \tag{D.2}$$

Using the change of variables $\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\boldsymbol{y}^*) = \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*, \cdot)^\sharp \pi_{\mathbf{Z}_2|\mathbf{Y}}(\cdot|\boldsymbol{y}^*)$, the second term in (D.2) is equivalent to

$$\mathbb{E}_{\pi_{\mathbf{X}|\boldsymbol{y}^*}}\left[\log \frac{\pi_{\mathbf{Z}_2}(\boldsymbol{z})}{\eta_{\mathbf{Z}_2}(\boldsymbol{z})}\bigg|_{\boldsymbol{z}=\widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\mathbf{X})}\right] = \mathbb{E}_{\pi_{\mathbf{Z}_2|\boldsymbol{y}^*}}\left[\log \frac{\pi_{\mathbf{Z}_2}(\mathbf{Z}_2)}{\eta_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right] \tag{D.3}$$

Taking an expectation on both sides of (D.2) over $\mathbf{Y}^* \sim \pi_{\mathbf{Y}}$ we have

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_c)] &= \mathbb{E}[\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_s)] - \mathbb{E}_{\pi_{\mathbf{Z}_2,\mathbf{Y}^*}}\left[\log \frac{\pi_{\mathbf{Z}_2}(\mathbf{Z}_2)}{\eta_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right] \\
&= \mathbb{E}[\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_s)] - \mathbb{E}_{\pi_{\mathbf{Z}_2}}\left[\log \frac{\pi_{\mathbf{Z}_2}(\mathbf{Z}_2)}{\eta_{\mathbf{Z}_2}(\mathbf{Z}_2)}\right] \\
&= \mathbb{E}[\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_s)] - \mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{Z}_2}||\eta_{\mathbf{Z}_2}) \\
&\leq \mathbb{E}[\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}^*)||\pi_s)],
\end{aligned} \tag{D.4}$$

where the last inequality follows from the positivity of the KL divergence. Furthermore, the inequality is strict if the marginal reference density $\pi_{\mathbf{Z}_2}$ is not equal to the standard Gaussian density $\eta_{\mathbf{Z}_2}$ so that $\mathcal{D}_{\mathrm{KL}}(\pi_{\mathbf{Z}_2}||\eta_{\mathbf{Z}_2}) \neq 0$. □

Proof of Proposition 12. From (5.25), the density for the output of the composed map is given by

$$\begin{aligned}
\pi_c(\boldsymbol{x}) &= \int \pi_{\mathbf{X}|\mathbf{Y}}(\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\boldsymbol{x})|\boldsymbol{y}) \left|\partial_{\boldsymbol{x}} \widehat{S}(\boldsymbol{y},\cdot)^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\boldsymbol{x})\right| \pi_{\mathbf{Y}}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \int \pi_{\mathbf{X}|\mathbf{Y}}(\widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)^{-1} \circ L \circ L^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\boldsymbol{x})|\boldsymbol{y}) \\
&\qquad \left|\partial_{\boldsymbol{z}} \widehat{S}^{\mathcal{X}}(\boldsymbol{y},\cdot)^{-1} \circ L(\boldsymbol{z})\right| \left|\partial_{\boldsymbol{x}} L^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{y}^*,\boldsymbol{x})\right| \pi_{\mathbf{Y}}(\boldsymbol{y}) d\boldsymbol{y}
\end{aligned} \tag{D.5}$$

Defining $S^{\mathcal{X}}(\boldsymbol{y},\boldsymbol{x}) = L^{-1} \circ \widehat{S}^{\mathcal{X}}(\boldsymbol{y},\boldsymbol{x})$, we have that $S^{\mathcal{X}}(\boldsymbol{y},\cdot)_\sharp \pi_{\mathbf{X}|\boldsymbol{y}} = \eta_{\mathbf{Z}_2}$ for all $\boldsymbol{y}$. Therefore, we can write

$$\begin{aligned}
\pi_c(\boldsymbol{x}) &= \int (S^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1})_\sharp S^{\mathcal{X}}(\boldsymbol{y},\cdot)_\sharp \pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) \pi_{\mathbf{Y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \\
&= \int (S^{\mathcal{X}}(\boldsymbol{y}^*,\cdot)^{-1})_\sharp \eta_{\mathbf{Z}_2}(\boldsymbol{x}) \pi_{\mathbf{Y}}(\boldsymbol{y}) d\boldsymbol{y} = \int \pi_{\mathbf{X}|\boldsymbol{y}^*}(\boldsymbol{x}) \pi_{\mathbf{Y}}(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} = \pi_{\mathbf{X}|\boldsymbol{y}^*}(\boldsymbol{x}).
\end{aligned} \tag{D.6}$$

□

Appendix E

Proofs for Chapter 7

Proof of Proposition 14. Let $\widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}$ be any density of the form $\widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) = f_1(\boldsymbol{y}_s, \boldsymbol{x}_r) f_2(\boldsymbol{x}_\perp, \boldsymbol{x}_r)$. Let $f_0(\boldsymbol{x}_r) = \int f_2(\boldsymbol{x}_\perp, \boldsymbol{x}_r) \mathrm{d}\boldsymbol{x}_\perp$ and

$$\begin{aligned}
\overline{f}_1(\boldsymbol{y}_s, \boldsymbol{x}_r) &= f_1(\boldsymbol{y}_s, \boldsymbol{x}_r) f_0(\boldsymbol{x}_r) \\
\overline{f}_2(\boldsymbol{x}_\perp, \boldsymbol{x}_r) &= f_2(\boldsymbol{x}_\perp, \boldsymbol{x}_r) / f_0(\boldsymbol{x}_r),
\end{aligned}$$

so that $\int \overline{f}_2(\boldsymbol{x}_\perp, \boldsymbol{x}_r) \mathrm{d}\boldsymbol{x}_\perp = 1$ for all $\boldsymbol{x}_r$ and $\int \overline{f}_1(\boldsymbol{y}_s, \boldsymbol{x}_r) \mathrm{d}\boldsymbol{x}_r = \int \widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1$ for all $\boldsymbol{y}_s$. Then $\boldsymbol{x}_r \mapsto \overline{f}_1(\boldsymbol{y}_s, \boldsymbol{x}_r)$ and $\boldsymbol{x}_r \mapsto \overline{f}_1(\boldsymbol{y}_s, \boldsymbol{x}_r)$ can be interpreted as conditional densities. From the definition of the KL divergence, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}} \left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}) || \widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})) \right] - \mathbb{E}_{\mathbf{Y}} \left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}) || \pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})) \right] \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Y}} [\log \pi^*_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{Y}) - \log \widetilde{\pi}_{\mathbf{X}|\mathbf{Y}}(\mathbf{X}|\mathbf{Y})] \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Y}} [\log \pi_{\mathbf{Y}_s|\mathbf{X}_r}(\mathbf{Y}_s|\mathbf{X}_r) \pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\mathbf{X}_\perp|\mathbf{X}_r) - \log \overline{f}_1(\mathbf{Y}_s, \mathbf{X}_r) \overline{f}_2(\mathbf{X}_\perp, \mathbf{X}_r)] \\
&= \mathbb{E}_{\mathbf{X},\mathbf{Y}} \left[\log \frac{\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\mathbf{Y}_s|\mathbf{X}_r)}{\overline{f}_1(\mathbf{Y}_s, \mathbf{X}_r)} - \log \frac{\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\mathbf{X}_\perp|\mathbf{X}_r)}{\overline{f}_2(\mathbf{X}_\perp, \mathbf{X}_r)} \right] \\
&= \mathbb{E}_{\mathbf{X}_r} \left[\mathbb{E}_{\mathbf{Y}_s|\mathbf{X}_r} \left[\log \frac{\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\mathbf{Y}_s|\mathbf{X}_r)}{\overline{f}_1(\mathbf{Y}_s, \mathbf{X}_r)} \right] + \mathbb{E}_{\mathbf{X}_\perp|\mathbf{X}_r} \left[\log \frac{\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\mathbf{X}_\perp|\mathbf{X}_r)}{\overline{f}_2(\mathbf{X}_\perp, \mathbf{X}_r)} \right] \right] \\
&= \mathbb{E}_{\mathbf{X}_r} \left[D_{\mathrm{KL}}(\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\cdot|\mathbf{X}_r) || \overline{f}_1(\cdot, \mathbf{X}_r)) + D_{\mathrm{KL}}(\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\cdot|\mathbf{X}_r) || \overline{f}_2(\cdot, \mathbf{X}_r)) \right] \qquad \text{(E.1)}
\end{aligned}$$

By the positivity of the two KL divergence terms in (E.1), we have the result in (7.6).

□

Proof of Proposition 16. Let $\pi_{\mathbf{X},\mathbf{Y}}$ be a joint density of $(\mathbf{X}, \mathbf{Y})$ with posterior density

$$\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})}{\pi_{\mathbf{Y}}(\boldsymbol{y})}.$$

Let the density for the optimal posterior approximation be

$$\pi^*_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)\pi_{\mathbf{X}_\perp|\mathbf{X}_r}(\boldsymbol{x}_\perp|\boldsymbol{x}_r)}{\pi^*_{\mathbf{Y}}(\boldsymbol{y}_s)},$$

where $\pi_{\mathbf{Y}_s|\mathbf{X}_r} = \int \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x}_\perp, \boldsymbol{x}_r)\pi(\boldsymbol{x}_\perp|\boldsymbol{x}_r)\mathrm{d}\boldsymbol{x}_\perp\mathrm{d}\boldsymbol{y}_\perp = \pi(\boldsymbol{y}_s|\boldsymbol{x}_r)$ is the approximate likelihood function. For this likelihood, the approximate data marginal satisfies

$$\begin{aligned}\pi^*_{\mathbf{Y}}(\boldsymbol{y}_s) &= \int \pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)\pi_{\mathbf{X}_r}(\boldsymbol{x}_r)\mathrm{d}\boldsymbol{x}_r \\ &= \int\int \pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{y}_\perp\mathrm{d}\boldsymbol{x} = \int \pi_{\mathbf{Y}_s|\mathbf{X}}(\boldsymbol{y}_s|\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \pi_{\mathbf{Y}}(\boldsymbol{y}_s).\end{aligned}$$

The KL divergence from the optimal posterior approximation to the true posterior in expectation over the data is then given by

$$\begin{aligned}&\mathbb{E}_{\mathbf{Y}}\left[D_{\mathrm{KL}}(\pi_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y})||\pi^*_{\mathbf{X}|\mathbf{Y}}(\cdot|\mathbf{Y}))\right] \\ &= \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{\pi_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})}{\pi^*_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})\pi_{\mathbf{X}}(\boldsymbol{x})/\pi_{\mathbf{Y}}(\boldsymbol{y})}{\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)\pi_{\mathbf{X}}(\boldsymbol{x})/\pi_{\mathbf{Y}_s}(\boldsymbol{y}_s)}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\mathbf{Y}}(\boldsymbol{y})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} - \int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)}{\pi_{\mathbf{Y}_s}(\boldsymbol{y}_s)}\mathrm{d}\boldsymbol{y} \\ &= \underbrace{\int \pi_{\mathbf{X},\mathbf{Y}}(\boldsymbol{x},\boldsymbol{y}) \log \frac{\pi_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\mathbf{Y}}(\boldsymbol{y})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}}_{I(\mathbf{X};\mathbf{Y})} - \underbrace{\int \pi_{\mathbf{X}_r,\mathbf{Y}_s}(\boldsymbol{x}_r,\boldsymbol{y}_s) \log \frac{\pi_{\mathbf{Y}_s|\mathbf{X}_r}(\boldsymbol{y}_s|\boldsymbol{x}_r)}{\pi_{\mathbf{Y}_s}(\boldsymbol{y}_s)}\mathrm{d}\boldsymbol{x}_r\mathrm{d}\boldsymbol{y}_s}_{I(\mathbf{X}_r;\mathbf{Y}_s)}.\end{aligned} \tag{E.2}$$

Lastly, from the chain rule for mutual information we have

$$I(\mathbf{X};\mathbf{Y}) - I(\mathbf{X}_r,\mathbf{Y}_s) = I(\mathbf{X}_r;\mathbf{Y}|\mathbf{X}_\perp) - I(\mathbf{X},\mathbf{Y}_s|\mathbf{Y}_\perp) - I(\mathbf{X}_r;\mathbf{Y}_s|\mathbf{X}_\perp,\mathbf{Y}_\perp).$$

□

Proof of Proposition 18. For the linear model $\mathbf{Y} = G\mathbf{X} + \boldsymbol{\mathcal{E}}$ with $\mathbf{X} \perp\!\!\!\perp \boldsymbol{\mathcal{E}}$, the covariance of $\mathbf{Y}$ is $\mathrm{Cov}(\mathbf{Y}) = G\Gamma_{\mathrm{pr}}G^T + \Gamma_{\mathrm{obs}}$, and the cross-covariance is $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \Gamma_{\mathrm{pr}}G^T$. Hence, the eigenvalue problems in (7.41) and (7.42) can be written as

$$\Gamma_{\mathrm{pr}}G^T(G\Gamma_{\mathrm{pr}}G^T + \Gamma_{\mathrm{obs}})^{-1}G\Gamma_{\mathrm{pr}}u_i^{\mathrm{CCA}} = \rho_i\Gamma_{\mathrm{pr}}u_i^{\mathrm{CCA}} \tag{E.3}$$

$$G\Gamma_{\mathrm{pr}}G^T v_i^{\mathrm{CCA}} = \rho_i(G\Gamma_{\mathrm{pr}}G^T + \Gamma_{\mathrm{obs}})v_i^{\mathrm{CCA}}. \tag{E.4}$$

Using the Sherman-Morrison-Woodbury formula, we have the matrix identity

$$G^T(G\Gamma_{\mathrm{pr}}G^T + \Gamma_{\mathrm{obs}})^{-1}G = \Gamma_{\mathrm{pr}}^{-1}(\Gamma_{\mathrm{pr}}^{-1} + G^T\Gamma_{\mathrm{obs}}^{-1}G)^{-1}G^T\Gamma_{\mathrm{obs}}^{-1}G.$$

Applying this identity to the left hand side of (E.3), the eigenvalue problems in CCA are also given by

$$G^T\Gamma_{\mathrm{obs}}^{-1}G\Gamma_{\mathrm{pr}}u_i^{\mathrm{CCA}} = \frac{\rho_i}{1-\rho_i}u_i^{\mathrm{CCA}},$$

$$G\Gamma_{\mathrm{pr}}G^T v_i^{\mathrm{CCA}} = \frac{\rho_i}{1-\rho_i}\Gamma_{\mathrm{obs}}v_i^{\mathrm{CCA}}.$$

The vectors $u_i^{\mathrm{CCA}}, v_i^{\mathrm{CCA}}$ are also eigenvectors of (7.31) and (7.32). Furthermore, given that $\rho_i/(1-\rho_i)$ is a monotonic function of $\rho_i \in [-1, 1]$, the eigenvectors are ordered in the same way as the solutions to (E.3) and (E.4). □

Appendix F

Additional calculations for Chapter 7

Gaussian subspace log-Sobolev constant (7.14). We begin by computing the eigenvalues of the joint covariance matrix $\mathrm{Cov}(\mathbf{X}, \mathbf{Y})$. The eigenvalues are given by the $d+m$ roots s of the equation $\det(\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) - s\mathrm{I}_{d+m}) = 0$. Without loss of generality, we let $d \geq m$.

From the matrix determinant lemma and the SVD of the forward model $G^T = U\Sigma V^T$ where $U \in \mathbb{R}^{d\times d}$, $V \in \mathbb{R}^{m\times m}$ are unitary matrices and $\Sigma \in \mathbb{R}^{d\times m}$ is a diagonal matrix containing zeros below row m, we have

$$\begin{aligned}
\det(\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) - s\mathrm{I}_{d+m}) &= \det((1-s)\mathrm{I}_d)\det(GG^T + \mathrm{I}_m - s\mathrm{I}_m - G(\mathrm{I}_d - s\mathrm{I}_d)^{-1}G^T). \\
&= \det((1-s)\mathrm{I}_d)\det(U(\Sigma^2 + (1-s)\mathrm{I}_m - \Sigma(1-s)^{-1}\Sigma)U^T) \\
&= \prod_{i=1}^{d}(1-s)\prod_{j=1}^{m}(\sigma_j^2(1-(1-s)^{-1}) + (1-s)). \\
&= \prod_{i=m+1}^{d}(1-s)\prod_{j=1}^{m}((1-s)^2 - \sigma_j^2 s).
\end{aligned}$$

Then, there are $d-m$ roots $s = 1$ and $2m$ roots

$$s = \frac{1}{2}\left(2 + \sigma_i \pm \sigma_i\sqrt{2+\sigma_i^2}\right), \tag{F.1}$$

from solving the quadratic equations $s^2 + (-2-\sigma_i)s + 1 = 0$ for $j = 1, \ldots, m$. The largest eigenvalues are the roots with the positive sign in (F.1). Given that $\sigma_i > 0$,

the roots are all greater than 1. Given that (F.1) with a positive sign is a monotonic functions of σ_i, it follows that the maximum eigenvalue is given by (7.14).

Expected KL divergence for a linear-Gaussian model (7.34). The difference of mutual information terms for Gaussian variables is given by

$$I(\mathbf{X}, \mathbf{Y}) - I(\mathbf{X}_r, \mathbf{Y}_s) = \frac{1}{2} \log \frac{|\Gamma_{\mathbf{X}}|}{|\Gamma_{\mathbf{X}|\mathbf{Y}}|} - \frac{1}{2} \log \frac{|\Gamma_{\mathbf{X}_r}|}{|\Gamma_{\mathbf{X}_r|\mathbf{Y}_s}|}, \tag{F.2}$$

where Γ represents a covariance. After the whitening transformations, we have the linear model $\mathbf{Y} = G\mathbf{X} + \boldsymbol{\mathcal{E}}$ with $G \leftarrow \Gamma_{\text{obs}}^{-1/2} G \Gamma_{\text{pr}}^{1/2}$, $\text{Cov}(\mathbf{X}) = \text{I}_d$ and $\text{Cov}(\boldsymbol{\mathcal{E}}) = \text{I}_m$. Then, the (conditional) covariances in (F.2) are given by

$$\begin{aligned}
\Gamma_{\mathbf{X}} &= \text{I}_d \\
\Gamma_{\mathbf{X}_r} &= U_r^T \text{I}_d U_r \\
\Gamma_{\mathbf{X}|\mathbf{Y}} &= \text{I}_d - G^T (GG^T + \text{I}_m)^{-1} G \\
\Gamma_{\mathbf{X}_r|\mathbf{Y}_s} &= U_r^T \text{I}_d U_r - U_r^T G^T V_s (V_s^T GG^T V_s + V_s^T V_s)^{-1} V_s^T G U_r.
\end{aligned}$$

Given that eigenvectors of $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$ are given by the left and right singular vectors U and V for G^T, respectively, we have $U_r = [u_1, \ldots, u_r]$ and $V_s = [v_1, \ldots, v_s]$. Then, the conditional covariances can be simplified as

$$\begin{aligned}
\Gamma_{\mathbf{X}|\mathbf{Y}} &= \text{I}_d - U\Lambda^2(\Lambda^2 + \text{I}_p)^{-1} U^T \\
\Gamma_{\mathbf{X}_r|\mathbf{Y}_s} &= \text{I}_r - \text{I}_{r,t} \Lambda_t^2 (\Lambda_t^2 + \text{I}_t)^{-1} \text{I}_{r,t}^T,
\end{aligned}$$

where $t = \min\{r, s\}$ and $\text{I}_{r,t}$ represents the first t columns of an identity matrix $\text{I}_r \in \mathbb{R}^{r \times r}$. By computing the determinants of the (conditional) covariances, we have

$$\begin{aligned}
I(\mathbf{X}, \mathbf{Y}) - I(\mathbf{X}_r, \mathbf{Y}_s) &= -\frac{1}{2} \log \left| \text{I}_p - \Lambda^2(\Lambda^2 + \text{I}_p)^{-1} \right| + \frac{1}{2} \log \left| \text{I}_t - \Lambda_t^2 (\Lambda_t^2 + \text{I}_t)^{-1} \right| \\
&= \frac{1}{2} \sum_{i>t}^{p} \log(1 + \lambda_i^2).
\end{aligned}$$

Bibliography

[1] Sigurd I Aanonsen, Geir Nævdal, Dean S Oliver, Albert C Reynolds, Brice Vallès, et al. "The ensemble Kalman filter in reservoir engineering–a review". In: *SPE Journal* 14.03 (2009), pp. 393–412.

[2] Luca Ambrogioni, Umut Güçlü, Marcel AJ van Gerven, and Eric Maris. "The kernel mixture network: A nonparametric method for conditional density estimation of continuous random variables". In: *arXiv:1705.07111* (2017).

[3] Brandon Amos, Lei Xu, and J Zico Kolter. "Input convex neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 146–155.

[4] Jeffrey L Anderson. "A Marginal Adjustment Rank Histogram Filter for Non-Gaussian Ensemble Data Assimilation". In: *Monthly Weather Review* 148.8 (2020), pp. 3361–3378.

[5] Jeffrey L Anderson. "A non-Gaussian ensemble filter update for data assimilation". In: *Monthly Weather Review* 138.11 (2010), pp. 4186–4198.

[6] Jeffrey L Anderson. "An ensemble adjustment Kalman filter for data assimilation". In: *Monthly Weather Review* 129.12 (2001), pp. 2884–2903.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. "Deep canonical correlation analysis". In: *International Conference on Machine Learning*. PMLR. 2013, pp. 1247–1255.

[8] Christophe Andrieu and Gareth O Roberts. "The pseudo-marginal approach for efficient Monte Carlo computations". In: *The Annals of Statistics* 37.2 (2009), pp. 697–725.

[9] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 214–223.

[10] Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data assimilation: methods, algorithms, and applications*. Vol. 11. SIAM, 2016.

[11] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*. Vol. 57. Springer Science & Business Media, 2007.

[12] Dominique Bakry and Michel Émery. "Diffusions hypercontractives". In: *Séminaire de Probabilités XIX 1983/84*. Springer, 1985, pp. 177–206.

[13] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data". In: *Journal of Machine Learning Research* 9.Mar (2008), pp. 485–516.

[14] Ricardo Baptista, Lianghao Cao, Joshua Chen, Omar Ghattas, Fengyi Li, Youssef Marzouk, and Tinsley Oden. "Statistical learning of models from diblock co-polymer images: Likelihood-free inference and information gain estimation via measure transport". In: *In preparation* (2022).

[15] Ricardo Baptista, Youssef Marzouk, Rebecca E Morrison, and Olivier Zahm. "Learning non-Gaussian graphical models via Hessian scores and triangular transport". In: *arXiv:2101.03093* (2021).

[16] Ricardo Baptista, Olivier Zahm, and Youssef Marzouk. "An adaptive transport framework for joint and conditional density estimation". In: *arXiv:2009.10303* (2020).

[17] Mark A Beaumont, Wenyang Zhang, and David J Balding. "Approximate Bayesian computation in population genetics". In: *Genetics* 162.4 (2002), pp. 2025–2035.

[18] Laurent Bertino, Geir Evensen, and Hans Wackernagel. "Sequential data assimilation techniques in oceanography". In: *International Statistical Review* 71.2 (2003), pp. 223–241.

[19] Dimitri P Bertsekas. "Nonlinear programming". In: *Journal of the Operational Research Society* 48.3 (1997), pp. 334–334.

[20] Peter J Bickel and Elizaveta Levina. "Regularized estimation of large covariance matrices". In: *The Annals of Statistics* 36.1 (2008), pp. 199–227.

[21] Chris Biemann. "Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems". In: *Proceedings of the first workshop on graph based methods for natural language processing*. Association for Computational Linguistics. 2006, pp. 73–80.

[22] Daniele Bigoni, Youssef Marzouk, Clémentine Prieur, and Olivier Zahm. "Nonlinear dimension reduction for surrogate modeling using gradient information". In: *arXiv:2102.10351* (2021).

[23] Christopher M Bishop. *Mixture density networks*. Tech. rep. Neural Computing Research Group report: NCRG/94/004. Aston University, 1994.

[24] Craig H Bishop, Brian J Etherton, and Sharanya J Majumdar. "Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects". In: *Monthly Weather Review* 129.3 (2001), pp. 420–436.

[25] Gary S Bloom and Solomon W Golomb. "Applications of numbered undirected graphs". In: *Proceedings of the IEEE* 65.4 (1977), pp. 562–570.

[26] Michael G.B. Blum. "Regression Approaches for ABC". In: *Handbook of Approximate Bayesian Computation*. Ed. by Scott A. Sisson, Yanan Fan, and Mark A. Beaumont. CRC Press, 2018. Chap. 3, pp. 71–86.

[27] Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. "Triangular transformations of measures". In: *Sbornik: Mathematics* 196.3 (2005), p. 309.

[28] Michael Brennan, Daniele Bigoni, Olivier Zahm, Alessio Spantini, and Youssef Marzouk. "Greedy inference with structure-exploiting lazy maps". In: vol. 33. 2020, pp. 8330–8342.

[29] Guy Bresler. "Efficiently learning Ising models on arbitrary graphs". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 2015, pp. 771–782.

[30] Tony Cai and Weidong Liu. "Adaptive thresholding for sparse covariance matrix estimation". In: *Journal of the American Statistical Association* 106.494 (2011), pp. 672–684.

[31] Lianghao Cao, Omar Ghattas, and J Tinsley Oden. "A Globally Convergent Modified Newton Method for the Direct Minimization of the Ohta–Kawasaki Energy with Application to the Directed Self-Assembly of Diblock Copolymers". In: *SIAM Journal on Scientific Computing* 44.1 (2022), B51–B79.

[32] Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. "Vector quantile regression: an optimal transport approach". In: *The Annals of Statistics* 44.3 (2016), pp. 1165–1192.

[33] Guillaume Carlier, Alfred Galichon, and Filippo 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.

[34] A. Carrassi, A. Trevisan, L. Descamps, O. Talagrand, and F. Uboldi. "Controlling instabilities along a 3DVar analysis cycle by assimilating in the unstable subspace: a comparison with the EnKF". In: *Nonlinear Processes in Geophysics* 15.4 (2008), pp. 503–521.

[35] Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. "Data assimilation in the geosciences: An overview of methods, issues, and perspectives". In: *Wiley Interdisciplinary Reviews: Climate Change* 9.5 (2018), e535.

[36] George Casella and Roger L Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002.

[37] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. "Latent variable graphical model selection via convex optimization". In: *Annals of Statistics* 40.4 (Aug. 2012), pp. 1935–1967.

[38] Seok-Ho Chang, Pamela C Cosman, and Laurence B Milstein. "Chernoff-type bounds for the Gaussian error function". In: *IEEE Transactions on Communications* 59.11 (2011), pp. 2939–2944.

[39] Abdellah Chkifa, Albert Cohen, and Christoph Schwab. "Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs". In: *Journal de Mathématiques Pures et Appliquées* 103.2 (2015), pp. 400–428.

[40] Rustum Choksi and Xiaofeng Ren. "On the derivation of a density functional theory for microphase separation of diblock copolymers". In: *Journal of Statistical Physics* 113.1 (2003), pp. 151–176.

[41] Alexandre Chorin, Matthias Morzfeld, and Xuemin Tu. "Implicit particle filters for data assimilation". In: *Communications in Applied Mathematics and Computational Science* 5.2 (2010), pp. 221–240.

[42] Albert Cohen. *Numerical analysis of wavelet methods*. Elsevier, 2003.

[43] Albert Cohen and Giovanni Migliorati. "Multivariate approximation in downward closed polynomial spaces". In: *Contemporary Computational Mathematics-A celebration of the 80th birthday of Ian Sloan*. Springer, 2018, pp. 233–282.

[44] David L Colton, Rainer Kress, and Rainer Kress. *Inverse acoustic and electromagnetic scattering theory*. Vol. 93. Springer, 1998.

[45] Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. "MCMC methods for functions: modifying old algorithms to make them faster". In: *Statistical Science* 28.3 (2013), pp. 424–446.

[46] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012. ISBN: 9781118585771.

[47] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. "The frontier of simulation-based inference". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30055–30062. DOI: 10.1073/pnas.1912789117.

[48] Tiangang Cui, Sergey Dolgov, and Olivier Zahm. "Conditional Deep Inverse Rosenblatt Transports". In: *arXiv:2106.04170* (2021).

[49] Tiangang Cui, Kody JH Law, and Youssef M Marzouk. "Dimension-independent likelihood-informed MCMC". In: *Journal of Computational Physics* 304 (2016), pp. 109–137.

[50] Tiangang Cui, James Martin, Youssef M Marzouk, Antti Solonen, and Alessio Spantini. "Likelihood-informed dimension reduction for nonlinear inverse problems". In: *Inverse Problems* 30.11 (2014), p. 114015.

[51] Tiangang Cui and Xin T Tong. "A unified performance analysis of likelihood-informed subspace methods". In: *arXiv:2101.02417* (2021).

[52] Tiangang Cui and Olivier Zahm. "Data-free likelihood-informed dimension reduction of Bayesian inverse problems". In: *Inverse Problems* 37.4 (2021), p. 045009.

[53] Niccolo Dalmasso, Rafael Izbicki, and Ann Lee. "Confidence sets and hypothesis testing in a likelihood-free inference setting". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2323–2334.

[54] Niccolò Dalmasso, Ann Lee, Rafael Izbicki, Taylor Pospisil, Ilmun Kim, and Chieh-An Lin. "Validation of approximate likelihood and emulator models for computationally intensive simulations". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3349–3361.

[55] Patrick Danaher, Pei Wang, and Daniela M Witten. "The joint graphical lasso for inverse covariance estimation across multiple classes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.2 (2014), pp. 373–397.

[56] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. "Density Estimation using Real NVP". In: *International Conference on Learning Representations*. 2017.

[57] Arnaud Doucet, Nando de Freitas, and Neil Gordon. *Sequential Monte Carlo methods in practice*. Springer-Verlag, 2001.

[58] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering". In: *Statistics and computing* 10.3 (2000), pp. 197–208.

[59] Mathias Drton and Marloes H Maathuis. "Structure learning in graphical modeling". In: *Annual Review of Statistics and Its Application* 4 (2017), pp. 365–393.

[60] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*. Vol. 39. Springer Science & Business Media, 2008.

[61] Mathias Drton, Han Xiao, et al. "Wald tests of singular hypotheses". In: *Bernoulli* 22.1 (2016), pp. 38–59.

[62] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. "Neural spline flows". In: *Advances in Neural Information Processing Systems*. 2019, pp. 7509–7520.

[63] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[64] Sacha Epskamp, Lourens J Waldorp, René Mõttus, and Denny Borsboom. "The Gaussian graphical model in cross-sectional and time-series data". In: *Multivariate Behavioral Research* 53.4 (2018), pp. 453–480.

[65] Geir Evensen. "The ensemble Kalman filter: Theoretical formulation and practical implementation". In: *Ocean dynamics* 53.4 (2003), pp. 343–367.

[66] Paul Fearnhead and Dennis Prangle. "Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3 (2012), pp. 419–474.

[67] Chi Feng and Youssef M Marzouk. "A layered multiple importance sampling scheme for focused optimal Bayesian experimental design". In: *arXiv:1903.11187* (2019).

[68] Andreas Fischer, Ching Y Suen, Volkmar Frinken, Kaspar Riesen, and Horst Bunke. "A fast matching algorithm for graph-based handwriting recognition". In: *International Workshop on Graph-Based Representations in Pattern Recognition*. Springer. 2013, pp. 194–203.

[69] Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. "Variational Bayesian Optimal Experimental Design". In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[70] Tomislav Fotak, Miroslav Bača, and Petra Koruga. "Handwritten signature identification using basic concepts of graph theory". In: *WSEAS Transactions on Signal Processing* 7 (2011), pp. 117–129.

[71] Marco Frei and Hans R Künsch. "Mixture ensemble Kalman filters". In: *Computational Statistics & Data Analysis* 58 (2013), pp. 127–138.

[72] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[73] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[74] Andrew Gelman and Xiao-Li Meng. "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling". In: *Statistical Science* (1998), pp. 163–185.

[75] Valeriy V Ginzburg, Jeffrey D Weinhold, and Peter Trefonas. "Computational modeling of block-copolymer directed self-assembly". In: *Journal of Polymer Science Part B: Polymer Physics* 53.2 (2015), pp. 90–95.

[76] Loïc Giraldi, Olivier P Le Maître, Ibrahim Hoteit, and Omar M Knio. "Optimal projection of observations in a Bayesian setting". In: *Computational Statistics & Data Analysis* 124 (2018), pp. 252–276.

[77] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69.2 (2007), pp. 243–268.

[78] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014.

[79] Neil J Gordon, David J Salmond, and Adrian FM Smith. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE proceedings F (radar and signal processing)*. Vol. 140. IET. 1993, pp. 107–113.

[80] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. "A kernel statistical test of independence". In: *Nips*. Vol. 20. Citeseer. 2007, pp. 585–592.

[81] Alice Guionnet and B Zegarlinksi. "Lectures on logarithmic Sobolev inequalities". In: *Séminaire de probabilités XXXVI*. Springer, 2003, pp. 1–134.

[82] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[83] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. "Joint estimation of multiple graphical models". In: *Biometrika* 98.1 (2011), pp. 1–15.

[84] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Vol. 1. Springer, 2002.

[85] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. "Canonical correlation analysis: An overview with application to learning methods". In: *Neural Computation* 16.12 (2004), pp. 2639–2664.

[86] Eugenio Hernández and Guido Weiss. *A first course on wavelets*. CRC Press, 1996.

[87] RA Holley and DW Stroock. "L_2 theory for the stochastic Ising model". In: *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 35.2 (1976), pp. 87–101.

[88] Richard Holley and Daniel Stroock. "Logarithmic Sobolev inequalities and stochastic Ising models". In: *Journal of Statistical Physics* 46.5 (1987), pp. 1159–1194.

[89] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.

[90] Harold Hotelling. "Analysis of a complex of statistical variables into principal components." In: *Journal of Educational Psychology* 24.6 (1933), p. 417.

[91] Harold Hotelling. "Relations between two sets of variates". In: *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[92] Peter L Houtekamer and Herschel L Mitchell. "A sequential ensemble Kalman filter for atmospheric data assimilation". In: *Monthly Weather Review* 129.1 (2001), pp. 123–137.

[93] Daniel Hsu, Sham Kakade, and Tong Zhang. "A tail inequality for quadratic forms of subgaussian random vectors". In: *Electronic Communications in Probability* 17 (2012), pp. 1–6.

[94] Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. "Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization". In: *International Conference on Learning Representations*. 2021.

[95] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. "Neural autoregressive flows". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2078–2087.

[96] Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. "Covariance matrix selection and estimation via penalised normal likelihood". In: *Biometrika* 93.1 (2006), pp. 85–98.

[97] Mitsuo Izuki. "The characterizations of weighted Sobolev spaces by wavelets and scaling functions". In: *Taiwanese Journal of Mathematics* 13.2A (2009), pp. 467–492.

[98] Jayanth Jagalur-Mohan and Youssef Marzouk. "Batch greedy maximization of non-submodular functions: Guarantees and applications to experimental design". In: *Journal of Machine Learning Research* 22.252 (2021), pp. 1–62.

[99] Priyank Jaini, Kira A Selby, and Yaoliang Yu. "Sum-of-squares polynomial flow". In: *International Conference on Machine Learning*. 2019, pp. 3009–3018.

[100] Ali Jalali, Christopher C Johnson, and Pradeep K Ravikumar. "On learning discrete graphical models using greedy methods". In: *Advances in Neural Information Processing Systems*. 2011, pp. 1935–1943.

[101] Andrew H Jazwinski. *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[102] Shengxiang Ji, Lei Wan, Chi-Chun Liu, and Paul F Nealey. "Directed self-assembly of block copolymers on chemical patterns: A platform for nanofabrication". In: *Progress in Polymer Science* 54 (2016), pp. 76–127.

[103] Ian T Jolliffe. *Principal component analysis*. Springer, 2002.

[104] Jari Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*. Vol. 160. Springer Science & Business Media, 2006.

[105] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2003.

[106] Xiaoning Kang and Xinwei Deng. "An Improved Modified Cholesky Decomposition Approach for Precision Matrix Estimation". In: *Journal of Statistical Computation and Simulation* 90.3 (2020), pp. 443–464.

[107] Matthias Katzfuss and Florian Schäfer. "Scalable Bayesian transport maps for high-dimensional non-Gaussian spatial fields". In: *arXiv:2108.04211* (2021).

[108] Hans Kellerer, Ulrich Pferschy, and David Pisinger. "Introduction to NP-Completeness of knapsack problems". In: *Knapsack problems*. Springer, 2004, pp. 483–493.

[109] Durk P Kingma and Prafulla Dhariwal. "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in Neural Information Processing Systems*. 2018, pp. 10215–10224.

[110] Herbert Knothe. "Contributions to the theory of convex bodies." In: *The Michigan Mathematical Journal* 1957.1028990175 (1957).

[111] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. "Normalizing flows: An introduction and review of current methods". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[112] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[113] Nikola Kovachki, Ricardo Baptista, Bamdad Hosseini, and Youssef Marzouk. "Conditional Sampling With Monotone GANs". In: *arXiv:2006.06755* (2020).

[114] Alois Kufner and Bohumír Opic. "How to define reasonably weighted Sobolev spaces". In: *Commentationes Mathematicae Universitatis Carolinae* 25.3 (1984), pp. 537–554.

[115] Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. "A large-scale study on regularization and normalization in GANs". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3581–3590.

[116] Remi R Lam, Olivier Zahm, Youssef M Marzouk, and Karen E Willcox. "Multifidelity dimension reduction via active subspaces". In: *SIAM Journal on Scientific Computing* 42.2 (2020), A929–A956.

[117] Kenneth Lange. *MM Optimization Algorithms*. SIAM, 2016.

[118] Beatrice Laurent and Pascal Massart. "Adaptive estimation of a quadratic functional by model selection". In: *Annals of Statistics* (2000), pp. 1302–1338.

[119] Steffen L Lauritzen. *Graphical models*. Vol. 17. Clarendon Press, 1996.

[120] Kody Law, Andrew Stuart, and Kostas Zygalakis. "Data assimilation". In: *Cham, Switzerland: Springer* (2015).

[121] Kody JH Law, Hamidou Tembine, and Raul Tempone. "Deterministic mean-field ensemble Kalman filtering". In: *SIAM Journal on Scientific Computing* 38.3 (2016), A1251–A1279.

[122] François Le Gland, Valerie Monbet, and Vu-Duc Tran. *Large sample asymptotics for the ensemble Kalman filter*. Research Report RR-7014. INRIA, 2009, p. 25.

[123] Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff Eldredge. "A low-rank nonlinear ensemble filter for vortex models of aerodynamic flows". In: *AIAA Scitech 2021 Forum*. 2021, p. 1937.

[124] Erich Leo Lehmann. *Elements of large-sample theory*. Springer Science & Business Media, 2004.

[125] Jing Lei and Peter Bickel. "A moment matching ensemble filter for nonlinear non-Gaussian data assimilation". In: *Monthly Weather Review* 139.12 (2011), pp. 3964–3973.

[126] Elizaveta Levina, Adam Rothman, Ji Zhu, et al. "Sparse estimation of large covariance matrices via a nested lasso penalty". In: *The Annals of Applied Statistics* 2.1 (2008), pp. 245–263.

[127] Mario Lezcano Casado. "Trivializations for gradient-based optimization on manifolds". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 9157–9168.

[128] Lina Lin, Mathias Drton, and Ali Shojaie. "Estimation of high-dimensional graphical models using regularized score matching". In: *Electronic Journal of Statistics* 10.1 (2016), p. 806.

[129] Dennis V Lindley. "On a measure of the information provided by an experiment". In: *The Annals of Mathematical Statistics* (1956), pp. 986–1005.

[130] Han Liu, Fang Han, and Cun-hui Zhang. "Transelliptical graphical models". In: *Advances in Neural Information Processing Systems*. 2012, pp. 800–808.

[131] Han Liu, John Lafferty, and Larry Wasserman. "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs". In: *Journal of Machine Learning Research* 10 (2009), pp. 2295–2328.

[132] Po-Ling Loh and Martin J Wainwright. "Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses." In: *NIPS*. 2012, pp. 2096–2104.

[133] Edward N Lorenz. "Deterministic nonperiodic flow". In: *Journal of the atmospheric sciences* 20.2 (1963), pp. 130–141.

[134] Edward N Lorenz. "Predictability: A problem partly solved". In: *Proc. Seminar on predictability*. Vol. 1. 1996.

[135] Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. "Likelihood-free inference with emulator networks". In: *Symposium on Advances in Approximate Bayesian Inference*. PMLR. 2019, pp. 32–53.

[136] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. "Benchmarking simulation-based inference". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 343–351.

[137] Andrew J Majda and John Harlim. *Filtering complex turbulent systems*. Cambridge University Press, 2012.

[138] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[139] Jan Mandel, Lynn S Bennethum, Jonathan D Beezley, Janice L Coen, Craig C Douglas, Minjeong Kim, and Anthony Vodacek. "A wildland fire model with data assimilation". In: *Mathematics and Computers in Simulation* 79.3 (2008), pp. 584–606.

[140] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. "Markov chain Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15324–15328.

[141] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. "Sampling via Measure Transport: An Introduction". In: *Handbook of Uncertainty Quantification*. Springer International Publishing, 2016, pp. 1–41. ISBN: 978-3-319-11259-6. DOI: 10.1007/978-3-319-11259-6_23-1.

[142] Robert J McCann. "Existence and uniqueness of monotone measure-preserving maps". In: *Duke Mathematical Journal* 80.2 (1995), pp. 309–323.

[143] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *The Annals of Statistics* (2006), pp. 1436–1462.

[144] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. "Which training methods for GANs do actually converge?" In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3481–3490.

[145] Tomer Michaeli, Weiran Wang, and Karen Livescu. "Nonparametric canonical correlation analysis". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 1967–1976.

[146] Giovanni Migliorati. "Adaptive approximation by optimal weighted least-squares methods". In: *SIAM Journal on Numerical Analysis* 57.5 (2019), pp. 2217–2245.

[147] Giovanni Migliorati. "Adaptive polynomial approximation by means of random discrete least squares". In: *Numerical Mathematics and Advanced Applications-ENUMATH 2013*. Springer, 2015, pp. 547–554.

[148] Rada Mihalcea and Dragomir Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

[149] Takemasa Miyoshi, Keiichi Kondo, and Toshiyuki Imamura. "The 10,240-member ensemble Kalman filtering with an intermediate AGCM". In: *Geophysical Research Letters* 41.14 (2014), pp. 5264–5271.

[150] Rebecca Morrison, Ricardo Baptista, and Youssef Marzouk. "Beyond normality: Learning sparse probabilistic graphical models in the non-Gaussian setting". In: *Advances in Neural Information Processing Systems 30*. 2017, pp. 2359–2369.

[151] Rebecca E Morrison, Ricardo Baptista, and Estelle L Basor. "Diagonal Nonlinear Transformations Preserve Structure in Covariance and Precision Matrices". In: *Journal of Multivariate Analysis* (2022, In Press).

[152] Matthias Morzfeld and Daniel Hodyss. "Gaussian approximations in filters and smoothers for data assimilation". In: *Tellus A: Dynamic Meteorology and Oceanography* 71.1 (2019), pp. 1–27.

[153] Benjamin Muckenhoupt. "Hardy's inequality with weights". In: *Studia Mathematica* 44.1 (1972), pp. 31–38.

[154] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[155] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[156] Gary W Oehlert. "A note on the delta method". In: *The American Statistician* 46.1 (1992), pp. 27–29.

[157] Takao Ohta and Kyozi Kawasaki. "Equilibrium morphology of block copolymer melts". In: *Macromolecules* 19.10 (1986), pp. 2621–2632.

[158] Dean S Oliver, Nanqun He, and Albert C Reynolds. "Conditioning permeability fields to pressure data". In: *ECMOR V-5th European conference on the mathematics of oil recovery*. 1996.

[159] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. "Parallel WaveNet: Fast high-fidelity speech synthesis". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 3918–3926.

[160] George Papamakarios and Iain Murray. "Fast ε-free inference of simulation models with Bayesian conditional density estimation". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1028–1036.

[161] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64.

[162] George Papamakarios, Theo Pavlakou, and Iain Murray. "Masked autoregressive flow for density estimation". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2338–2347.

[163] Matthew D Parno and Youssef M Marzouk. "Transport map accelerated markov chain Monte Carlo". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (2018), pp. 645–682.

[164] Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016. DOI: 10.1017/CBO9781316219232.

[165] Gabriel Peyré, Marco Cuturi, et al. "Computational optimal transport". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[166] Natesh S Pillai and Xiao-Li Meng. "An unexpected encounter with Cauchy and Lévy". In: *The Annals of Statistics* (2016), pp. 2089–2097.

[167] Christian Pinto-Gómez, Francesc Pérez-Murano, Joan Bausells, Luis Guillermo Villanueva, and Marta Fernández-Regúlez. "Directed self-assembly of block copolymers for the fabrication of functional devices". In: *Polymers* 12.10 (2020), p. 2432.

[168] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. "On Variational Bounds of Mutual Information". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5171–5180.

[169] Jonathan Poterjoy. "A localized particle filter for high-dimensional nonlinear systems". In: *Monthly Weather Review* 144.1 (2016), pp. 59–76.

[170] Leah F Price, Christopher C Drovandi, Anthony Lee, and David J Nott. "Bayesian synthetic likelihood". In: *Journal of Computational and Graphical Statistics* 27.1 (2018), pp. 1–11.

[171] Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff D Eldredge. "A low-rank ensemble Kalman filter for elliptic observations". In: *arXiv:2203.05120* (2022).

[172] Jian Qin, Gurdaman S Khaira, Yongrui Su, Grant P Garner, Marc Miskin, Heinrich M Jaeger, and Juan J de Pablo. "Evolutionary pattern design for copolymer directed self-assembly". In: *Soft Matter* 9.48 (2013), pp. 11467–11472.

[173] Tom Rainforth, Rob Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. "On nesting Monte Carlo estimators". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4267–4276.

[174] James O Ramsay. "Estimating smooth monotone functions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2 (1998), pp. 365–375.

[175] Garvesh Raskutti and Caroline Uhler. "Learning directed acyclic graph models based on sparsest permutations". In: *Stat* 7.1 (2018), e183.

[176] Herbert E Rauch, CT Striebel, and F Tung. "Maximum likelihood estimates of linear dynamic systems". In: *AIAA Journal* 3.8 (1965), pp. 1445–1450.

[177] Patrick Rebeschini, Ramon Van Handel, et al. "Can local particle filters beat the curse of dimensionality?" In: *The Annals of Applied Probability* 25.5 (2015), pp. 2809–2866.

[178] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2015. ISBN: 9781107069398.

[179] Sebastian Reich. "A nonparametric ensemble transform method for Bayesian inference". In: *SIAM Journal on Scientific Computing* 35.4 (2013), A2013–A2024.

[180] Sebastian Reich. "Data assimilation: The Schrödinger perspective". In: *Acta Numerica* 28 (2019), pp. 635–711.

[181] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR, 2015, pp. 1530–1538.

[182] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The Annals of Applied Probability* 7.1 (1997), pp. 110–120.

[183] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton University Press, 2015.

[184] Donald J Rose, R Endre Tarjan, and George S Lueker. "Algorithmic aspects of vertex elimination on graphs". In: *SIAM Journal on Computing* 5.2 (1976), pp. 266–283.

[185] Murray Rosenblatt. "Remarks on a multivariate transformation". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 470–472.

[186] Jonas Rothfuss, Fabio Ferreira, Simon Walther, and Maxim Ulrich. "Conditional Density Estimation with Neural Networks: Best Practices and Benchmarks". In: *arXiv:1903.00954* (2019).

[187] Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[188] Thomas L Saaty and Robert G Busacker. *Finite Graphs and Networks: an introduction with applications*. McGraw-Hill Book Company, 1965.

[189] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015.

[190] Narayana P Santhanam and Martin J Wainwright. "Information-theoretic limits of selecting binary graphical models in high dimensions". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4117–4134.

[191] Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. "Sparse Cholesky Factorization by Kullback–Leibler Minimization". In: *SIAM Journal on Scientific Computing* 43.3 (2021), A2019–A2046.

[192] Byron Schmuland. "Dirichlet forms with polynomial domain". In: *Math. Japon* 37.6 (1992), pp. 1015–1024.

[193] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. "A generalized representer theorem". In: *International Conference on Computational Learning Theory*. Springer. 2001, pp. 416–426.

[194] Huawei Shen, Xueqi Cheng, Kai Cai, and Mao-Bin Hu. "Detect overlapping and hierarchical community structure in networks". In: *Physica A: Statistical Mechanics and its Applications* 388.8 (2009), pp. 1706–1712.

[195] Bernard W Silverman. "On the estimation of a probability density function by the maximum penalized likelihood method". In: *The Annals of Statistics* (1982), pp. 795–810.

[196] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[197] Scott A Sisson, Yanan Fan, and Mark M Tanaka. "Sequential Monte Carlo without likelihoods". In: *Proceedings of the National Academy of Sciences* 104.6 (2007), pp. 1760–1765.

[198] Laura Slivinski, Elaine Spiller, Amit Apte, and Björn Sandstede. "A hybrid particle–ensemble Kalman filter for Lagrangian data assimilation". In: *Monthly Weather Review* 143.1 (2015), pp. 195–211.

[199] Kathrin Smetana and Olivier Zahm. "Randomized residual-based error estimators for the proper generalized decomposition approximation of parametrized problems". In: *International Journal for Numerical Methods in Engineering* 121.23 (2020), pp. 5153–5177.

[200] Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. "Obstacles to high-dimensional particle filtering". In: *Monthly Weather Review* 136.12 (2008), pp. 4629–4640.

[201] Chris Snyder, Thomas Bengtsson, and Mathias Morzfeld. "Performance bounds for particle filters using the optimal proposal". In: *Monthly Weather Review* 143.11 (2015), pp. 4750–4761.

[202] Yang Song and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution". In: *Advances in Neural Information Processing Systems* 32 (2019).

[203] Alessio Spantini, Ricardo Baptista, and Youssef Marzouk. "Coupling techniques for nonlinear ensemble filtering". In: *SIAM Review* (2022, In Press).

[204] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. "Inference via low-dimensional couplings". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 2639–2709.

[205] Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. "Optimal low-rank approximations of Bayesian linear inverse problems". In: *SIAM Journal on Scientific Computing* 37.6 (2015), A2451–A2487.

[206] Liangjun Su and Halbert White. "A consistent characteristic function-based test for conditional independence". In: *Journal of Econometrics* 141.2 (2007), pp. 807–834.

[207] Liangjun Su and Halbert White. "Testing conditional independence via empirical likelihood". In: *Journal of Econometrics* 182.1 (2014), pp. 27–44.

[208] Arun Suggala, Mladen Kolar, and Pradeep K Ravikumar. "The Expxorcist: Nonparametric Graphical Models Via Conditional Exponential Densities". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 4446–4456.

[209] Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. "Conditional density estimation and simulation through optimal transport". In: *Machine Learning* 109.4 (2020), pp. 665–688.

[210] Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. "Distributional barycenter problem through data-driven flows". In: *arXiv:2104.14329* (2021).

[211] Esteban G Tabak and Cristina V Turner. "A family of nonparametric density estimation algorithms". In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164.

[212] Amirhossein Taghvaei, Jana De Wiljes, Prashant G Mehta, and Sebastian Reich. "Kalman filter and its modern extensions for the continuous-time nonlinear filtering problem". In: *Journal of Dynamic Systems, Measurement, and Control* 140.3 (2018), p. 030904.

[213] Tao Tang and Tao Zhou. "On discrete least-squares projection in unbounded domain with random evaluations and its application to parametric uncertainty quantification". In: *SIAM Journal on Scientific Computing* 36.5 (2014), A2272–A2295.

[214] VN Temlyakov. "Greedy approximation in convex optimization". In: *Constructive Approximation* 41.2 (2015), pp. 269–296.

[215] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg. "Protein design by sampling an undirected graphical model of residue constraints". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 6.3 (2008), pp. 506–516.

[216] Transport Maps Team. *TransportMaps v2.0*. http://transportmaps.mit.edu/docs/example-inverse-sparsity-identification.html. 2018.

[217] Brian L Trippe and Richard E Turner. "Conditional density estimation with Bayesian normalising flows". In: *Bayesian Deep Learning: NIPS 2017 Workshop*. 2018.

[218] Tuyen Trung Truong and Hang-Tuan Nguyen. "Backtracking Gradient Descent Method and Some Applications in Large Scale Optimisation. Part 2: Algorithms and Experiments". In: *Applied Mathematics & Optimization* 84.3 (2021), pp. 2557–2586.

[219] Benigno Uria, Iain Murray, and Hugo Larochelle. "RNADE: The real-valued neural autoregressive density-estimator". In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013.

[220] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[221] Brani Vidakovic. *Statistical modeling by wavelets*. Vol. 503. John Wiley & Sons, 2009.

[222] Cédric Villani. *Optimal transport: Old and New*. Vol. 338. Springer Science & Business Media, 2008.

[223] Martin J Wainwright, John D. Lafferty, and Pradeep K. Ravikumar. "High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression". In: *Advances in Neural Information Processing Systems*. MIT Press, 2007, pp. 1465–1472.

[224] Dilin Wang, Hao Liu, and Qiang Liu. "Variational inference with tail-adaptive f-divergence". In: *Advances in Neural Information Processing Systems* 31 (2018).

[225] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. "Information-theoretic bounds on model selection for Gaussian Markov random fields". In: *2010 IEEE International Symposium on Information Theory*. IEEE. 2010, pp. 1373–1377.

[226] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[227] Antoine Wehenkel and Gilles Louppe. "Unconstrained monotonic neural networks". In: *Advances in Neural Information Processing Systems*. 2019, pp. 1543–1553.

[228] Li Wenliang, Dougal Sutherland, Heiko Strathmann, and Arthur Gretton. "Learning deep kernels for exponential family densities". In: *International Conference on Machine Learning*. 2019, pp. 6737–6746.

[229] Jeffrey S Whitaker and Thomas M Hamill. "Ensemble data assimilation without perturbed observations". In: *Monthly Weather Review* 130.7 (2002), pp. 1913–1924.

[230] Alan S Willsky. "Multiresolution Markov models for signal and image processing". In: *Proceedings of the IEEE* 90.8 (2002), pp. 1396–1458.

[231] CF Jeff Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.

[232] Wei Biao Wu and Mohsen Pourahmadi. "Nonparametric estimation of large covariance matrices of longitudinal data". In: *Biometrika* 90.4 (2003), pp. 831–844.

[233] Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. "Graphical models via univariate exponential family distributions". In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 3813–3847.

[234] Hongkang Yang and Esteban G Tabak. "Conditional Density Estimation, Latent Variable Discovery, and Optimal Transport". In: *Communications on Pure and Applied Mathematics* 75.3 (2022), pp. 610–663.

[235] Tao Yang, Prashant G Mehta, and Sean P Meyn. "Feedback particle filter". In: *IEEE transactions on Automatic control* 58.10 (2013), pp. 2465–2480.

[236] Mihalis Yannakakis. "Computing the minimum fill-in is NP-complete". In: *SIAM Journal on Algebraic Discrete Methods* 2.1 (1981), pp. 77–79.

[237] Ming Yuan and Yi Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (2007), pp. 19–35.

[238] Olivier Zahm, Paul G Constantine, Clémentine Prieur, and Youssef M Marzouk. "Gradient-based dimension reduction of multivariate vector-valued functions". In: *SIAM Journal on Scientific Computing* 42.1 (2020), A534–A558.

[239] Olivier Zahm, Tiangang Cui, Kody Law, Alessio Spantini, and Youssef Marzouk. "Certified dimension reduction in nonlinear Bayesian inverse problems". In: *Mathematics of Computation* (2022, In Press).

[240] Jakob Zech and Youssef Marzouk. "Sparse Approximation of Triangular Transports, Part II: The Infinite-Dimensional Case". In: *Constructive Approximation* (2022), pp. 1–50.

[241] Tong Zhang. "Adaptive forward-backward greedy algorithm for learning sparse representations". In: *IEEE transactions on information theory* 57.7 (2011), pp. 4689–4708.

[242] David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B Lee. "Diagnostics for conditional density models and Bayesian inference algorithms". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 1830–1840.

[243] Shuheng Zhou, John Lafferty, and Larry Wasserman. "Time varying undirected graphs". In: *Machine Learning* 80.2-3 (2010), pp. 295–319.

[244] Mengbin Zhu, Peter Jan Van Leeuwen, and Javier Amezcua. "Implicit equal-weights particle filter". In: *Quarterly Journal of the Royal Meteorological Society* 142.698 (2016), pp. 1904–1919.