Parameterizing transport maps for ensemble data assimilation

by

Daniel Sharp

Submitted to the Center for Computational Science and Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Computational Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2023

Author .....
Center for Computational Science and Engineering
September 11, 2023

Certified by .....
Youssef M. Marzouk
Professor
Thesis Supervisor

Accepted by .....
Nicolas Hadjiconstantinou
Director, Center for Computational Science and Engineering

Parameterizing transport maps for ensemble data assimilation

by

Daniel Sharp

Submitted to the Center for Computational Science and Engineering
on May 24, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Computational Science and Engineering

Abstract

This thesis discusses methods for Bayesian parameter estimation, particularly in the case of state space models (SSMs). We begin by reviewing established methods for filtering in SSMs, and by examining the graphical model structure of a parameterized SSM. Then we discuss established methods for estimating the parameters of such an SSM, making use of its graphical structure. Next we employ monotone triangular transport maps as a method of estimating conditional probability densities and performing conditional sampling, and relate these tasks to the original filtering problem. We provide some practical results and experiments for employing these maps for inference, particularly examining the map parameterization for this function approximation problem. Using these ingredients, we introduce and discuss an algorithm that uses transport to perform online inference of the static parameters of an SSM, and relate this algorithm to prior methods. Finally, we tie the problems of function approximation and static parameter inference together with numerical examples of transport for sequential inference.

Most of the results in this thesis are powered by two software packages that were developed at length over the course of the thesis work: `EnsembleFiltering.jl`, written in Julia for performing automatically-differentiable ensemble-based filtering on the CPU and GPU; and `MParT`, written in C++ for evaluating and training monotone triangular transport maps.

Thesis Supervisor: Youssef M. Marzouk
Title: Professor

Acknowledgments

It takes a community to produce a workable thesis, and so I am thankful for my family and friends who all have been supportive. My parents and brother have all been dependable figures my entire life, enthusiastic about helping me advance my education, and for that I am thankful. Academically, I must admit a certain level of dependence on my colleagues both inside and outside the uncertainty quantification group. In particular, I am indebted to Dallas Foster, Michael Brennan, Paul-Baptiste Rubio, Max Ramgraber, Josh White, and Matt Parno for my understanding of virtually all that lies herein. I would be remiss to forget the support of my advisor, Youssef– both helping when I am in dire need, as well as fostering an incredible group of students of uncertainty (all of whom I would name were the margins not so large). In part, I must also mention Jean Sofronas and Kate Nelson, who have been more than accommodating when I needed flexibility to meet one important deadline or another. I must mention the prior and ongoing support from colleagues, friends, and supervisors from my prior institution, Virginia Tech.

Finally, I could not possibly overstate my thanks towards my partner, Savannah. She contributed constant and enduring support as needed, and certainly received more than an earful about "moving dirt" on several occasions. This was not the most straighforward thesis to write, but thanks to her, my family, and everyone else, it certainly was not the hardest.

Contents

1 Introduction **13**

1.1 Motivation 13

1.2 Literature Review 14

2 Background **17**

2.1 Data Assimilation 18

2.1.1 Kalman Filtering 18

2.1.2 Ensemble Kalman Filtering 20

2.1.3 Parameterized SSM Example 21

2.2 Measure Transport 23

2.2.1 Ensemble Transport for Bayesian Inference 24

2.2.2 Map Parameterization and Estimation 27

3 Approximation Behavior of Transport Map Estimation **33**

3.1 Approximation Bases 33

3.1.1 Hermite Polynomials 34

3.1.2 Hermite Functions 35

3.2 Estimating Expectations with Transport 38

3.3 Numerical Results 43

3.3.1 Hermite Functions 43

3.3.2 Transport Expectations 43

4 Ensemble Transport for State Space Models **47**

4.1 Ensemble Transport for Filtering 47
4.2 Ensemble Transport for Static Parameter Estimation 48
4.3 Numerical Results 51
4.3.1 Ensemble Transport Filtering 51
4.3.2 Augmented State Ensemble Transport Filtering 55

5 Conclusions and Future Work **61**

5.1 Conclusions 61
5.2 Future Work 62
5.2.1 Chaotic State Space Models 62

List of Figures

2-1 Diagram illustrating a triangular transport map $S : \nu \to \mu$ 25

2-2 Diagram of a rectified expansion for the dth map component 30

3-1 Example degree 5 polynomial constructed from Hermite basis 36

3-2 Examples of Hermite functions up to degree 5 37

3-3 Diagram comparing true and estimated densities of pushforward measures 41

3-4 (Left): Creating estimates of the maximizer of ψ_α furthest from the origin, (right): Showing numerical maxima of ψ_α. 44

3-5 Evaluating the error for increasing evaluations; the horizontal axis measures the ratio of evaluations for the estimation method compared to the number of training samples for the map 45

3-6 Evaluating the error for increasing dimension; the number of evaluations is fixed at $10^7 \pm 10\%$ as possible (dimensions 7,8,9 maximize Q such that the number of evaluations Q^n does not exceed 10^7) 46

4-1 Performance of EnTF (Transport filter) versus EnKF and ETKF . . . 54

4-2 Calculating D_J statistic on testing and training datasets over ensemble size J in 100 trials 56

4-3 Pushforward of S_W 57

4-4 Sequential estimation of parameters in linear system 58

4-5 Comparing density estimates of 4e4 samples from the analytical posterior $\pi(W|Y_{1:K})$ to 2e5 samples of the transport-based approximate posterior $\widehat{\pi}(W|Y_{1:K})$ 60

List of Tables

4.1 Summary of distributions in Ensemble Transport Filtering 48

4.2 Frequency of capturing true parameter values for example trajectory with $K = 2000$ and $N = 1000$ 59

List of Algorithms

1 Observation space stochastic EnKF 21

2 Bayesian Inference from Samples using Measure Transport 26

3 Joint state and parameter ensemble transport filtering 49

Chapter 1

Introduction

1.1 Motivation

A major motivation for this work, particularly for static parameter estimation, is in the interest of developing solutions for geophysical problems. The numerical weather prediction (NWP) community in particular has vested interest in tuning complex mathematical models to capture borderline-chaotic behavior, with these estimations happening in an online and computationally tractable manner. NWP is an interesting domain that has well-understood models of how geophysical dynamics should behave in theory, in addition to physical infrastructure observing these dynamics in real time. Since weather behavior is inherently global, one must understand large-scale behavior to predict what happens in local areas. While NWP is on the forefront of data assimilation, many of their methods tend to be variational and could benefit from further explorations into quantifying uncertainty not just on physical behavior, but propagating uncertainty in the geophysical model itself.

Recent work in the vein of measure transport for data assimilation has made it clear that, if one is interested in employing techniques with transport maps, the user must be able to use both scalable *algorithms* and scalable *software*. As such, it was deemed useful to work on software for measure transport that scales well due to a high-performance underlying base and is flexible for users interested in the methods with different applications. For that reason, the tool `MParT` is inherently tied to the

applications discussed in this document and the package `EnsembleFiltering.jl` is an implementation of filtering algorithms to ensure consistent reproducible behavior for all kinds of problems for in state-space models.

1.2 Literature Review

While schemes for estimating the state in a State Space Model (SSM) and quantifying the relevant uncertainty of said states date back to the original Kalman Filter [1], estimating a static parameter jointly with these states in a parameterized SSM has been a challenge with most progress much more recently. Reasonable summaries of traditional solutions to the joint state-and-parameter estimation problem in a Bayesian filtering perspective come from [2], [3]. Särkkä [2] discusses a few different approaches; the simplest being a state augmentation approach, where we treat the parameters as states themselves that vary with the "true" states. However, most algorithms presented end up using variational methods similar to expectation maximization for creating a maximum-likelihood estimator of the true parameter values. To contrast this, Liu and West [3] summarize prior work on estimating states and parameters jointly using sequential Monte Carlo. While these methods are robust to non-Gaussian noise and nonlinear behavior, they suffer the same problems as traditional particle filtering when faced with scalability, particle degeneracy, and sensitivity to proposal distributions. With the exception of state augmentation, all methods relied on within these works require a "smoothing-like" correction—estimating parameters and states jointly at step k requires k iterations of an algorithm, which is substantial for algorithms running over long periods of time or with large state spaces. The fact that these algorithms become more expensive in time makes them "offline", or at least intractable for real-time assimilation of data for static parameter inference. To contrast, [4] covers a variety of algorithms that are both "offline" and "online", which revolve around sequential Monte Carlo methods. This discusses several algorithms similar to [2] from a particle perspective, in addition to extensions for online methods, though several are also MLE-based methods. In particular, it includes discussions on a few

papers that use pseudo-marginal Markov chain Monte Carlo (MCMC) and particle MCMC [5], [6], which gave way to state-of-the-art algorithms. The MCMC in these papers is used in order to correct degeneracy and particle collapse while asymptotically sampling from the correct distribution. The state-of-the-art in this vein is then presented at length in [7], which uses the particle MCMC scheme in a hierarchical sequential Monte Carlo scheme to approximate the states and parameters. The costs of these methods based on particle MCMC is that, in addition to having the drawbacks to general SMC methods, they are still offline and have increasing cost at each timestep.

In parallel with the developments from the particle filtering community, there has been a proliferation of data assimilation methods for general SSMs proposed recently since the introduction of the stochastic Ensemble Kalman Filter (EnKF) in [8], [9]. Some broad characterizations of these "standard" EnKF methods are discussed in [10]–[13], as well as any number of variational methods (which do not approximately or adequately capture uncertainty) such as 3DVAR, 4DVAR, and many of their posterity. Not captured in these texts is [14], which notably ties the data assimilation problem to a larger measure transport framework for sequentially conditioning on observations, then assessing uncertainty in the state space. Some applications come up in [15], [16], which use data assimilation techniques for estimating states of interest, as well as static parameters. These applications largely use state augmentation in practice because, for physical problems, SMC methods are far too expensive to collect a sufficiently large ensemble.

Tying all of these works together are some interesting results from [17], [18], which (similar to [14]) give novel ways of framing not only the state estimation problem, but the problem of jointly estimating parameters as well. By using modern computing automatic differentiability tools, [17] has a very simple hybrid ensemble/variational framework which allows for reasonably advanced computational tools for parameter estimation while retaining uncertainty in the states, but the parameter estimation step has some offline cost. In contrast, by taking a functional tensor-train approach, [18] has a very complex way of relating density function approximation to the se-

quential Bayesian posterior estimation step seen in traditional filtering. By using an expressive enough basis, the density function approximation method extends readily to approximation of the static parameters in the state space model. Particle flow filters [13], [19] are based on Stein variational gradient descent (SVGD) [20] to solve a flow problem similar to, e.g., normalizing flows and score-matching, developing a method to sequentially condition the states on the observations by estimating the flow of particles through a transport PDE determined by the prior and posterior. This method, similar to the measure transport approaches, is a powerful tool in extending the traditional EnKF and variational methods to nonlinear problems.

Chapter 2

Background

As with many other problems in uncertainty quantification, we take a Bayesian approach to solving a type of inverse problem; in particular, we are interested in Bayesian data assimilation for SSMs. Suppose we have a Gaussian SSM and access to functions $F : \mathbb{R}^n \to \mathbb{R}^n$, $H : \mathbb{R}^n \to \mathbb{R}^m$ that describe a process

$$\begin{aligned} X_k &= F(X_{k-1}) + \xi_k, \quad \xi_k \sim \mathcal{N}(0, \mathbf{Q}) \\ Y_k &= H(X_k) + \gamma_k, \qquad \gamma_k \sim \mathcal{N}(0, \mathbf{R}) \end{aligned} \tag{2.1}$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$, $\mathbf{R} \in \mathbb{R}^{m \times m}$. The subscript $(\cdot)_k$ indicates the variable is associated with some time t_k. In this framework, we assume a prior on X_0 and that we observe $\mathbf{y}_k$ (realizations of Y_k) for $k = 1, \cdots, K$. While we know the covariance matrices $\mathbf{Q}, \mathbf{R}$, and the operators F, H, we do not know X_k with certainty for any k.

To demonstrate, imagine the motivating example of weather prediction. Supposing X_k is a vector of wind velocities on a spatial domain at time k, one can take F as the mathematical model of how the wind velocities evolve in time. This often takes the form of solving a partial differential equation (PDE) from time t_{k-1} to time t_k. Finally, we place down physical wind velocity sensors on a coarse grid to gather observation $\mathbf{y}_k$ of the wind velocities at time k. In this setting, X_k would be a PDE solution on a (possibly unstructured) mesh and the placement of observation sensors corresponding to Y_k may not be exactly on the mesh corresponding to X_k,

therefore our observation operator H would be some sparse interpolation operator for X_k. Further, the sensors are likely to have inaccuracies, hence the observation noise covariance $\mathbf{R}$. Therefore, this example SSM estimates wind velocities on a coarse grid given an approximate state model as well as noisy and sparse observations.

In this example and many others, F is likely to be a function of different "static" parameters W; a time-dependent weather forecasting PDE will likely be parameterized by viscosity, humidity, and temperature for example. While these are not truly static, they change at a much longer timescale than the wind velocities. Therefore, as we may be uncertain about the exact value of W, the underlying motivation for this thesis is to be able to jointly estimate X_k and W as we record new realizations of the observations Y_k.

2.1 Data Assimilation

Most generally, data assimilation is the study and field of using both mathematical models and observations to infer quantities of interest and gauge uncertainty on these estimations. The goal of data assimilation is obtaining results that are more informative than using models or observations on their own. It is clear from the above that sequentially estimating the state X_k in the SSM shown in eq. (2.1) is one form of data assimilation.

2.1.1 Kalman Filtering

Suppose we know that X_0 follows some distribution D; then, $p(X_k|X_{k-1} = \mathbf{x}_{k-1}) = \mathcal{N}(F(\mathbf{x}_{k-1}), \mathbf{Q})$, so we could sample $\widehat{X}_0 \sim D$ and propagate it through the model K times to obtain a sample of the marginal $p(X_K)$. However, if $\mathbf{Q}$ is nonzero, the distribution of $\widehat{X}_K$ will be extremely diffuse. Further, it would be entirely independent of the data realizations $\{\mathbf{y}_k\}$, which would hopefully give information as to where the X_K should be concentrated.

Calculating conditional covariance, it follows

$$\begin{aligned}\mathbb{C}\text{ov}(Y_k|X_k) &= \mathbb{C}\text{ov}(H(X_k)+\gamma_k)\\ &= \mathbb{C}\text{ov}(H(X_k)) + 2\mathbb{C}\text{ov}(H(X_k),\gamma_k) + \mathbb{C}\text{ov}(\gamma_k)\\ &= \mathbb{C}\text{ov}(H(X_k)) + \mathbf{R},\end{aligned}$$

since γ_k is independent from X_k. If $H(\mathbf{x}) = \mathbf{H}\mathbf{x}$, then we get $\mathbb{C}\text{ov}(H(X_k)) = \mathbf{H}\mathbb{C}\text{ov}(X_k)\mathbf{H}^\top$ by the properties of the covariance. Finally,

$$\mathbb{C}\text{ov}(X_k, Y_k) = \mathbb{C}\text{ov}(X_k, H(X_k)+\gamma_k) = \mathbb{C}\text{ov}(X_k, H(X_k)).$$

Now we assume that X_k, Y_k are **jointly normally distributed**. Let $\mathbb{E}[X_k] = \mathbf{m}_k^-$ and $\mathbb{C}\text{ov}(X_k) = \mathbf{P}_k^-$. In the normal X_k, Y_k case, we see

$$p\left(\begin{bmatrix}X_k\\ Y_k\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_k^-\\ \mathbb{E}[H(X_k)]\end{bmatrix}, \begin{bmatrix}\mathbf{P}_k^- & \mathbb{C}\text{ov}(X_k, H(X_k))\\ \mathbb{C}\text{ov}(H(X_k), X_k) & \mathbb{C}\text{ov}(H(X_k)) + \mathbf{R}\end{bmatrix}\right),$$

where, when $H(\mathbf{x}) = \mathbf{H}\mathbf{x}$,

$$p\left(\begin{bmatrix}X_k\\ Y_k\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_k^-\\ \mathbf{H}\mathbf{m}_k^-\end{bmatrix}, \begin{bmatrix}\mathbf{P}_k^- & \mathbf{P}_k^-\mathbf{H}^T\\ \mathbf{H}\mathbf{P}_k^- & \mathbf{H}\mathbf{P}_k^-\mathbf{H}^T + \mathbf{R}\end{bmatrix}\right),$$

which agrees with what one might see in, e.g., [2]. Then, using facts of Gaussian distributions [21],

$$\begin{aligned}p(X_k|Y_k = \mathbf{y}_k) = \mathcal{N}\left(\mathbf{m}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbb{E}[H(X_k)]), \mathbf{P}_k^- - \mathbf{K}_k\mathbb{C}\text{ov}(H(X_k), X_k)\right),\\ \mathbf{K}_k = \mathbb{C}\text{ov}(X_k, H(X_k))(\mathbb{C}\text{ov}(H(X_k)) + \mathbf{R})^{-1},\end{aligned} \tag{2.2}$$

where $\mathbf{K}_k$ is defined as the *Kalman gain*. Again taking the H linear case, note

$$\begin{aligned}p(X_k|Y_k = \mathbf{y}_k) = \mathcal{N}(\mathbf{m}_k^- + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}\mathbf{m}_k^-),\ \mathbf{P}_k^- - \mathbf{K}_k\mathbf{H}\mathbf{P}_k^-),\\ \mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}^\top(\mathbf{H}\mathbf{P}_k^-\mathbf{H}^\top + \mathbf{R})^{-1}.\end{aligned}$$

We can relax the constraint of the joint distribution $p(X_k, Y_k)$ to say that X_k, Y_k are jointly normal conditioned on $Y_{1:k-1} = \mathbf{y}_{1:k-1}$. The difficulty here is that the covariances, e.g., $\mathbb{C}\text{ov}(H(X_k)|Y_{1:k-1})$, are very difficult to calculate for nonlinear operators. Further, the joint conditional Gaussianity constraint on X_k, Y_k is difficult to assume without linear F, H anyway. Therefore, this joint normality of X_k, Y_k conditioned on $Y_{1:k-1}$ is a **vital** assumption to easily guarantee the behavior of the Kalman filter.

2.1.2 Ensemble Kalman Filtering

Suppose we have $\{\mathbf{x}_{k-1}^{(j)}\}_{j=1}^{J} \sim p(X_{k-1}|Y_{1:k-1} = \mathbf{y}_{1:k-1})$, which are samples of the filtering (also known as "analysis") distribution at time $k-1$. Then, if $\widehat{\xi}_k^{(j)} \sim \mathcal{N}(0, \mathbf{Q})$ and $\widehat{\mathbf{x}}_k^{(j)} = F(\mathbf{x}_{k-1}^{(j)}) + \widehat{\xi}_k^{(j)}$, our samples must follow $\{\widehat{\mathbf{x}}_k^{(j)}\}_{j=1}^{J} \sim p(X_k|Y_{1:k-1})$ by definition. We can create Monte Carlo estimates of covariance terms used in Kalman filtering via the empirical distribution of $\{\widehat{\mathbf{x}}_k^{(j)}\}$ to estimate the Kalman gain $\widehat{\mathbf{K}}_k$. Finally, if we know that the mean of the distribution is shifted by $\mathbf{K}_k(\mathbf{y}_k - \mathbb{E}[H(X_k)])$ in the traditional Kalman filter, we can shift each particle $\widehat{\mathbf{x}}_k^{(j)}$ by $\widehat{\mathbf{K}}_k\left(\mathbf{y}_k - \widehat{\mathbf{y}}_k^{(j)}\right)$, where

$$\widehat{\mathbf{y}}_k^{(j)} \sim p(Y_k|X_k = \widehat{\mathbf{x}}_k^{(j)}, Y_{1:k-1} = \mathbf{y}_{1:k-1}) = \mathcal{N}(H(\widehat{\mathbf{x}}_k^{(j)}), \mathbf{R}).$$

This is called the **stochastic ensemble Kalman filter** (usually just EnKF). The term "stochastic" comes from the fact that $\widehat{\mathbf{y}}_k^{(j)}$ is a sample from the *forecasted* observation distribution and it is not a statistic of said distribution. Because this will be referenced diligently throughout the text, we write out algorithm 1 explicitly.

The description of algorithm 1 as "observation space" comes from the fact that, in order to work with general observation operators, we work with the sample covariance of an observation ensemble. The traditional state space EnKF estimates $\mathbf{P}_k^- \in \mathbb{R}^{n\times n}$ directly as opposed to estimating $\mathbf{HP}_k^- \in \mathbb{R}^{m\times n}$; when H is not linear, this is impossible to perform. While we are forced to work in observation-space from a practical standpoint, this algorithm is generally more scalable than the state space equivalent for the motivating physical applications, as $X_k \in \mathbb{R}^n$ and $Y_k \in \mathbb{R}^m$, with $m \ll n$, i.e,

Algorithm 1: Observation space stochastic EnKF

input : Initial ensemble $\{\mathbf{x}_{k-1}^{(j)}\}_{j=1}^{J} \sim p\left(X_{k-1}|Y_{1:k-1} = \mathbf{y}_{1:k-1}\right)$, new data $\mathbf{y}_k$

output: $\{\mathbf{x}_k^{(j)}\}_{j=1}^{J} \overset{approx}{\sim} p\left(X_k|Y_{1:k} = \mathbf{y}_{1:k}\right)$

`/* Forecast */`

Sample $\widehat{\xi}_k^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$;

$\widehat{\mathbf{x}}_k^{(j)} \leftarrow F(\mathbf{x}_k) + \widehat{\xi}_k^{(j)}$;

`/* Analysis */`

$\widehat{\mathbf{m}}_k^X \leftarrow \mathbb{E}[\widehat{\mathbf{x}}_k^{(j)}], \quad \widehat{\mathbf{m}}_k^H \leftarrow \mathbb{E}[H(\widehat{\mathbf{x}}_k^{(j)})]$;

Form matrices $\widehat{\mathbf{X}}_k \leftarrow [\widehat{\mathbf{x}}_k^{(j)} - \widehat{\mathbf{m}}_k^X], \quad \widehat{\mathbf{H}}_k \leftarrow [H(\widehat{\mathbf{x}}_k^{(j)}) - \widehat{\mathbf{m}}_k^H]$;

Sample $\widehat{\gamma}_k^{(j)} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$;

$\mathbf{x}_k^{(j)} \leftarrow \widehat{\mathbf{x}}_k^{(j)} + \widehat{\mathbf{X}}_k \widehat{\mathbf{H}}_k^\top \left(\widehat{\mathbf{H}}_k \widehat{\mathbf{H}}_k^\top + \mathbf{R}\right)^{-1} \left(\mathbf{y}_k - H(\widehat{\mathbf{x}}_k^{(j)}) + \widehat{\gamma}_k\right)$;

we observe on a much smaller dimension than we would like to predict.

2.1.3 Parameterized SSM Example

Suppose that we take the SSM expressed in eq. (2.1) and parameterize it by some random variable $W \in \mathbb{R}^{n_w}$, which correspond to "static parameters" as described earlier.

$$\begin{aligned} X_k &= F(X_{k-1}; W) + \xi_k, \quad \xi_k \sim \mathcal{N}(0, \mathbf{Q}_W) \\ Y_k &= H(X_k; W) + \gamma_k, \qquad \gamma_k \sim \mathcal{N}(0, \mathbf{R}_W) \end{aligned} \tag{2.3}$$

We note now that, in this setting, X_k is a function of W for all k, so inferring or estimating, e.g., $\mathbf{K}_k$ requires knowledge of W and thus samples of X_k are only valid if W stays constant in time. As an example, take the linear system

$$\begin{aligned} X_k &= \theta X_{k-1} + q\xi_k, \quad \xi_k \sim \mathcal{N}(0, \mathbf{I}) \\ Y_k &= X_k + r\gamma_k, \qquad \gamma_k \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{2.4}$$

with $X_0 \sim \mathcal{N}(0, \eta_0 \mathbf{I})$. Supposing that $\mathbf{P}_{k-1} = \eta_{k-1}\mathbf{I}$ for the purpose of induction, observe $\mathbf{P}_k^- = (\theta^2 \eta_{k-1} + q^2)\mathbf{I} =: \lambda_k \mathbf{I}$. Thus, since $\mathbf{H} = \mathbf{I}$,

$$\mathbf{K}_k = \mathbf{P}_k^-(\mathbf{P}_k^- + r^2\mathbf{I})^{-1} = \lambda_k(\lambda_k + r^2)^{-1}\mathbf{I} =: \omega_k \mathbf{I},$$

Finally, this leads to

$$\mathbf{m}_k = \theta\mathbf{m}_{k-1} + \mathbf{K}_k(\mathbf{y}_k - \theta\mathbf{m}_{k-1}) = \theta(1-\omega_k)\mathbf{m}_{k-1} + \omega_k\mathbf{y}_k$$
$$\mathbf{P}_k = \mathbf{P}_k^- - \mathbf{K}_k\mathbf{P}_k^- = (1-\omega_k)\lambda_k\mathbf{I} =: \eta_k\mathbf{I}$$

Observe that λ_k, η_k both depend heavily on the parameters W. Finally, note that we desire the joint posterior $p(X_k, W|Y_{1:k}) \propto p(X_k|Y_{1:k}, W)p(Y_{1:k}|W)p(W)$. Since we condition on the data $Y_{1:k} = \mathbf{y}_{1:k}$, this normalization is constant. Analytically, we can use the Kalman filter in eq. (2.2) to find

$$p(Y_{1:k}|W) = \prod_{j=1}^{k} p(Y_j|Y_{1:j-1}, W) = \prod_{j=1}^{k} \mathcal{N}(Y_j; \theta\mathbf{m}_{j-1}, (\lambda_j + r^2)\mathbf{I}) \tag{2.5}$$

Then, the procedure to sample from the joint distribution $p(X_k, W|Y_{1:k})$ becomes trivial in practice, and we can use a standard method (e.g., maximum-likelihood, maximum a posteriori estimation, conditional expectation) to estimate the realization of W given our observations, then filter X_k for future time with fixed parameters $\mathbf{w}$. However, note two connected computational difficulties in this near-trivial example; first, λ_k and $\mathbf{m}_k$ are both heavily dependent on the realization of W. Estimating the posterior $p(W|Y_{1:k})$ requires filtering all k steps to get the intermediate λ_k and $\mathbf{m}_k$ variables. Second, while we can easily calculate the unnormalized PDF of the posterior for a particular value $\mathbf{w}$, if we wanted to learn from the data and create a new estimate of the parameters at step k, say $\mathbf{w}'$, we would have to recognize that this would change each and every $\mathbf{m}_k$ and thus the posterior estimation of X_k would be different at any given step. While we know that the true posterior will asymptotically concentrate around the true value of $\mathbf{w}$, this may not be the case if we were to empirically estimate the quantities in an "online" fashion (i.e. without reexamining X_j for $j = 1, \cdots, k$). If we create a new estimate $\mathbf{w}^{(k)}$ at each timestep, we forecast samples of the joint distribution $p(X_{k-1}|Y_{1:k-1}, W)p(W|Y_{1:k-1})$ to the next timestep via the forward operator $p(X_k|X_{k-1}, Y_{1:k-1}, W')$. Note, however, that X_{k-1} has explicit parametric dependence on random variable W and X_k now has explicit parametric dependence only on W', which has an unknown relationship to

W. Therefore, since we do not explicitly marginalize out our realization of W from the previous timestep when estimating the current parameter value W', we introduce some error into our estimate. We speculate that, by neglecting to take the history of $\mathbf{w}^{(k)}$ into account, an online method forfeits the asymptotic concentration that a traditional posterior would display.

2.2 Measure Transport

Measure transport is a field of mathematics often framed in the optimality setting. At the most theoretic level, given two sigma algebras $\mathcal{F}_1, \mathcal{F}_2$ and sample spaces Ω_1, Ω_2, we take two measures μ, ν to create measurable spaces $(\mu, \mathcal{F}_1, \Omega_1)$ and $(\nu, \mathcal{F}_2, \Omega_2)$. Then, we seek to find a meaningful coupling $\gamma_{\mu,\nu} \in \Gamma(\mu, \nu)$ where

$$\Gamma(\mu,\nu) = \{\gamma : \mathcal{F}_1 \times \mathcal{F}_2 \to \mathbb{R} \mid \gamma(A, \Omega_2) = \mu(A)\ \forall A \in \mathcal{F}_1,\ \gamma(\Omega_1, B) = \nu(B)\ \forall B \in \mathcal{F}_2\}.$$

One example of such a coupling $\gamma \in \Gamma(\mu, \nu)$ is the trivial $\gamma(A, B) = \mu(A)\nu(B)$, thus $\Gamma(\mu, \nu)$ cannot be empty. Another example would be finding the optimal coupling $\widehat{\gamma}_{\mu,\nu} = \arg\min_\gamma \int_{\Omega_1 \times \Omega_2} c(z, x)\, d\gamma(z, x)$, which is exactly the Kantorovich optimal transport problem for some cost function $c : \Omega_1 \times \Omega_2 \to \mathbb{R}^+$. In particular, if $c(z, x) = \|z - x\|^2$, then we get the optimal transport coupling induced by the Wasserstein-2 distance [22]. Assuming that μ is absolutely continuous under the Lebesgue measure, we can further try to find some deterministic transformation $S : \Omega_1 \to \Omega_2$ such that $Z = S(X)$ for random variables Z, X on the sample spaces Ω_2, Ω_1 respectively; this transformation S is often called a *transport map* from ν to μ, and induces the coupling of $\gamma(A, B) = \mu(A)\mu(S(B))$ since $\nu(B) = \mu(S(B))$ must hold. Intuitively, this is saying that there can be some transformation S that maps between two distributions with random variables Z and X. This transformation S can exist with fairly mild assumptions [22] and often is not unique in the same way that couplings between the measures are not unique. Given some transport map S which pushes ν to μ, we can define the measures that the image and preimage of

S live on. The *pushforward* of ν under S is given by $\mu = S_\sharp \nu := \nu \circ S^{-1}$, i.e., the image of S lives in the pushforward measure. The *pullback* of μ under S is given by $\nu = S^\sharp \mu = \mu \circ S$, i.e., the preimage of S lives in the pullback measure. Intuitively, if $Z = S(X)$, then Z lives in the pushforward distribution and $X = S^{-1}(Z)$ lives in the pullback distribution.

In this document, we focus on the Knothe-Rosenblatt (KR) rearrangement for measures with support over all real numbers, i.e., $\Omega_1 = \Omega_2 = \mathbb{R}^n$. We define a KR transport constructively using a map componentwise in n dimensions: if S is a deterministic transport map defining a KR rearrangement, then $[S(x_1, \ldots, x_n)]_d = C_d(x_d; x_1, \ldots, x_{d-1})$ where the function $x_d \mapsto C_d(x_d \cdot; x_1, \ldots, x_{d-1})$ is a monotone function. Colloquially, this means that C_d, the dth *map component* of S, is independent of $d+1, d+2, \ldots, n$ and importantly that C_d is monotone in its last argument. Note that, if μ has support on the entirety of $\mathbb{R}^n$, then $C_d(\cdot; x_1, \ldots, x_{d-1})$ must be a bijective function as the monotonicity forces injectivity. Since there must be some X such that $C_d(x_d; x_1, \ldots, x_{d-1}) = z_d$ for any valid value of z_d, C_d must be surjective as well. This kind of transport has earned a moniker of "monotone triangular transport", or just "triangular transport", due to the behavior of the Jacobian of S. Since C_{d_1} is independent of x_{d_2} when $d_2 > d_1$, all entries of the Jacobian above the diagonal are zero, i.e., $\partial_{d_2} C_{d_1} = 0$ for $d_2 > d_1$. Further, $\partial_d C_d > 0$ from the monotonicity constraint.

In fig. 2-1, we see a two-dimensional triangular transport example where the pushforward of ν under S is μ, so $(Y, X) \sim \nu$ implies then $S(Y, X) \sim \mu$. If we have sets $A_X, A_Y, B_1, B_2 \subseteq \mathbb{R}$ where B_1, B_2 are respectively the images of A_X, A_Y under S, then

$$\nu(A_Y \otimes A_X) = S_\sharp \nu(B_2 \otimes B_1) = S^\sharp \mu(A_Y \otimes A_X) = \mu(B_2 \otimes B_1).$$

2.2.1 Ensemble Transport for Bayesian Inference

For this thesis, we are primarily concerned with using triangular transport for sample-based Bayesian inference, which we can imagine extending from the cartoon drawn in fig. 2-1. Suppose we have random variables X, Y where $Y \in \mathbb{R}^m$ following distribution

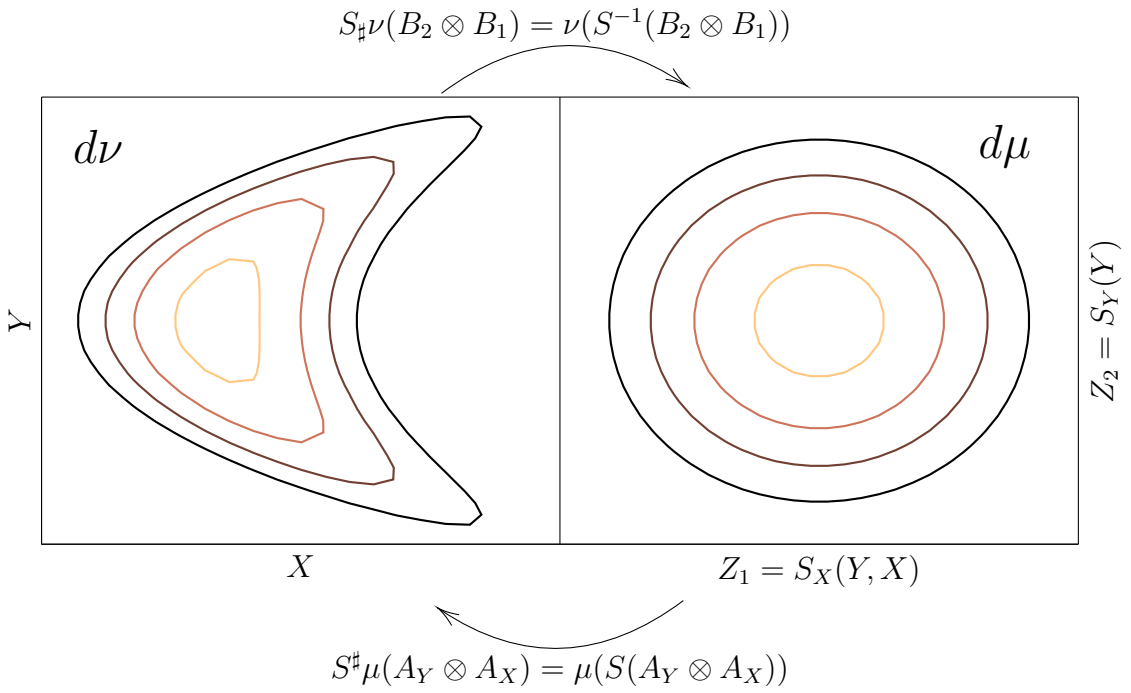

Figure 2-1: Diagram illustrating a triangular transport map $S : \nu \to \mu$

π_Y is an observation of a phenomenon and $X \in \mathbb{R}^n$ following distribution π_X is the corresponding state, so we can easily draw samples from the likelihood $p(Y|X = \mathbf{x})$ given a sample $\mathbf{x}$. Then, with sufficient samples $(X, Y) \in \mathbb{R}^{n+m}$ of the joint distribution on the measure $\nu = \nu_X \otimes \nu_{Y|X}$, we can construct a map $S(Y, X) : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ that pushes ν to some "reference" measure $\mu_1 \otimes \mu_2$, where μ_1 and μ_2 are respectively induced by independent distributions η_{Z_y} and η_{Z_x}. Note that S has a particular structure

$$S(Y, X) = \begin{bmatrix} S_Y(Y) \\ S_X(Y, X) \end{bmatrix} = \begin{bmatrix} C_1(Y_1) \\ \vdots \\ C_m(Y_m; Y_1, \ldots, Y_{m-1}) \\ C_{m+1}(X_1; Y) \\ \vdots \\ C_{m+n}(X_n; Y, X_1, \ldots, X_{n-1}) \end{bmatrix}, \quad \partial_d C_d > 0,$$

where $d = 1, \ldots, m+n$. The pushforward identities in this case are $S_\sharp \nu = \mu$, i.e., μ is the pushforward of ν under S, $S_{Y\sharp}\nu_Y = \mu_1$, i.e., S_Y pushes ν_Y forward to μ_1, and $S_{X\sharp}\nu = \mu_2$, i.e., S_X pushes ν forward to μ_2.

Suppose we have samples $\{\widehat{\mathbf{x}}^{(j)}\}_{j=1}^J$ and want to generate samples from a distribution $p(X|Y = \mathbf{y})$. Then, we draw $\widehat{\mathbf{y}}^{(j)} \sim p(Y|X = \widehat{\mathbf{x}}^{(j)})$ to create valid samples $(\widehat{\mathbf{x}}^{(j)}, \widehat{\mathbf{y}}^{(j)})$ from the joint distribution of X and Y. The map "block" S_X creates valid samples $\mathbf{z}^{(j)} = S_X(\widehat{\mathbf{y}}^{(j)}, \widehat{\mathbf{x}}^{(j)})$ from distribution η_{Z_x}, which follows from looking at the pushforward measure of ν under $S_X : \mathbb{R}^{m+n} \to \mathbb{R}^n$. Finally, we recall the distribution of interest to be $p(X|Y = \mathbf{y})$ which induces measure $\nu_{X|\mathbf{y}}$. Intuitively, we note that $S_X(\mathbf{y}, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ pushes $\nu_{X|\mathbf{y}}$ to μ_2. Why is this? Formally, S_X only pushes from the joint measure to the reference measure. However, if we fix $\mathbf{y}$, then any corresponding coupled input $\mathbf{x}$ must satisfy that $(\mathbf{y}, \mathbf{x})$ is a valid sample from the joint distribution, and thus $\mathbf{x}$ must be sampled from the conditional distribution $p(X|Y = \mathbf{y})$ by the definition of the conditional distribution. The function $S_X(\mathbf{y}, \cdot)$ is indeed invertible, so we can use the pullback definition to observe $\nu_{X|\mathbf{y}} = S_X(\mathbf{y}, \cdot)^\sharp \mu_2 = S_X(\mathbf{y}, \cdot)^\sharp \left(S_{X\sharp}\nu\right)$. Stated more clearly, we can find that $S_X(\mathbf{y}, \cdot)^{-1}(\mathbf{z}^{(j)})$ is a valid sample of the conditional distribution $p(X|Y = \mathbf{y})$. This suggests a workhorse conditional sampling algorithm [23], [24] using transport maps in algorithm 2 used in, e.g., [25], [26].

Algorithm 2: Bayesian Inference from Samples using Measure Transport

input : Prior samples $\{\widehat{\mathbf{x}}_{k-1}^{(j)}\}_{j=1}^J \sim p(X)$, observation data $\mathbf{y}$, observation/likelihood model $p(Y|X)$
output: $\{\mathbf{x}^{(j)}\}_{j=1}^J \sim p\left(X|Y\right)$
Sample $\widehat{y}^{(j)} \sim p(Y|X = \mathbf{x}^{(j)})$;
Create transport map $S_X : \mathbb{R}^{m+n} \to \mathbb{R}^n$ satisfying $S_{X\sharp}\nu = \mu_2$;
Create ensemble $\mathbf{z}^{(j)} = S_X(\widehat{\mathbf{x}}^{(j)}, \widehat{\mathbf{y}}^{(j)})$;
Generate samples $\mathbf{x}^{(j)} = S_X(\mathbf{y}, \cdot)^{-1}(\mathbf{z}^{(j)})$

Algorithm 2, an example of *simulation-based inference*, has a few interesting properties. We note that the only structural assumption was really that the original S is "block triangular" and $S_X(\mathbf{y}, \cdot)$ is invertible for any fixed $\mathbf{y}$. The exact structure of S_X is entirely left to implementation. Indeed, other types of transport [27] can be used in practice. Further, we note that S_Y is entirely unused, so we can choose

to only estimate S_X and not S entirely. Additionally, assuming that S_X is an exact transport map and the prior ensemble is valid, this indeed gives an algorithm to sample from the posterior $p(X|Y)$ exactly. Finally, this map S_X is very flexible: if one were interested in generatively sampling from $p(X|Y)$, we need only to be able to sample from η_{Z_x} (often assumed to be Gaussian); if we are interested in conditioning on a different observation $\mathbf{y}^*$, we can reuse the ensemble of samples $\mathbf{z}^{(j)}$; if we are interested in density estimation, we need only to be able to evaluate η_{Z_x}, the density of our reference measure μ, and the determinant of our map's jacobian, $|\nabla S_X|$.

2.2.2 Map Parameterization and Estimation

Map Estimation

With the exception of single-dimensional problems, we seldom know the exact monotone triangular transport map S that pushes ν to μ, or even the structure of such a transport map. Therefore, the estimation of S ends up becoming a multidimensional function approximation problem. Here, we strive to find a parametric solution by restricting ourselves to a finite-dimensional set of admissible functions, then attempting to find the best solution in that set based on some objective. At this point, *we restrict ourselves to solving problems to estimate transport maps from samples following measure* ν, where we assume μ is well-understood.

In this scenario, we start by assuming μ admits a density η and ν admits a density π (where we abuse notation and use π and η to denote the distributions that they determine). Unlike many function approximation problems, there are generally no samples $(Z^{(j)}, X^{(j)})$ following a coupling of μ and ν. For example, the simple least-squares regression problem of the form

$$\arg\min_{U \in \mathcal{F}} \sum_{j=1}^{J} \|\mathbf{z}_j - U(\mathbf{x}_j)\|^2$$

is entirely insufficient for our setting of measure transport, due to this lack of coupling. Further, even if we induced matching samples somehow, we have no intuition

why our transport map approximation might work for samples outside of the training dataset since the sample matching could be entirely pathological. Using this uninformed objective disregards our substantial knowledge of the structure of probability distributions and their properties. Therefore, we instead opt to choose something derived from literature and recall the Kullback-Leibler divergence $\mathcal{D}_{KL}$. Since we assume μ is well-understood, we just need to estimate how well the pullback is approximated; i.e., for arbitrary map U, we compare π, the density of our data, with $U^\sharp\eta$, the pullback density from our reference, if we want to evaluate how close $U^{-1}(Z)$ is to following density π. Formally, given an identically and independently sampled set $\{\mathbf{x}^{(j)}\}_{j=1}^J \sim \pi$, we seek to estimate the solution to the following minimization problem via its approximation

$$S = \arg\min_{U\in\mathcal{F}} \mathcal{D}_{KL}(\pi||U^\sharp\eta) = \int \pi(x)\log\left(\frac{\pi(x)}{U^\sharp\eta(x)}\right)\,dx \tag{2.6}$$

$$\begin{aligned}\widehat{S} &= \arg\min_{U\in\mathcal{F}} \sum_{j=1}^{J} \log\pi\left(\mathbf{x}^{(j)}\right) - \log U^\sharp\eta\left(\mathbf{x}^{(j)}\right) \\ &= \arg\max_{U\in\mathcal{F}} \sum_{j=1}^{J} \log\eta\left(U(\mathbf{x}^{(j)})\right) + \log\det\nabla U(\mathbf{x}^{(j)}),\end{aligned} \tag{2.7}$$

where, using the change of variables formula for a density, eq. (2.7) is the Monte Carlo approximation of eq. (2.6) via random samples $\mathbf{x}^{(j)}$, so $\widehat{S}$ is a Monte Carlo approximation of S. Therefore, the difficulty of this approximation problem comes down to the difficulty of working with the density η of our reference measure μ, and the complexity of the function space $\mathcal{F}$. Assuming that π has support over the entire real line, the first issue can easily be resolved by choosing the standard normal measure for μ, so $\eta(x) = \mathcal{N}(x; \mathbf{0}, \mathbf{I})$, which has many desirable properties. This in fact allows one to take the n-dimensional problem eq. (2.7) and make it into n independent problems, since η acts independently on each dimension of $U(x)$, which is well covered [23]–[26].

Map Parameterization

When estimating a monotone triangular transport map, the approximating function class is perhaps the most important and most complicated choice. Ensuring a sufficiently expressive function space can be difficult and one has to further ensure that monotonicity is never violated. We focus here on "rectified basis expansions" [28], [29]. As in many approximation problems, we choose some finite set of one dimensional functions $\mathcal{F}_d = \{\psi_\alpha\}_\alpha$ hierarchically indexed by integer α. We then work with $\mathcal{F}_d$ when approximating the map component $d \leq n$, i.e., $\mathcal{F}_d$ determines the class of function C_d satisfying $C_d(X_d; X_1, \ldots, X_{d-1}) \sim \mathcal{N}$ which is monotone in its X_d. We generally will consider polynomial-like functions, but $\mathcal{F}_d$ can instead be made up of radial basis functions (RBFs), wavelets, or other approximation bases. It is not actually necessary that we keep $\mathcal{F}_d$ identical for each output dimension d, and local functions like wavelets and RBFs may admit a different method of representing the functional hierarchy. Then, we choose a set of "multiindices" (often called a multiindex set) $\mathcal{A}_d$ as our finite basis for approximation for component d, which we give cardinality $|\mathcal{A}_d| = L$. One multiindex is the multivariate extension of an index in one dimension. For example, if the ℓth multiindex is $(2, 4, 7)$, we know that we are in dimension $d = 3$ and the ℓth multivariate basis function is expressed as $f^{(\ell)}(x, y, z) = \psi_2(x)\psi_4(y)\psi_7(z)$. Formally, if we have the ℓth multiindex $\boldsymbol{\alpha}^{(\ell)}$, we create the ℓth multivariate basis function as $f^{(\ell)}(t_1, \ldots, t_d) = \psi_{\alpha_1^{(\ell)}}(t_1)\psi_{\alpha_2^{(\ell)}}(t_2)\cdots\psi_{\alpha_d^{(\ell)}}(t_d)$. One can imagine this as a product of, e.g., orthogonal polynomials of t_1, t_2, etc. From here we have L real-valued functions $f^{(\ell)} : \mathbb{R}^d \to \mathbb{R}$, which we take as a basis for our expansion, or $\widehat{C}_d(\mathbf{x}; \mathbf{c}) = \sum c_\ell f^{(\ell)}(\mathbf{x})$. However, one can easily note that this is not generally monotone without further restrictions on $\mathcal{F}_d$ and our coefficients $\mathbf{c} \in \mathbb{R}^L$.

To enforce monotonicity, we choose a rectifier g that is invertible and positive everywhere. While we might desire certain properties of g, we need only that it is positive, invertible, and differentiable. Then, we can create an integral expression

that guarantees monotonicity by construction,

$$C_d(\mathbf{x}, \mathbf{c}) = \widehat{C}_d(0; x_1, \ldots, x_{d-1}, \mathbf{c}) + \int_0^{x_d} g\left(\partial_t \widehat{C}_d(t; x_1, \ldots, x_{d-1}, \mathbf{c})\right) \, dt,$$

which follows from the fact that $\int_0^s g(r(t))\, dt$ is an increasing function in s for any choice function $r : \mathbb{R} \to \mathbb{R}$, since $g(r(t)) > 0$ for all t comes for free from the construction of g as a positive bijector. This "monotone multivariate expansion" is summarized by fig. 2-2. Even though $\mathbf{c}$ gives coefficients for $\widehat{C}_d$, they become parameters of C_d that affect the function in a nonlinear way (i.e., they are no longer simple "coefficients"). Similarly, though $\widehat{C}_d \in \operatorname{span} \mathcal{F}_d$, we no longer have that C_d is in the same function space due to the nonlinear behavior of the rectification and integration.

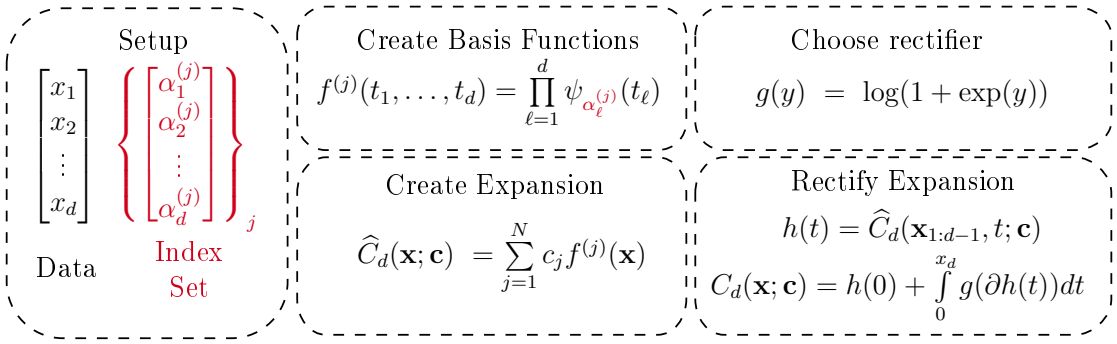

Figure 2-2: Diagram of a rectified expansion for the dth map component

Finally, on the note of approximation choice, the set of multiindex sets $\{\mathcal{A}_d\}_{d=1}^n$ (i.e., there is one multiindex set for each output of a transport map, $\{C_d\}$) is a difficult choice in practice. There is one coefficient c_j to estimate for each $\boldsymbol{\alpha}^{(j)}$, so there is a computational cost by introducing large multiindex sets. Multiple methods of choosing the multiindex set reflect and extend how one might choose a basis for one-dimensional function approximation, by choosing a maximum complexity p of the expansion (e.g., maximum degree of a polynomial) and truncating at that point, but this is not a well defined imposition for multiple dimensions. The simplest adaptation to this is choosing a set $\mathcal{A}_d^{tensor} = \{\boldsymbol{\alpha}^{(\ell)} | \max_i \alpha_i^{(\ell)} \leq p\}$, which is usually referred to as the "full tensor product expansion" as it just truncates each dimension at p, but forms

multivariate expansions with full complexity up to dp, where "full complexity" is the sum of each dimension's complexity. Further, the tensor product expansion induces a cardinality of the multiindex set as $|\mathcal{A}_d| = p^d$, which clearly becomes cost-prohibitive when estimating one coefficient for each member of the set for even a moderate selection of d. Instead of choosing to truncate each dimension independently, one can choose to truncate based on the *full complexity*. A total order p multiindex set would be given by $\mathcal{A}_d^{total} = \{\boldsymbol{\alpha}^{(\ell)} | \sum_{i=1}^d \alpha_i^{(\ell)} \leq p\}$. We can easily observe that $\mathcal{A}_d^{total} \subset \mathcal{A}_d^{tensor}$, but is strictly much smaller. Using a "stars and bars" approach from traditional combinatorics, one can verify that the number of coefficients corresponding to such a set is given by $\left|\mathcal{A}_d^{total}\right| = \sum_{i=0}^p \binom{d+i-1}{i} = \binom{d+p}{p}$, which can be shown to be bounded by p^d for reasonably small p, d. For example, $p = 5, d = 4$ gives that $|\mathcal{A}_d^{tensor}| = 625$ and $\left|\mathcal{A}_d^{total}\right| = 126$. This total order multiindex set selection style can be extended to work with other total order limiting styles, e.g., limiting by the criterion $\sum_\ell \exp(\alpha_\ell^{(j)}) \leq p$. Finally, we often choose a "separable" multiindex set, which can be defined as

$$\mathcal{A}_d^{sep} \subset \mathcal{A}_d \text{ where } \mathcal{A}_d^{sep} = \left\{\boldsymbol{\alpha}^{(j)} \in \mathcal{A}_d \mid \alpha_d^{(j)} = 0 \text{ if } (\exists \ell < d)\ \alpha_\ell^{(j)} \neq 0\right\},$$

intuited as "the separable multiindex set $\mathcal{A}_d^{sep}$ takes elements from $\mathcal{A}_d$ corresponding to basis functions that are either constant in x_d or constant in x_ℓ for every $\ell < d$".

While these are acceptable attempts at creating a set of multiindices with minimal cardinality that might be adequately expressive, we can also use the set $\mathcal{A}_d$ to describe *the relationship* that X_d has with $X_{1:d-1}$. Consider three random variables Y, X_1, X_2 such that X_1 and X_2 are conditionally independent given Y; we can observe that a monotone triangular map can be structured as such

$$S(Y, X_1, X_2) = \begin{bmatrix} S_Y(Y) \\ S_{X_1}(Y, X_1) \\ S_{X_2}(Y, X_2) \end{bmatrix} \Rightarrow \nabla S = \begin{bmatrix} \nabla_Y S_Y & 0 & 0 \\ \nabla_Y S_{X_1} & \nabla_{X_1} S_{X_1} & 0 \\ \nabla_Y S_{X_2} & 0 & \nabla_{X_2} S_{X_2} \end{bmatrix}$$

where we note that S_{X_2} is only a function of Y, X_2. The existence of such a S_{X_2} is easily provable. Suppose S is a proper KR rearrangement with third compo-

nent $S_{X_2}(Y, X_1, X_2)$; then, using a similar approach as algorithm 2, we know that $S_{X_2}(y, x_1, \cdot)$ pushes $p(X_2|Y = y, X_1 = x_1)$ to some distribution η_3. Since $X_2 \perp X_1|Y$ is given to us, we know S_{X_2} *must* be constant in X_1. Therefore, we expect valid subset of the multiindex set for the third output $\mathcal{A}_3^{ind} \subseteq \mathcal{A}_3$ to have the form $\mathcal{A}_3^{ind} = \{\boldsymbol{\alpha}^{(j)} \in \mathcal{A}_3 | \alpha_2^{(j)} = 0\}$. Using this imposes structure that we know a priori as opposed to hoping that the solution to the optimization problem posed in eq. (2.6) gives us parameters $\mathbf{c}$ reflecting the true independence structure when estimated numerically. Further, imposing this conditional independence a priori decreases the number of parameters we must estimate so, for problems dependent with many conditionally independent variables (i.e. sparse precision matrices), we can more tractably work with expansions of the form C_d due to a "small" cardinality $|\mathcal{A}_d|$. Several algorithms exist for "adapting" the multiindex set to take advantage of the structure of the problem, either via imposing conditional independence or via adding multiindices to ameliorate where the error in the function approximation is high [30]–[32]. There has been recent work suggesting algorithms specific to the triangular transport map framework described above [25], [33].

Chapter 3

Approximation Behavior of Transport Map Estimation

3.1 Approximation Bases

In fig. 2-2, we sketch out creating a transport map from an expansion of basis functions; in particular, each basis function $f^{(\ell)}$ is a tensor product of different one-dimensional basis functions drawn from an indexed set $\{\phi_\alpha\}_\alpha$. It is not actually necessary that $\{\phi_{\alpha,i,d}\}$, the univariate functions for approximation in the ith input for the dth map component, are drawn from the same set for all i or d; in fact, it is often advantageous to choose a specific basis in input i for some component C_d based on the dependence relationship of input d and input i of the data-generating distribution π. In previous works on using Knothe-Rosenblatt transport, substantial gains were seen choosing a special basis for approximation when $i = d$, particularly using "local approximators" such as wavelets expansions and RBFs [14], [25], [26].

In this text, we specifically consider at the softplus rectifier, $g(y) = \log(1 + \exp(y)) > 0$ for the "positive bijector", as illustrated in fig. 2-2. The softplus function is asymptotically linear as y gets large, and asymptotically zero as $-y$ gets large. For all y, we see $g(y) - y = \log(\exp(-y) + 1) \leq \exp(-y)$, meaning that the softplus function reverts to the identity at an exponential rate for $y > 0$. Similarly, $g(y) - 0 < \exp(y)$ for all y, meaning that g decays exponentially for decreasing y.

3.1.1 Hermite Polynomials

Probabilist Hermite polynomials, with α polynomial denoted He_α, can be defined as polynomials orthogonal under the standard normal probability measure. Going forward, Hermite polynomial refers to the probabilist species, and physicist Hermite polynomials are identified explicitly. The reason that one might be interested in using Hermite polynomials is due to a multitude of results from the field of polynomial chaos expansions (PCEs) [34], [35]. If our *data* $X \sim \pi$ follows a standard normal distribution, the transport problem is similar to trying to find the PCE estimating a transformation of X, to some extent. We generally assume that ν is not the normal measure; in fact, we often want the pushforward measure $S_\sharp\nu$ to be standard normal when training a transport map from samples. We choose such polynomials as they should be orthogonal on a measure that is absolutely continuous with respect to ν, and π (the assumed density of ν) is assumed in this work to have support on all of $\mathbb{R}^n$ unless otherwise stated.

Hermite polynomials can be defined constructively using the recurrence relation $\mathrm{He}_{\alpha+1}(x) = x\mathrm{He}_\alpha(x) - \mathrm{He}'_\alpha(x)$, where He_α is the Hermite polynomial with index $\alpha = 0, 1, 2\ldots$ (and $\mathrm{He}_0(x) :\equiv 1$). We note that the *physicist* Hermite polynomial at index α, denoted H_α, satisfies $\mathrm{H}_\alpha(x) = 2^{\alpha/2}\mathrm{He}_\alpha(x\sqrt{2})$. Further, these follow the Appell property, where $\mathrm{He}'_\alpha(x) = \alpha\mathrm{He}_{\alpha-1}(x)$. This can easily be seen by inductively assuming the property holds for α, then observing

$$\begin{aligned}\mathrm{He}'_{\alpha+1}(x) &= \frac{d}{dx}\left[x\mathrm{He}_\alpha(x) - \mathrm{He}'_\alpha(x)\right] = \mathrm{He}_\alpha(x) + x\mathrm{He}'_\alpha(x) - \alpha\mathrm{He}'_{\alpha-1}(x)\\ &= \mathrm{He}_\alpha(x) + \alpha x\mathrm{He}_{\alpha-1}(x) - \alpha\mathrm{He}'_{\alpha-1}(x) = \mathrm{He}_\alpha(x) + \alpha\mathrm{He}_\alpha(x)\\ &= (\alpha+1)\mathrm{He}_\alpha(x).\end{aligned}$$

The Appell property shows that $\mathrm{He}_{\alpha+1}$ has local maxima wherever He_α has a zero, which is vital when considering the bounds of where functions involving Hermite polynomials are maximized or minimized. Further, there are several results showing the behavior of the roots of He_α. Since it can be shown that the function He_α

has the parity of α (i.e., He_{2k} is an even function and He_{2k+1} is an odd function), the roots must be symmetric around the origin. We then refer to $r_\alpha = \max\{x \in \mathbb{R} \,|\, \mathrm{He}_\alpha(x) = 0\}$ as "the largest root" of He_α without sign ambiguity. A primitive result due to Laguerre's theorem is that $r_\alpha \geq \sqrt{\alpha - 1}$, but one can show much more exact behavior as $r_\alpha = 2\sqrt{\alpha} - i_1(9\alpha)^{-1/6} + o(\alpha^{-1/6})$ where $i_1 = 3.3721\ldots$ is the first root of a particular function given in [36], [37, (6.32.8)] (we note that the bounds in literature are given for *physicist* Hermite polynomials and note that the roots differ by a scaling of $\sqrt{2}$).

Figure 3-1 gives an example of what a well-behaved combination of the first five Hermite polynomials looks like. Note that ∂S matches $\partial \widehat{S}$ when they're both positive, but ∂S is marginally above zero when $\partial \widehat{S} < 0$. Further, notice that this example admits $\partial S \approx 0$ for $|x| > 2.5$, as $\partial \widehat{S} \ll 0$. When demonstrating the power of these rectified polynomials, we consider that this can be easily used for general inference, but we can see the possible problem if using this function with triangular transport. If we try to apply $\hat{S}(\cdot)^{-1}(z)$, we will get a numerically unstable inversion for z outside a given region.

3.1.2 Hermite Functions

We define the α Hermite function as $\psi_\alpha(x) = \mathrm{He}_\alpha(x)g(x)$ for $\alpha = 0, 1, \ldots$, where $g(x) = \sqrt{\mathcal{N}(x; 0, 1)}$ is a squared exponential kernel, i.e., the square root of the Gaussian probability density function. First, note that

$$\langle \psi_{\overline{\alpha}}, \psi_\alpha \rangle = \int \mathrm{He}_{\overline{\alpha}}(x)\mathrm{He}_\alpha(x)g^2(x)\,dx \equiv \delta_{\overline{\alpha}\alpha}.$$

Note that this Hermite function localizes the polynomial He_α around the origin and tapers off to zero. This can be adjusted to fit around any interval with any spread as desired, but the results shown below must be extended accordingly. These functions can be used to remedy the problem with tails demonstrated by fig. 3-1; suppose that we knew the one-dimensional expansion $\widehat{S}$ had bounded support, so $\widehat{S}(x) = 0$ for

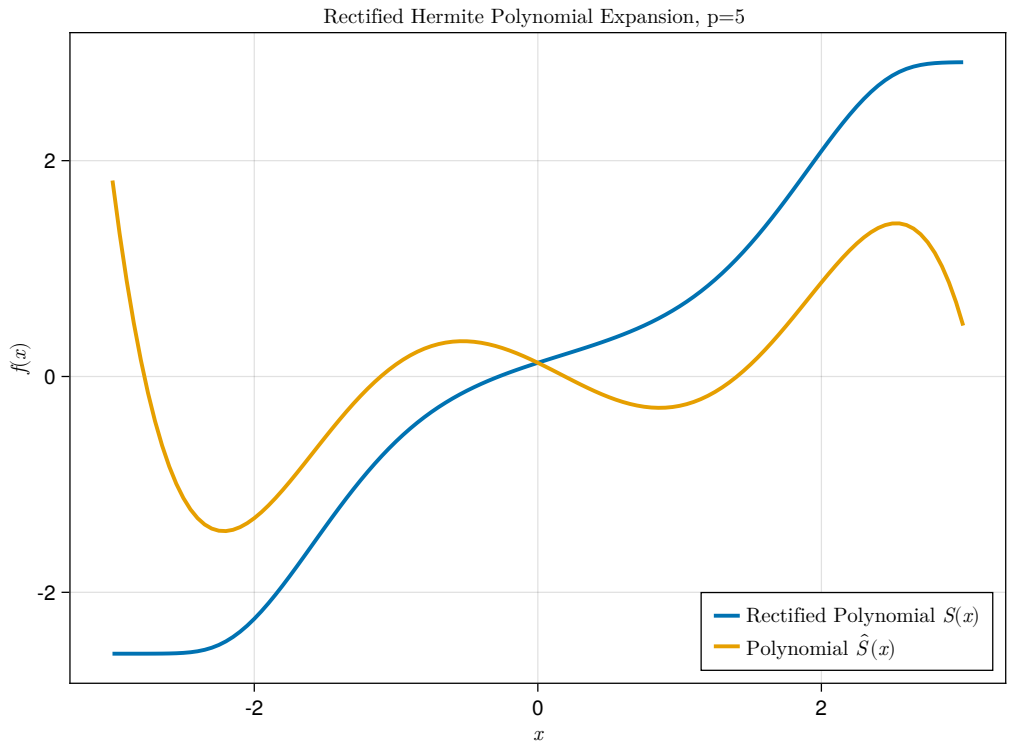

Figure 3-1: Example degree 5 polynomial constructed from Hermite basis

$x > a$. Then, for $x > a$,

$$S(x) = \widehat{S}(0) + \int_0^x g(\partial \widehat{S}(t))\, dt = \widehat{S}(0) + \int_0^a g(\partial \widehat{S}(t))\, dt + \int_a^x g(0)\, dt = \beta_1 + (x-a)\beta_2,$$

for some numbers $\beta_1, \beta_2 \in \mathbb{R}$. Therefore, if our basis functions satisfy $\psi_\alpha \approx 0$ for sufficiently large x and a fixed basis, we can approximate our rectified function as "linear in the tails". Hermite functions show these desirable properties, as seen in fig. 3-2.

For any ε, we can find x_α^ε such that $|\psi_\alpha(x)| < \varepsilon$ for all $x > x_\alpha^\varepsilon$, so rectified basis expansions of Hermite functions (when using softplus) ostensibly revert to linear past some x_α^ε defined by choice of degrees and coefficients. Recalling r_α as the largest root

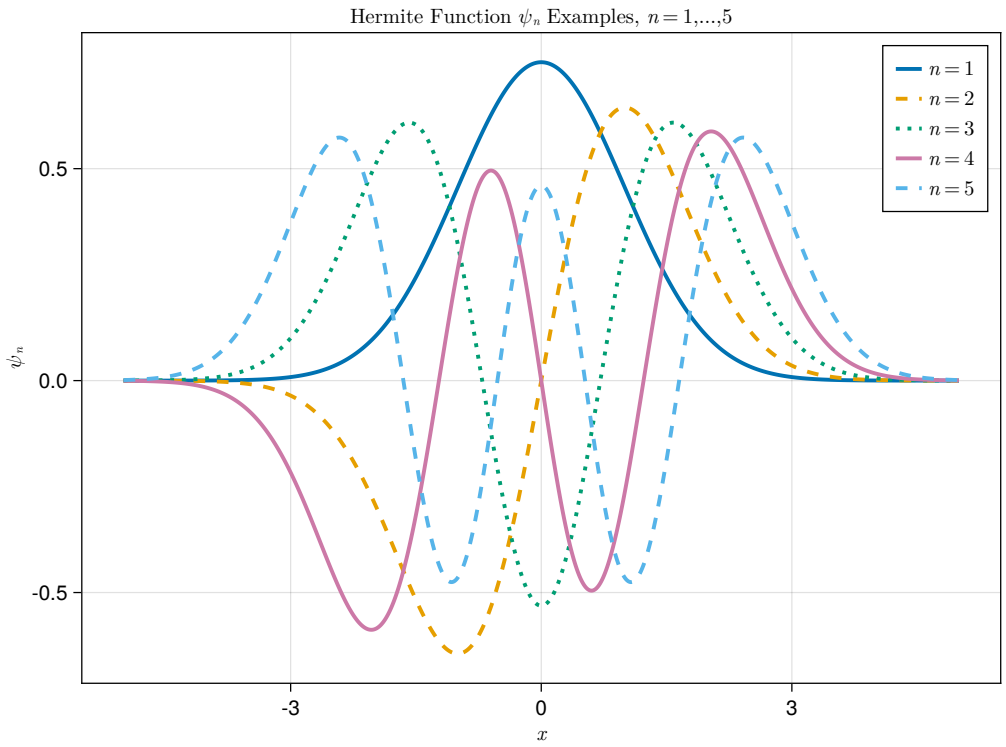

Figure 3-2: Examples of Hermite functions up to degree 5

of ψ_α, we see the maximizer must satisfy the equation

$$\begin{aligned} 0 &:= \frac{d}{dt}\left[\psi_\alpha(t)\right]_{t=x} \\ &= \mathrm{He}'_\alpha(x)\exp(-x^2/4) - \frac{x}{2}\mathrm{He}_\alpha(x)\exp(-x^2/4) \\ 0 &= x\mathrm{He}_\alpha(x) - 2\mathrm{He}'_\alpha(x) = (x\mathrm{He}_\alpha(x) - \mathrm{He}'_\alpha(x)) - \mathrm{He}'_\alpha(x) \\ &= \mathrm{He}_{\alpha+1}(x) - \alpha\mathrm{He}_{\alpha-1}(x). \end{aligned}$$

If we examine the function $h_\alpha(x) := \mathrm{He}_{\alpha+1}(x) - \alpha\mathrm{He}_{\alpha-1}(x)$, it is apparent that $h_\alpha(r_{\alpha+1}) < 0$ as $r_{\alpha+1} > r_{\alpha-1}$ and $\mathrm{He}_{\alpha-1}(x) > 0$ for $x > r_{\alpha-1}$. Further, we know that there exists some $y > r_{\alpha+1}$ such that $h_\alpha(y) > 0$, since h_α is a polynomial with a positive leading coefficient, indicating h_n has a root x^*_α satisfying $x^*_\alpha > r_{\alpha+1}$, which precisely reflects the location of a maximum of ψ_α.

Recall from section 3.1.1 that the last root of He_α behaves as $r_\alpha = \sqrt{\alpha} - \varepsilon_\alpha$ for $\varepsilon_\alpha > 0$ that decays at a rate of $\alpha^{-1/6}$ (and He_α grows as x^α for $x > r_\alpha$). The

idea that the maximizers of these functions propagate away from the origin can be illustrated by fig. 3-2 and, further, we do not see a substantial decay in the size of the corresponding maxima as the degree becomes larger. Qualitatively, this has to do with the movement of the maxima to the right at a slow rate similar to $\sqrt{\alpha}$, while simultaneously tuning up the polynomial degree. If the movement were faster, then the last maxima of ψ_α would decay faster in α; the slow decay of the furthest maxima seems, then, to be a *feature* of the method in some sense. Otherwise these could not possibly be appropriate "global approximators".

Recently, it has been proposed to use *compact* or *localized* Hermite functions as an alternative to the traditional operator [38], where we produce $\widetilde{\psi}_\alpha(x) = \mathrm{He}_\alpha(x)\widetilde{g}(x)$, where $\widetilde{g}$ is an approximation of $\exp(-x^2/4)$ but lives on some compact domain $[-x_0, x_0]$, which tells us that the maximizer of $\widetilde{\psi}_\alpha$ must live in $[-x_0, x_0]$ for any α. Choices for $\widetilde{g}$ include but are not limited to spline and B-spline approximations of $\exp(-x^2/4)$, which are easily represented but only piecewise differentiable—possibly problematic when considering, e.g., the approximation guarantees shown in section 3.2—or the Gaspari-Cohn function, coincidentally used in data assimilation literature for localization [10], [11], [13].

3.2 Estimating Expectations with Transport

While Monte Carlo estimates of expectations are appealing for their dimension-free convergence rates, oftentimes we might be able to impose more structure on the function we integrate over to require significantly fewer function evaluations if we pick educated inputs to the function. Suppose we have data that can be represented as $Y \sim \nu$ for some measure ν, and we create a transport map S such that $\nu = S^\sharp \mu$, where μ is the standard normal distribution. Then, for an arbitrary function f, we

may be interested in

$$\begin{aligned}\mathbb{E}_\nu[f(Y)] &= \int f(y)\, d(S^\sharp\mu)(y) = \int f(y)\, d(\mu \circ S)(y) = \int f(S^{-1}(x))\, d\mu(x) \\ &= \frac{1}{\sqrt{2\pi}} \int f(S^{-1}(x)) \exp(-x^2/2)\, dx = \frac{1}{\sqrt{\pi}} \int f(S^{-1}(\sqrt{2}z)) \exp(-z^2)\, dz.\end{aligned} \tag{3.1}$$

With no further assumptions on S besides the coupling and invertibility, we are motivated to use Gauss-Hermite quadrature to find such a quantity of interest. This is particularly interesting for polynomial f (i.e., moments of ν), or f with swiftly decaying Taylor series, due to well-established properties of Gauss quadrature: if we use a rule of order Q, then our rule will be exact in one dimension on polynomials up to degree $2Q-1$. Therefore, assuming an exact transport map S coupling ν and μ, we know that $S^{-1}(\mathbf{0})$ gives us the exact expected value of ν, for example. We can get the uncentered variance by evaluating $S^{-1}/2$ at every possible combination of ± 0.5 in all dimensions. This extends to insinuate that, for a general map, we need Q^n inversions of the full transport map. However, if we assume S to have a KR rearrangement structure, then we can use the structure of the transport map's evaluation at each quadrature point for an efficient rule. Suppose we have one-dimensional quadrature points $\{\zeta_q\}$ and weights $\{w_q\}$ for a marginal of an isotropic measure μ (for example, Gauss-Hermite quadrature with μ being a Gaussian measure), and a monotone triangular transport map $S : \mathbb{R}^n \to \mathbb{R}^n$ that pushes ν to measure μ. Then, we know from the above that we can estimate

$$\mathbb{E}_\nu[f] \approx \sum_{q_n=1}^{Q} \sum_{q_{n-1}=1}^{Q} \cdots \sum_{q_1=1}^{Q} w_{q_n} w_{q_{n-1}} \cdots w_{q_1} f\left(S^{-1}\left(\zeta_{q_1}, \zeta_{q_2}, \ldots, \zeta_{q_n}\right)\right).$$

The discerning reader will recall that

$$S^{-1}(\zeta_1, \ldots, \zeta_n) = \begin{bmatrix} C_1^{-1}(\zeta_1) \\ S_2(C_1^{-1}(\zeta_1), \cdot)^{-1}(\zeta_2, \ldots, \zeta_n) \end{bmatrix}.$$

In a tensor-product quadrature scheme, then, we only need to calculate $S_1^{-1}(\zeta_q)$ for

each one-dimensional point (instead of recalculating the same thing for all Q^n points). Then, we can use those Q cached values for inputs into S_2, which gives a clear way to tabulate or cache evaluations of inverses of each C_d at each dimension d to only employ $Q + Q^2 + \cdots + Q^d = Q(Q^n - 1)/(Q - 1)$ one-dimensional function inversions. Employing this scheme improves on general inversion of the KR rearrangement, as it uses the caching *across* quadrature points, where the usual one-dimensional inversion in these maps only employs the structure at each point in parallel. The full algorithm must then have nQ^n one-dimensional function inversions in our fully tensorized case, substantially larger for large n and small Q. This strategy can be used for further savings if the function we use for our expectation, f itself, has triangular structure. Further, there are more sophisticated sparse quadrature schemes for computing expectations in high dimensional space, but the fundamental structure of a triangular map offers significant benefit in evaluations since each component C_d needs not consider x_i for $d < i \leq n$.

An interesting question one can pose is, in practice, how much can we say about the error imposed by these "exact" quadrature rules when we only have an estimate of the exact transport map. Given ν, μ, consider a transport map S and the inverse map $T := S^{-1}$ such that $S_\sharp \nu = \mu$ and $T_\sharp \mu = \nu$. Suppose we estimate such a transport map with a function $\widehat{S}$ and associated inverse $\widehat{T} := \widehat{S}^{-1}$. Then, we will have some estimated pushforward and pullback distributions, $\widehat{\mu} := \widehat{S}_\sharp \nu$ and $\widehat{\nu} := \widehat{T}_\sharp \mu$. Note that, trivially, $\widehat{T}_\sharp \widehat{S}_\sharp \nu := \nu$, since $\widehat{T}$ is the *exact* inverse of the approximate map $\widehat{S}$, so we hope that $\widehat{S}_\sharp \nu = \widehat{\mu} \approx \mu$ and $\widehat{T}_\sharp \mu = \widehat{\nu} \approx \nu$ in some sense. These maps and such estimates are visualized in fig. 3-3, which shows what one might see when plotting the pushforward or pullback distribution's density under an estimate of a transport map (as opposed to the true measures). Intuitively, the problem of estimating a transport map from samples will produce a function that might be biased unless the true KR rearrangement lives in our approximation space, and the estimate could also have high variance [39].

Now suppose we have a square-integrable function f that is ℓ-Lipschitz. If we want $\mathbb{E}[f(Y)]$, we see

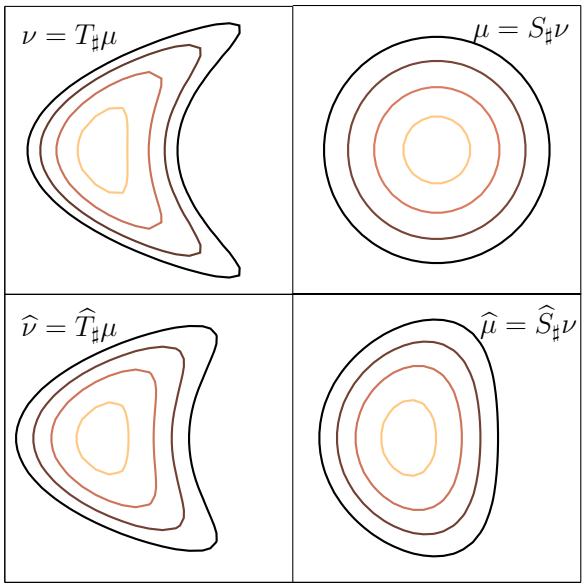

Figure 3-3: Diagram comparing true and estimated densities of pushforward measures

$$
\begin{aligned}
\|\mathbb{E}_\nu[f] - \mathbb{E}_{\widehat{\nu}}[f]\|^2 &= \left\| \int f \, d\nu - \int f \, d\widehat{\nu} \right\|^2 \\
&= \left\| \int f \circ T \, d\mu - \int f \circ \widehat{T} \, d\mu \right\|^2 \\
&\leq \int \left\| f \circ T - f \circ \widehat{T} \right\|^2 d\mu \int 1 \, d\mu \qquad \text{Cauchy-Schwarz} \\
&\leq \int \ell^2 \left\| T(x) - \widehat{T}(x) \right\|^2 d\mu(x) \\
&= \ell^2 \int \left\| y - \widehat{T}(S(y)) \right\|^2 d\nu(y)
\end{aligned}
$$

This can easily be extended to work for any distance in a Banach space, and it holds for all maps S and $\widehat{S}$. Now, specializing μ as the standard normal distribution and $S, \widehat{S}$ as KR rearrangements, we can use Proposition 1 of [25] to see

$$\|\mathbb{E}_\nu[f] - \mathbb{E}_{\widehat{\nu}}[f]\|^2 \leq 2\ell^2 \mathcal{D}_{KL}(\nu || \widehat{\nu}).$$

Proposition 3.2.1. *If $f \in L^2$ is ℓ-Lipschitz, S is the KR rearrangement with push-forward measure $S_\sharp \nu = \mu$ where $\mu := \mathcal{N}$, and there exists $\widehat{T}$ is a monotone triangular*

map that defines measure $\widehat{\nu} := \widehat{T}_\sharp \mu$*, then*

$$\|\mathbb{E}_\nu[f] - \mathbb{E}_{\widehat{\nu}}[f]\|^2 \leq 2\ell^2 \mathcal{D}_{KL}(\nu||\widehat{\nu}).$$

This has resounding implications for any approximation of the exact expectation $\mathbb{E}_\nu[f(Y)]$, since we can use, e.g., eq. (3.1) by substituting in $\widehat{S}^{-1}$. Then, if $f(x) = x$ (i.e., we want the expected value of $Y \sim \pi$), we perform one evaluation $\widehat{S}(\mathbf{0})$ to get an approximation with error bounded by $2\mathcal{D}_{KL}(\nu||\widehat{\nu})$. There is much to gain from improving such a bound—even simple functions such as $f(x) = x^2$ (which would give the variance) are not globally Lipschitz, and globally Lipschitz functions make a very narrow subset of all functions we may be interested in. It should be noted that proposition 3.2.1 can be applied to transport-powered simple Monte Carlo schemes as well as quadrature. However, being able to perform efficient quadrature rules can certainly beat the number of evaluations required to create a simple Monte Carlo estimate of the expectation for sufficiently regular f and small dimension. We can use basic approximation theory results from, e.g., [40], to find that, in one dimension, the approximate error bound is

$$\begin{aligned} E_Q[f] &\leq \frac{\| \left(f \circ S^{-1}\right)^{(2Q)} \|_\infty}{(2Q)!} \langle \mathrm{He}_Q, \mathrm{He}_Q \rangle = \frac{\| \left(f \circ S^{-1}\right)^{(2Q)} \|_\infty}{(2Q)!} \frac{Q!}{2^Q} \\ &\approx \sqrt{2} \| \left(f \circ S^{-1}\right)^{(2Q)} \|_\infty \left(8Q/e\right)^{-Q}, \end{aligned}$$

where the approximation comes from Stirling's formula. Then, since we scale up to dimension d, we say that the error grows decays with exponential rate Q/d, scaled by the regularity of the derivatives of $f \circ S^{-1}$. It is generally observed [40] that this regularity term makes the bound reasonably loose, but it is important to observe the regularity of S^{-1} and its derivatives not only causes numerical difficulties when evaluating an estimate of S as in section 3.1, but could also increase the error of the quadrature even if the estimate is exact.

3.3 Numerical Results

3.3.1 Hermite Functions

First, we look at the value of x_α^*, the largest maximizer of the α Hermite function ψ_α, and note that the value of x_α^* indeed grows very consistently with the result on r_α, the largest root of the α Hermite polynomial. Here, we use the form $\tilde{x}_\alpha^* = s_1\sqrt{\alpha} + s_2\alpha^{-1/6} + s_3$ and fit our coefficients $\{s_i\}$ via a least-squares regression on the data, with $\{s_i\} = \{1.27, -5.53, 3.34\}$. Then, fig. 3-4 shows that the maximizers indeed do empirically grow at a rate comparable to what [37] tells us about the roots of the α polynomial. Further, to showcase where the problems can occur, fig. 3-4 also plots the maxima $\psi_\alpha(x_\alpha^*)$ as a function of the maximizers and we do indeed see that, not only do the Hermite functions have maximizers that grow similar to $\sqrt{n}$, but we can actually use an approximation $\psi_\alpha(x_\alpha^*) \approx t_1 x^{t_2}$ where a least-squares regression gives $\{t_i\} = \{0.655, -0.153\}$. Since the maximizers' growth decays rapidly (i.e., the derivative of our approximation of x_α^* in α is a function of the form $1/p(\alpha)$ for polynomial p), the maxima of $\psi_\alpha(x_\alpha^*)$ seems to asymptotically go to zero. However, the Hermite functions propagate enough mass far from the origin for large α to make using these functions in practice extraordinarily problematic. For example, when $\alpha = 100$, it is easy to computationally estimate $x_\alpha^* \approx 13.77$ and $\psi_\alpha(x_\alpha^*) = 0.43$, meaning that, if we rectify and invert this, it is likely to have problems with outliers. This indicates the need for explicit adjustments when including function spaces spanned by high-degree Hermite functions, e.g., the "localized Hermite functions" presented above. The "exact" x_α^* in fig. 3-4 is estimated by numerically finding the maximum of the Hermite functions on a uniform one dimensional grid with resolution $h = 0.01$.

3.3.2 Transport Expectations

We experiment with the results on transport-based expectation estimation by creating a numerical example of working with $\mathbf{X} \in \mathbb{R}^n$ following a "multidimensional" banana

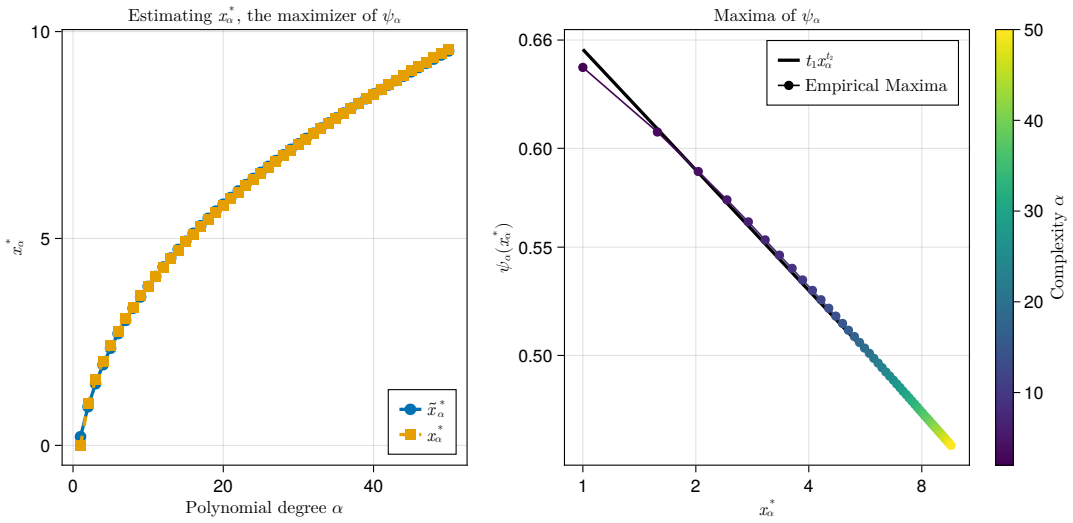

Figure 3-4: (Left): Creating estimates of the maximizer of ψ_α furthest from the origin, (right): Showing numerical maxima of ψ_α.

distribution, where

$$X_1 = Z_1, \quad X_d = Z_d + Z_{d-1}^2 \; \forall d = 2, \ldots, n, \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n).$$

This admits a simple analytic triangular transport map, and allows us to use simple multiindex sets for scalable estimation and evaluation of the transport map. In these experiments, we use 2e4 samples from the banana distribution to train a map, then compare different methods of estimating expectations to the pure Monte Carlo expectation estimate from using the samples from training the map as a reference value. Since the divergence of a pushforward distribution from the reference is directly related with the numerical estimate of a transport map via samples, these experiments directly can be related to proposition 3.2.1.

First, generating and evaluating an increasing number of points is shown in fig. 3-5. For a given number of quadrature points Q, we use Q^n evaluations of the entire map $\widehat{S}$ (as opposed to the tabulation strategy introduced), hence the exponential scaling. We compare the simple tensor-product Gaussian quadrature, a trapezoid rule (which assumes the support lies in $[-5, 5]$), a pure Monte Carlo scheme using generative sampling, and a reference Monte Carlo estimate from the samples used to

train the map. Further, we look at the mean $\mathbb{E}[\mathbf{X}]$ as well as the uncentered variance $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$, where the errors are measured using the RMSE and Förstner distance d_F [41], respectively. These are given by

$$\text{RMSE}(X, \bar{X}) = \frac{\|X - \bar{X}\|}{\sqrt{n}}, \qquad d_F(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{d=1}^{n} \log^2 \lambda_d(\mathbf{A}, \mathbf{B})},$$

where $\lambda_d(\mathbf{A}, \mathbf{B})$ is the dth generalized eigenvalue of symmetric positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$.

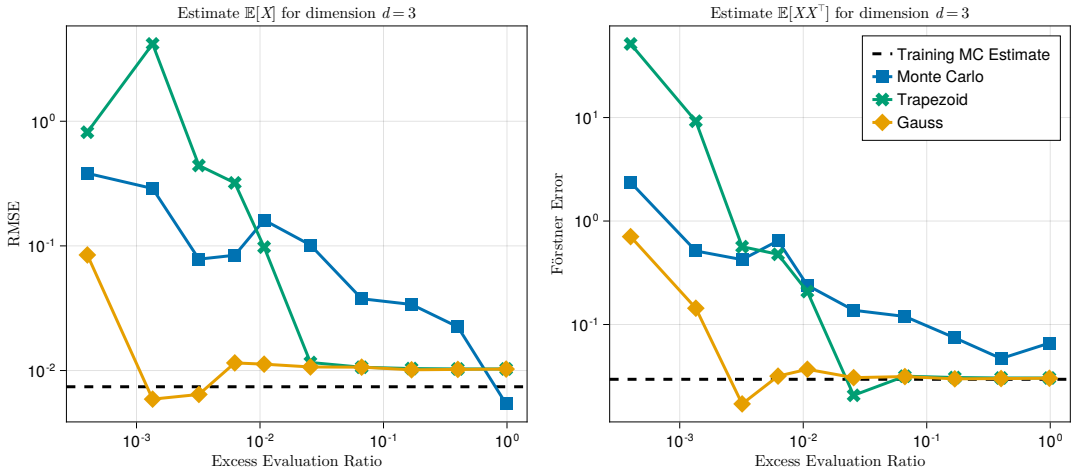

Figure 3-5: Evaluating the error for increasing evaluations; the horizontal axis measures the ratio of evaluations for the estimation method compared to the number of training samples for the map

We note that, for a fixed dimension, we can gain substantial savings in number of evaluations of f by choosing a smart scheme of integration when estimating $\mathbb{E}[f(\mathbf{X})]$ instead of using the samples we use to train the transport map $\widehat{S}$. This can be substantial if f is an expensive function. Figure 3-5 suggests we need only 1% of the number of evaluations to get a comparable estimate of $\mathbb{E}[f(\mathbf{X})]$ as pure Monte Carlo using the samples used to train the transport map. We observe, however, that we cannot substantially improve the error from this reference estimate which follows intuition (otherwise we would gain accuracy "for free"). In fact, we observe a slight bias from our reference error when estimating $\mathbb{E}[X]$, but this is marginal in comparison

to the overall error.

It is imperative to simultaneously examine how these estimates perform if we instead fix the number of *testing* evaluations for the estimates and scale the dimension of the random variable we are concerned with. When scaling the dimension and fixing the number of *training* samples, we note that the transport map induces a pullback distribution $\widehat{S}^\sharp\eta$ that has increasing divergence from π; this can be estimated using the exact value of the minimization problem expressed in eq. (2.7) since we can express the target distribution π in closed form here. We note that only the trapezoid rule seems to deviate strongly from the error of the reference estimate of the expectation from training samples. This is expected because, as the dimension n increases, the amount of mass of the Gaussian located outside $[-5, 5]$ grows exponentially with n. Additionally, the slight bias in dimensions eight and nine in fig. 3-6 is likely attributable to the inability to effectively employ the same number of transport map evaluations across all dimensions.

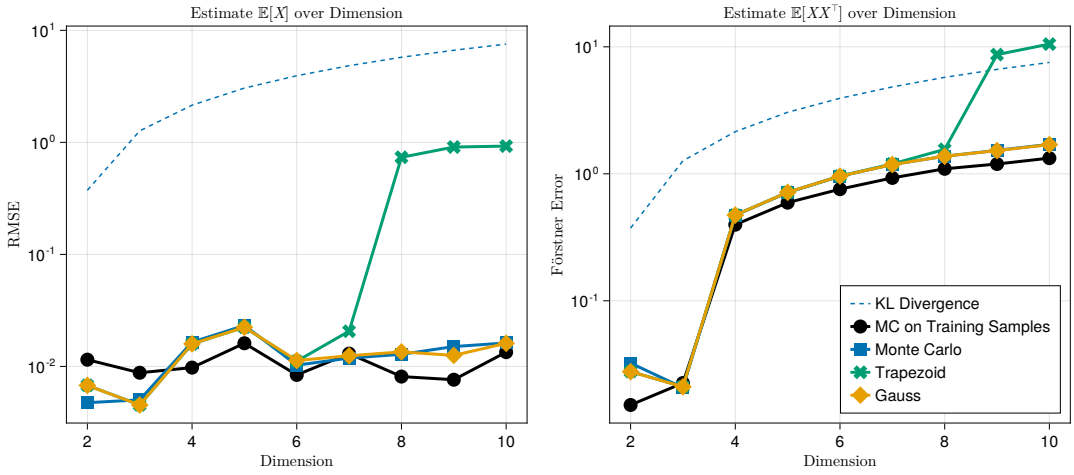

Figure 3-6: Evaluating the error for increasing dimension; the number of evaluations is fixed at $10^7 \pm 10\%$ as possible (dimensions 7,8,9 maximize Q such that the number of evaluations Q^n does not exceed 10^7)

Chapter 4

Ensemble Transport for State Space Models

In a sense, it is easy to see that the Bayesian data assimilation problem is a simple sequence of finding posteriors, which is well suited for the monotone triangular transport framework we have built up so far.

4.1 Ensemble Transport for Filtering

Intuitively, we recall from section 2.1 that the problem of interest in Bayesian filtering is posed as, given the conditional distribution $p(X_{k-1}|Y_{1:k-1})$ and new observation Y_k, we wish to estimate the new conditional distribution $p(X_k|Y_{1:k}) = p(X_k|Y_{1:k-1}, Y_k)$. We further recall from the EnKF in section 2.1.2 that we assume a collection or ensemble of samples $\{\mathbf{x}_{k-1}^{(j)}\} \sim p(X_{k-1}|Y_{1:k-1})$, from which it is easy to simulate samples of the distribution $p(X_k, Y_k|Y_{1:k-1})$. The problem of having samples of a joint distribution and wanting to create samples of a conditional is neatly solved by the Knothe-Rosenblatt transport framework, as outlined in section 2.2, so we create and train a map S_k for generating samples from $p(X_k|Y_{1:k-1}, Y_k)$. This framework for Bayesian filtering was introduced in [14], and can be summarized by the sequence of ensemble members and their distributions summarized in table 4.1. We note that, at a high level, the map composition $S_k(\mathbf{y}_k, \cdot)^{-1} \circ S_k$ can be thought of as the nonlinear

extension of applying a Kalman gain operator $\mathbf{K}_k$, and there are distinct parallels shown in [14].

Ensemble Member	$\widehat{\mathbf{x}}_k^{(j)}$	$\widehat{\mathbf{y}}_k^{(j)}$	$\mathbf{z}_k^{(j)}$	$\mathbf{x}_k^{(j)}$
Dimension	n	m	n	n
Definition	$F(\mathbf{x}_{k-1}^{(j)}) + \widehat{\xi}_k^{(j)}$	$H(\widehat{\mathbf{x}}_k^{(j)}) + \widehat{\gamma}_k^{(j)}$	$S_k(\widehat{\mathbf{y}}_k^{(j)}, \widehat{\mathbf{x}}_k^{(j)})$	$S_k(\mathbf{y}_k, \cdot)^{-1}(\mathbf{z}_k^{(j)})$
Distribution	$p(X_k \vert \mathbf{y}_{1:k-1})$	$p(Y_k \vert \widehat{\mathbf{x}}_k^{(j)})$	$\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$	$p(X_k \vert \mathbf{y}_{1:k})$

Table 4.1: Summary of distributions in Ensemble Transport Filtering

4.2 Ensemble Transport for Static Parameter Estimation

Note that in the above section, at timestep k, we create an estimate of a map component $S_k(Y, X)$ to infer parameter X_k. Importantly, we make no linear assumption on the relationship between $Y_k \in \mathbb{R}^m$ and $X_k \in \mathbb{R}^{n_x}$, nor do we make a linear assumption on $F(X_{k-1})$. Therefore, there could theoretically be elements of X_{k-1} that are related with one another in a highly nonlinear fashion. One very simple idea, then, is to adapt the method discussed by the above section to use for the parameterized state space problem given in section 2.1.3, as it is just a state that relates highly nonlinearly. Suppose we take the SSM given in eq. (2.3) parameterized by random variable W on $\mathbb{R}^{n_w}$. For a given timestep k, if we have joint samples $\{(\widehat{\mathbf{y}}_k^{(j)}, \mathbf{w}^{(j)}, \widehat{\mathbf{x}}_k^{(j)})\} \sim p(Y_k, W, X_k | Y_{1:k-1} = \mathbf{y}_{1:k-1})$, it is easy to see how to extend algorithm 2 toward sampling the posterior of interest $p(W, X_k | Y_{1:k} = \mathbf{y}_{1:k})$.

This algorithm ties back to state augmentation, a strategy employed in traditional data assimilation by practitioners. The problem with state augmentation for many EnKF-like methods is that it has little ability to capture nonlinear elements, and most parameters enter in extraordinarily nonlinearly.

Algorithm 3: Joint state and parameter ensemble transport filtering

input : Samples $\{(\widehat{\mathbf{x}}_{k-1}^{(j)}, \mathbf{w}^{(j)}\}_{j=1}^{J} \sim p(X_{k-1}, W|Y_{1:k-1})$, observation $\mathbf{y}_k$, SSM of form 2.3

output: $\{(\mathbf{x}_k^{(j)}, \mathbf{w}'^{(j)})\}_{j=1}^{J} \sim p\left(X_k, Z|Y_{1:k}\right)$

Sample $\widehat{\mathbf{x}}_k^{(j)} \sim p(X_k|X_{k-1} = \mathbf{x}_{k-1}^{(j)}, W = \mathbf{w}^{(j)})$;

Sample $\widehat{y}_k^{(j)} \sim p(Y_k|X = \widehat{\mathbf{x}}_k^{(j)}, W = \mathbf{w}^{(j)}, Y_{1:k-1})$;

Estimate $S_k : \mathbb{R}^{m+n_x+n_w} \to \mathbb{R}^{n_x+n_w}$ satisfying $S_{k\sharp}p(Y_k, W, X_k|Y_{1:k-1}) = \mathcal{N}$;

Create ensemble $(\mathbf{z}_w^{(j)}, \mathbf{z}_x^{(j)}) = S_k(\widehat{\mathbf{y}}_k^{(j)}, \mathbf{w}_k^{(j)}, \widehat{\mathbf{x}}_k^{(j)})$;

Generate samples $(\mathbf{w}'^{(j)}, \mathbf{x}^{(j)}) = S_k(\mathbf{y}_k, \cdot, \cdot)^{-1}(\mathbf{z}_w^{(j)}, \mathbf{z}_x^{(j)})$

Parameterized SSM Example Revisited

Recall eq. (2.4), a parameterized SSM with form

$$\begin{aligned} X_k &= \theta X_{k-1} + q\xi_k, \quad \xi_k \sim \mathcal{N}(0, \mathbf{I}) \\ Y_k &= X_k + r\gamma_k, \quad \gamma_k \sim \mathcal{N}(0, \mathbf{I}), \end{aligned}$$

In the method detailed by algorithm 3, we must find some kind of transport for the "augmented" state vector (W, X_k). Assuming that this transport is block triangular, we can think of this as creating a block transport map of the form

$$S_k(Y_k, W, X_k) = \begin{bmatrix} S_{W_k}(Y_k, W) \\ S_{X_k}(Y_k, W, X_k) \end{bmatrix},$$

where S_{W_k} transports (Y_k, W) jointly to the n_w dimensional reference distribution and S_{X_k} transports (Y_k, W, X_k) jointly to the n_x dimensional reference distribution, where our reference of choice will be the standard normal. While the form of component S_{W_k} is not readily available, the component S_{X_k} is more clear. By definition, $Y_k = X_k + r\gamma_k$, where $\gamma_k \sim \mathcal{N}(0, \mathbf{I})$, so the function $S_{X_k}(Y_k, W, X_k) = (X_k - Y_k)/r$ transports the joint to a standard normal distribution. This clearly satisfies the conditions of a KR rearrangement, as S_{X_k} is monotone in X_k. Then, assuming $X_k \in \mathbb{R}$, we recall the form of the function space for the estimation of S_{X_k} when using a rectified basis expansion,

$$\widehat{S}_{X_k}(Y_k, W, X_k) = \widehat{C}_{X_k}(Y_k, W, 0) + \int_0^{X_k} g\left(\widehat{C}_{X_k}(Y_k, W, t)\right) \, dt,$$

where f comes from a finite dimensional function space, e.g., the span of Hermite polynomials. Even assuming that our estimation $\widehat{S}_{X_k}$ found the indepence θ, q and X_k conditional on r, we still retain the conundrum of estimating a rational function of r with polynomials. If we were to use polynomials, we would see extremely poor behavior in approximating this transport, which can be intuitively understood by examining the Taylor series $1/x = \sum_{j=0}^{\infty}(1-x)^j$, converging when $|1-x| < 1$. Further, this mandates "nonseparable" multiindices in the multiindex sets, i.e., multiindices with nonzero entries in the last element (representing X_k), as the exact transport map has a term X_k/r, which is a product of a functions of r and X_k.

While it is not easy to find a transport $S_{W_k}(Y_k, W)$ that is both *monotone* and *triangular* in W, we can at least find a transport that might give a heuristic of the difficulty of approximating the triangular transport solution. We use the form of Y_k expressed above, then recall that we can in fact know the distribution $p(Y_k|W)$ in this system exactly due to the closed form expressions, so we see that, imposing the assumption that $X_0 \sim \mathcal{N}(0, s_0^2)$ for some fixed prior variance s_0^2

$$\begin{aligned}
Y_k &= X_k + r\gamma_k \\
&= (\theta X_{k-1} + q\xi_k) + r\gamma_k \\
&= (\theta(\theta(\ldots(\theta X_0 + q\xi_1)\ldots) + q\xi_{k-1}) + q\xi_k) + r\gamma_k \\
&= \theta^k X_0 + \sum_{j=1}^{k} \theta^{j-1} q\xi_j + r\gamma_k \\
&\sim \mathcal{N}\left(0, s_{Y_k}^2\right) \text{ where } s_{Y_k}^2 = \theta^{2k} s_0^2 + \sum_{j=1}^{k} \theta^{2(j-1)} q^2 + r^2,
\end{aligned} \tag{4.1}$$

therefore $S'_{W_k} = Y_k / s_{Y_k}$ is a valid transport to a standard normal. This has the stipulations that it is a real-valued function of all the parameters θ, q, r and not monotone in any of the parameters; thus it does not satisfy the necessary conditions for $S_{W_k} : \mathbb{R}^3 \to \mathbb{R}^3$, which lives in a larger dimension and is comprised of monotone components. Despite the fact that S'_{W_k} is not a triangular transport map we seek, it is likely the most intuitive possible transport component. We note that this map has

frequency modes that increase as k grows, meaning that the maps approximating it would grow in complexity as k grows. This is also a fairly complicated function of the parameter vector W, which faces the same problems as the transport shown for S_{X_k} but with added complexity since S'_{W_k} is a function of more variables and lives outside even the "rational" function class (due to the square root $\sqrt{s^2_{Y_k}}$). **It is vital to understand that this does not prove anything about the form of** S_{W_k}, but rather gives an intuition as to why S_{W_k} might be difficult to approximate well when using a parameterization of $\widehat{S}_{W_k}$ that does not accurately represent the function space that S_{W_k} lives in.

4.3 Numerical Results

4.3.1 Ensemble Transport Filtering

First, we benchmark the ensemble transport filtering (EnTF) algorithm error metrics against traditional EnKF and the observation-space Ensemble Transform Kalman Filter (ETKF) algorithms [42]. A standard benchmark problem for this is to perform the filter on the Lorenz-63 system, which can be described as

$$\begin{aligned}\frac{dx_1}{dt} &= \sigma(x_2 - x_1)\\ \frac{dx_2}{dt} &= x_1(\rho - x_3) - x_2\\ \frac{dx_3}{dt} &= x_1 x_2 - \beta x_3,\end{aligned}$$

where we describe the forecast operator $F(\mathbf{x}; \sigma, \rho, \beta, \Delta t)$ as the solution of the Lorenz system at time Δt with initial condition $\mathbf{x}$; note that the system is autonomous, so the choice of start time is arbitrary. In this scenario, we pick the standard choice of parameters $\rho = 28$, $\sigma = 10$, and $\beta = 8/3$ as well as the step size $\Delta t = 0.001$. We use random initial condition $\mathbf{X}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and for the forecast operator F, we use two Euler integration steps to progress from t to $t + \Delta t$, where our true solution $\mathbf{x}(t)$ is solved using the Runge-Kutta scheme proposed by [43]. We observe

at every 100 forecast steps (i.e., we assimilate at every $100\Delta t = 0.1$). We initialize our transport filter with a valid ensemble of particles from our Lorenz manifold by assimilating the first 100 observations of the system with the standard EnKF to get an appropriately distributed ensemble. Finally, we then compare the performance of a quadratic complexity Ensemble Transport filter (EnTF) to the EnKF and the ETKF (using the same starting ensemble for each one) from step 101 onward. Finally, we employ ensemble inflation, a standard tool in data assimilation for artificially increasing variance of an ensemble which may be too small to sufficiently capture the spread of the ensemble. Given an inflation coefficient τ, we perform inflation on an ensemble $\{\mathbf{a}_j\}$ with ensemble mean $\overline{\mathbf{a}}$ by creating a new ensemble $\{\widehat{\mathbf{a}}_j\}$ such that

$$\widehat{\mathbf{a}}_j = (\mathbf{a}_j - \overline{\mathbf{a}})\tau + \overline{\mathbf{a}}.$$

The ensemble that we inflate could either be the *forecast* ensemble, $\{\widehat{\mathbf{x}}_k^{(j)}\}$ or the *analysis* ensemble $\{\mathbf{x}_k^{(j)}\}$. Since we require sufficient variance in the ensemble to adequately capture the behavior of the Lorenz problem, we choose to inflate the forecast ensemble, which generally leads to $\widehat{S}_k$ (the numerical approximation to the transport map S_k) having a more stable inverse when performing the analysis step to obtain $\mathbf{x}_k^{(j)}$; τ is chosen to be 0.05 for ensembles with $J \geq 100$ and 0.2 for smaller ensembles. We gauge our error via time-averaged Root Mean Squared Error (RMSE) and time-averaged Continuous Ranked Probability Score (CRPS), which have the forms

$$\begin{aligned}
\text{RMSE}(\{\mathbf{x}_k^{(j)}\}_{j,k}, \{\mathbf{x}_k\}_k) &= \mathbb{E}_k\left[\frac{\|\mathbf{x}_k - \overline{\mathbf{x}}_k\|}{\sqrt{D}}\right] = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\sqrt{D}}\left\|\mathbf{x}_k - \frac{1}{J}\sum_{j=1}^{J}\mathbf{x}_k^{(j)}\right\| \\
\text{CRPS}(\{\mathbf{x}_k^{(j)}\}_{j,k}, \{\mathbf{x}_k\}_k) &= \mathbb{E}_k\left[0.5\mathbb{E}_{j,\ell}[\|\mathbf{x}_k^{(j)} - \mathbf{x}_k^{(\ell)}\|] - \mathbb{E}_j[\|\mathbf{x}_k^{(j)} - \mathbf{x}_k\|]\right]
\end{aligned}$$

The boxplots shown in fig. 4-1 show how the error scales in these RMSE and CRPS metrics across several trials of performing these filtering schemes for exponentially increasing ensemble sizes. We note that the transport scheme has a much higher spread of errors compared to the standard assimilation schemes, generally with a right

skew. However, the median error for each given step seems to perform comparable to the reference methods. The major difference here from more remarkable results on filtering using transport [14], [26] is the choice of parameterization of the map. Even though the transport maps included here can have complex parameterizations with high-degree polynomials, the high performance of a filter on Lorenz is largely dependent on its ability to capture highly localized phenomena. Thus, the functions chosen in these other results reflecting the ability to capture details in localized regions (oftentimes radial basis functions or wavelets) are able to perform local approximation very well, whereas the Hermite functions used here are empirically observed to be somewhat robust approximators of more global behavior.

These results should be qualified by the stability of the filtering regime. In the low-rank ensemble case (e.g., 30 or 100 ensemble members), it is often the case that the function approximation problem is not robust to outliers; this leads to an unstable inverse when applying $\widehat{S}_X(\mathbf{y}_k, \cdot)^{-1}$. This is indeed due to the behavior of a multivariate expansion $\widehat{C}_d$ in some dimension j having strictly and strongly negative slope for such an outlier, which is how the discussion of integrated Hermite functions as possibly poor monotone function approximators is extremely relevant to our numerical applications of the filtering regime. Taking measures such as inflation (to ensure the transport map can learn on a patch of the joint distribution) can help, as well as further measures not taken here such as regularization on the coefficients in the components $\{C_d\}$. However, even then, the inversion step did entirely fail numerically several times when experimenting for fig. 4-1, though it was a rare event with a probability of less than $2.5\text{e} - 5$ for any given assimilation step (i.e., for every 40,000 map inversions, there was less than one numerical failure on average).

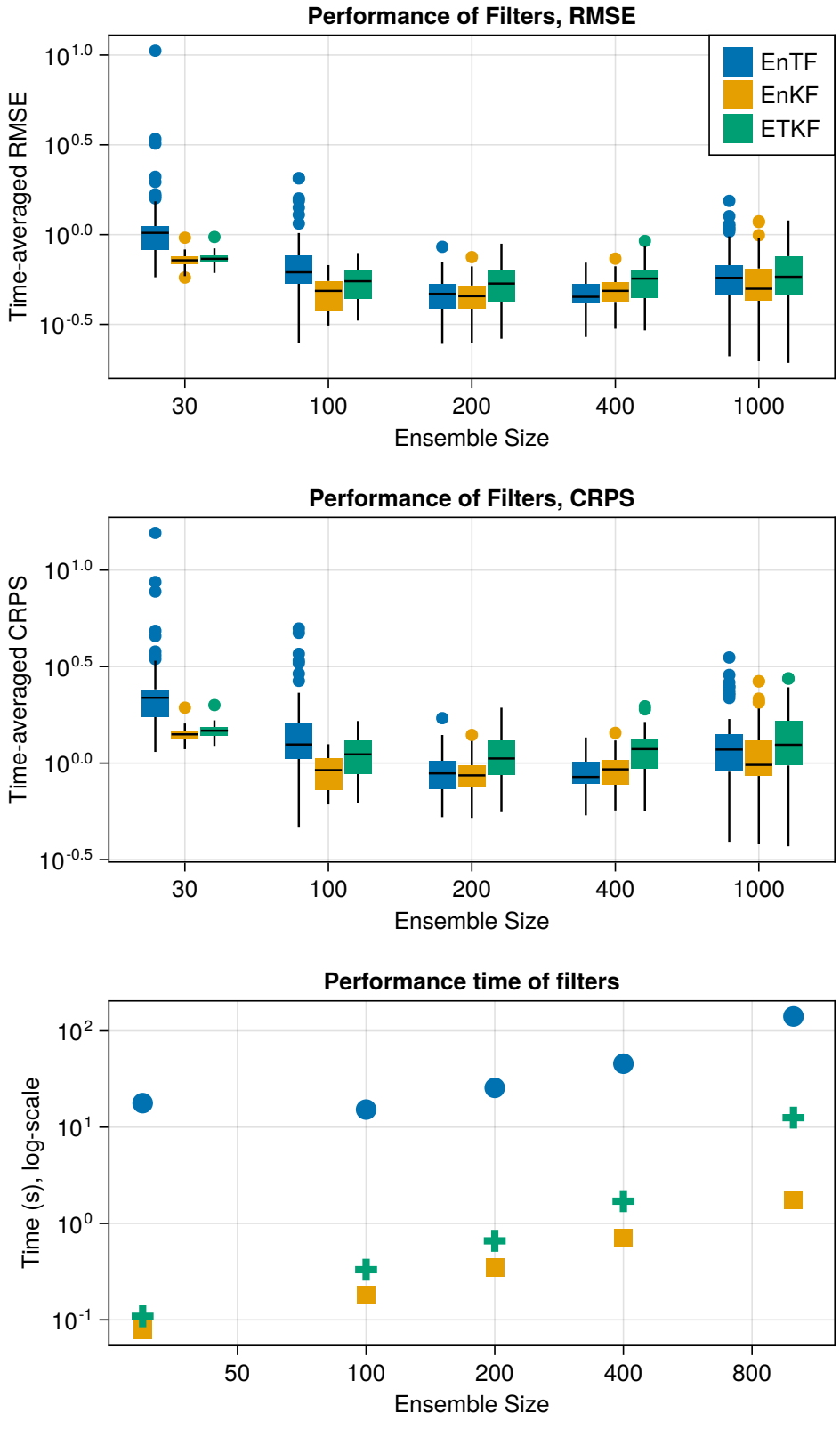

Figure 4-1: Performance of EnTF (Transport filter) versus EnKF and ETKF

4.3.2 Augmented State Ensemble Transport Filtering

Now we evaluate the performance of the parameter-state joint inference algorithm using measure transport. We set priors on our parameters $X_0 \sim \mathcal{N}(0,1)$, $\theta \sim \text{Beta}(7,3)$, $q^2, r^2 \sim \Gamma^{-1}(4,5)$, and $q = |\sqrt{q^2}|$, $r = |\sqrt{r^2}|$, noting that these priors enforce the bounds $\theta \in [0,1]$, $q, r > 0$. Then, $\mathbb{E}[\theta] = 0.7, \mathbb{E}[q] = \mathbb{E}[r] = \sqrt{5}\Gamma(3/2)/\Gamma(4) \approx 1.23854$; the noise parameters q, r are then quite large relative to the expectation of the state X_k, so we expect a rather crude approximation of both the state and the parameters. Further, this invites instability to appear in such a noise dominated problem.

To first evaluate the performance of the transport map on our problem, we set up a one-step Bayesian inference problem by creating a transport map S_1 which satisfies $S_1(Y_1, \theta, q, r, X_1) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_4)$. From our analytical solution of S_{X_1}, we know the form of one multiindex set $\mathcal{A}_{X_1}$ (though we do not know the exact set) and use a total order multiindex set for the nonzero components; we use total order multiindex sets for our parameters $\mathbf{w}$. In these experiments, we choose a map that has total order two in each map component to ensure sufficient functional complexity in the parameters. For a one-dimensional distribution, we can evaluate the quality of the pushforward via the Kolmogorov-Smirnov statistic D_J on J samples, the maximum discrepancy between the empirical CDF of the pushforward $S_{X_1\sharp}\pi$ and the reference density η. This can be formulated as

$$D_J = \sup_z |F_J(z) - F(z)|, \quad F_N(z) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{1}_{<z}\left(S_{X_1}(y_1^{(j)}, \mathbf{w}^{(j)}, x_1^{(j)})\right),$$

where F is the CDF of η, in this case the standard normal distribution. This statistic is a simple and effective distance between data and a reference distribution.

Despite the intricacies of estimating a rational function, we can still find S_{X_1} with reasonable confidence as seen in fig. 4-2, which evaluates the average Kolmogorov-Smirnov statistic D_J over 100 trials for increasing training ensemble size J with one standard deviation band (and a fixed 2e4 testing samples). The plotted training error, which comes from eq. (2.7), is partially omitted as it can be negative, and it is observed to generally increase due to better resolution as J increases. At $J = 1000$,

we see that there are diminishing returns and thus consider that a sufficiently large ensemble size.

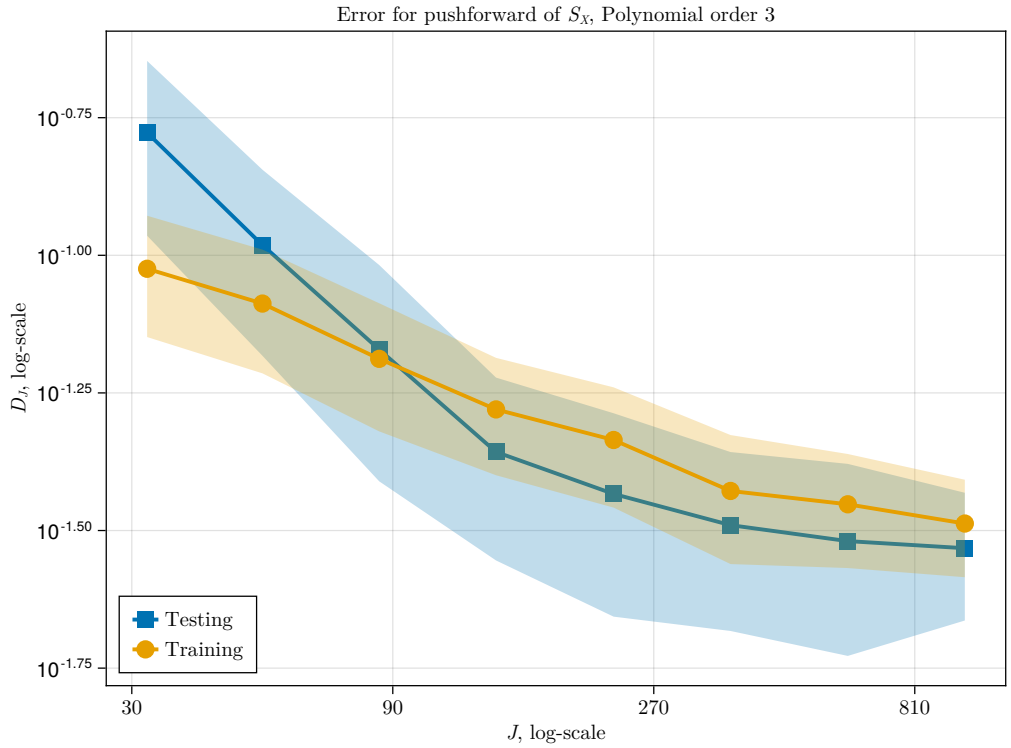

Figure 4-2: Calculating D_J statistic on testing and training datasets over ensemble size J in 100 trials

While the performance of S_{X_1} (the "lower block" of our desired map S_1) seems to exceeds expectations, this was also expected to be the easier component to train. In fig. 4-3, we see on the left that, left to its own devices, the map will not produce a pushforward that looks anything remotely like a standard normal. Note that Hermite polynomials, as remarked in section 3.1, are used in PCE for data that originate from the normal distribution (due to being orthogonal under the Gaussian measure), and Hermite functions are orthogonal under L^2; however, θ, q, r all originate from compact domains so, in order to ensure that they match the approximation basis of Hermite functions, we must transform them to live on the L^2 space. Therefore, we arbitrarily choose bijections $\mathbf{f}(\theta, q, r) = [f_1(\theta), f_2(q), f_3(r)]$ that achieve this desired property, then train the transport map $S_{W,1}(\theta, q, r) = S^{\mathbf{f}}_{W,1}(\mathbf{f}(\theta, q, r))$—we choose the logistic sigmoid function $f_1(\theta) := 1/(1 + \exp(-\theta))$, then $f_2 \equiv f_3 \equiv \log$. Note that

all of these functions are analytically invertible and only depend on one variable, so $S_{W,1}^{-1} = \mathbf{f}^{-1} \circ \left(S_{W,1}^{\mathbf{f}}\right)^{-1}$ is easily computable. Then, we see on the right in fig. 4-3 that the distribution qualitatively achieves significantly more normality.

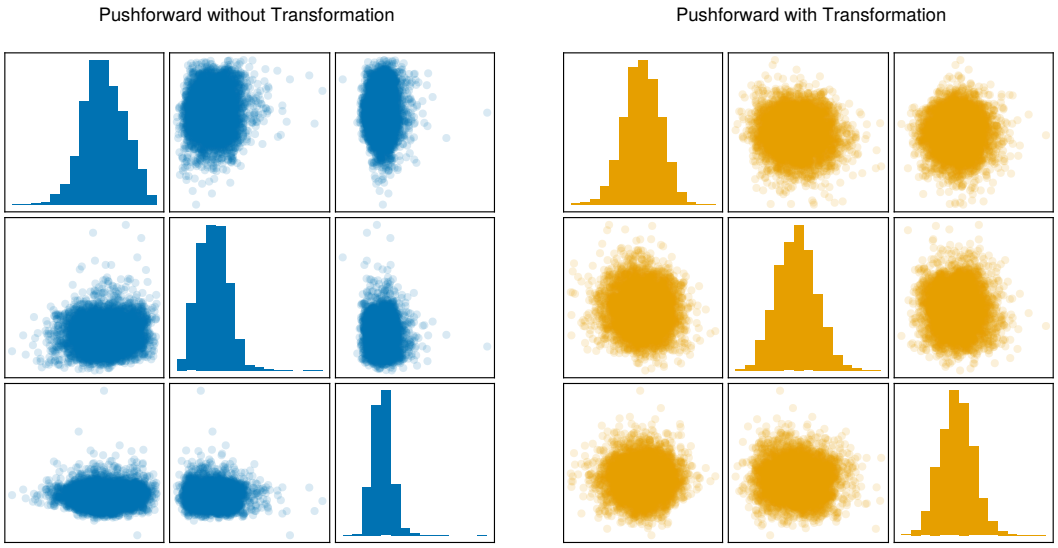

Figure 4-3: Pushforward of S_W

Therefore, for each independent step, we know that there is a series of transformations to reasonably estimate a transport map S_k for sufficiently large ensemble sizes. However, in practice, this is still insufficient to ensure a stable estimation of the state and parameters sequentially; at a given step $\widehat{k}$, there is insufficient variance in the ensemble to capture the behavior of $Y_{\widehat{k}}$, leading to an unstable calculation of $S_{\widehat{k}}(Y_{\widehat{k}}, \cdot)^{-1}(Z_{\widehat{k}})$. The solution to this from traditional data assimilation literature is, similar to inflation in the filtering case, artificially increasing the variance of the ensemble. In literature on parameter estimation [16], artificial dynamics of the parameters are often introduced. By perturbing the parameters at each step a small amount, we can ensure that the variance of the ensemble of parameter values increases to adequately capture significantly more behavior. We cannot simply add Gaussian noise to constrained parameters, as they might escape the bounds we put in place, therefore we add Gaussian noise in the "transformed space" induced by our sequence of bijections $\mathbf{f}$ introduced above. However, this might induce bias on our posterior distribution $p(W|Y_{1:k})$, therefore we strive to keep this noise to a minimum. Since all

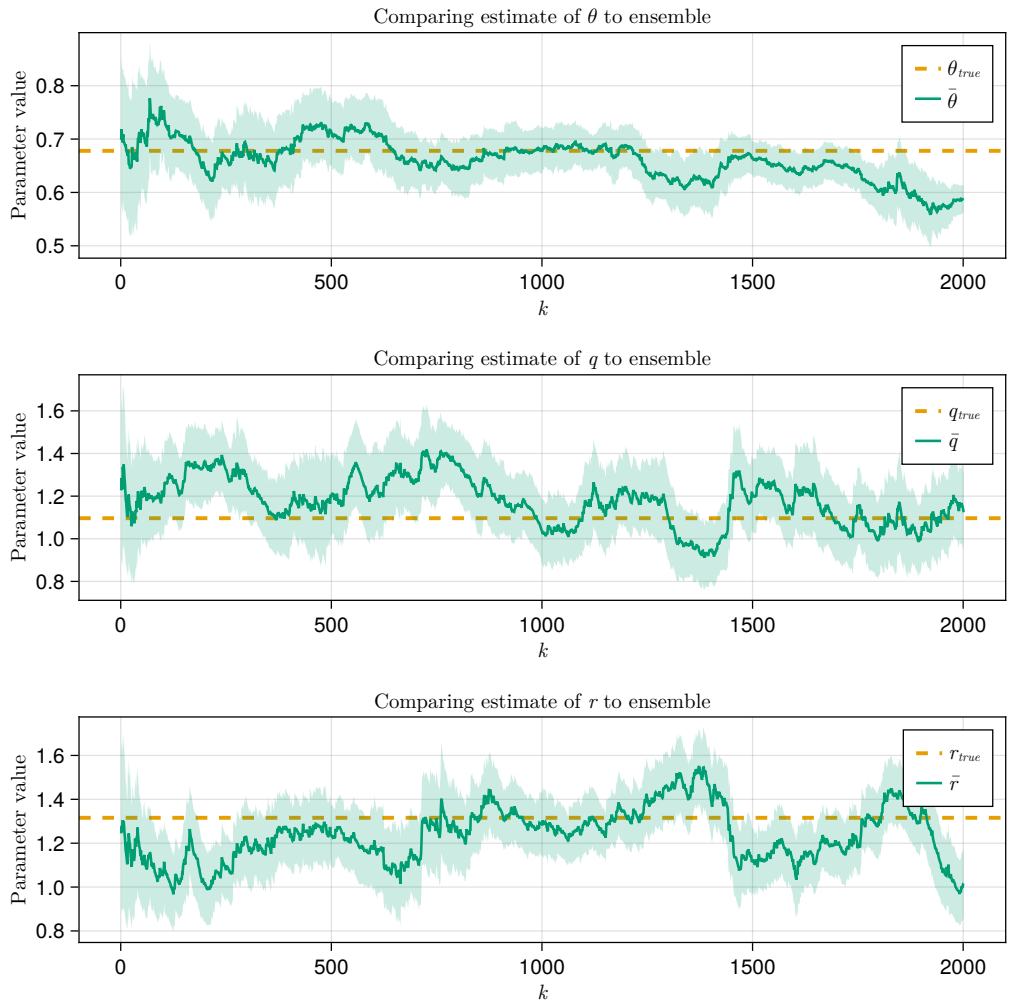

Figure 4-4: Sequential estimation of parameters in linear system

of the transformations $\mathbf{f}$ are smooth, we can expect such a bias from the transformed perturbation negligibly for sufficiently small noise, which we observe to be the case.

We tune the transformed noise to have variance 10^{-4}, then examine the empirical distribution of the ensemble of $w^{(k)}$ parameters at each timestep k in fig. 4-4. This includes the ensemble average as well as its standard deviation. Table 4.2 shows that the ensemble average of each parameter tends to be in j deviations of the true value as frequently as the central limit theorem might suggest.

As we know the prior $p(W)$, we can use an exact likelihood from eq. (2.5) for the purpose of sampling from the posterior $\pi(W|Y_{1:K}) \propto p(Y_{1:K}|W)p(W)$, then compare the analytical posterior on W to the approximate posterior $\widehat{\pi}(W|Y_{1:K})$ we get from

	$\mathbb{P}\left(\left\|\mathbb{E}[w^{(k)}] - w_{true}\right\| < j\sigma_{w^{(k)}}\right)$		
	$j=1$	$j=2$	$j=3$
Parameter w			
$w=\theta$	0.82059	0.96402	0.98551
$w=q$	0.78661	0.99550	1.00000
$w=r$	0.73663	0.99100	1.00000

Table 4.2: Frequency of capturing true parameter values for example trajectory with $K = 2000$ and $N = 1000$

algorithm 3. Since the parameters are estimated in this algorithm as fixed, we aggregate samples from the last $\tau = 200$ timesteps to ensure that we have a sufficiently expressive ensemble of samples, and they avoid being biased by the last assimilation step. Using an adaptive random walk Metropolis algorithm to sample from the analytical posterior π, we can compare empirical distributions as seen in fig. 4-5 for the performance of our parameter estimation scheme.

We note that working in the space transformed by $\mathbf{f}$ preserves desirable numerical properties to keep the parameters consistently in the correct domain, and that the maximum a posteriori (MAP) estimate, i.e. the mode of the posterior, seems to perform well in this scenario. In these scenarios, while uncertainty on the parameter of interest helps, a practitioner would likely be more interested in a point estimate for the purpose of usage in the SSM. Further, we see a consistent behavior of our posterior distributions consistently overestimating the spread of our parameters relative to samples of the analytical posterior. We speculate this diffuse estimate has two origins: the artificial measures we take to compute these quantities, i.e., the error of the transport map as well as the noise we add at each step. The second possible source for this error would be the natural and philosophical consequences of online parameter estimation ignoring all prior observations when assimilating data at step k, discussed in section 2.1.3.

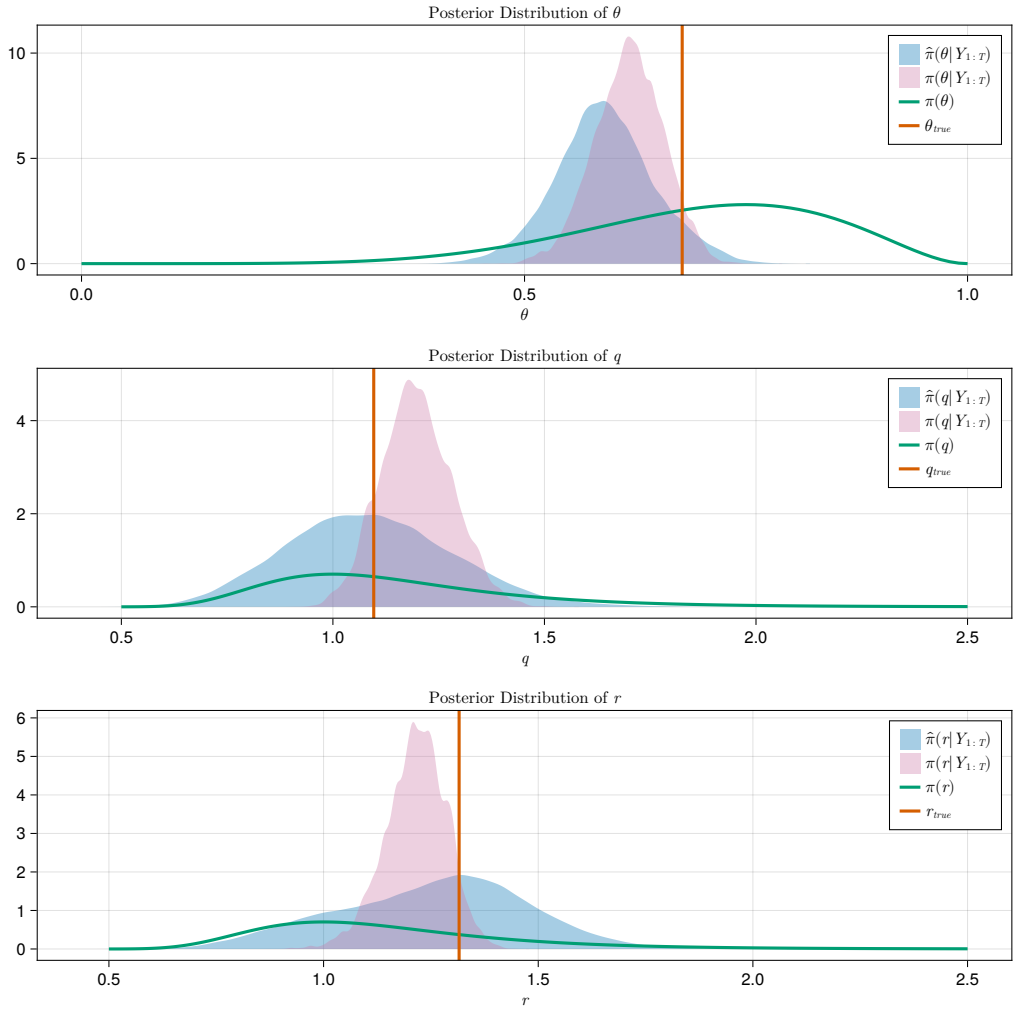

Figure 4-5: Comparing density estimates of 4e4 samples from the analytical posterior $\pi(W|Y_{1:K})$ to 2e5 samples of the transport-based approximate posterior $\widehat{\pi}(W|Y_{1:K})$

Chapter 5

Conclusions and Future Work

5.1 Conclusions

While it may not be straightforward, the data assimilation problem for parameterized SSMs can be amenable to solutions using transport maps, similar to filtering and smoothing. However, the problem comes down to efficiently and effectively parameterizing the functions that perform measure transport in a way that is aware of the problem at hand. In the numerical examples, we show that blindly using a high complexity map with an ineffective basis may not lead to results, but using heuristics that are problem-dependent can give reasonable performance. Further, in this setting, we still must make sacrifices in formulating a solution to a problem that approximates our problem at hand. We also demonstrate that transport can create convincing estimations of a quantity of interest $\mathbb{E}_\nu[f(X)]$ with a relatively miniscule number of evaluations of f; however, this can be intractable in higher dimensions unless a smarter high-dimensional quadrature scheme is used, or it could be impractical as the cost of quadrature could exceed the cost of employing a Monte Carlo estimate using the samples a map is trained on.

5.2 Future Work

There is substantial work to be done both for explicit and intelligent choice of map parameterizations; implementing and experimenting with compact Hermite functions is one clear future step. Further, improving on RBF placement and implementations from prior works is a topic of interest, particularly for `MParT`. On the side of parameter estimations, it would be interesting to compare the algorithm proposed here to the SMC^2 algorithm for checking *efficiency* instead of just *accuracy* as above.

5.2.1 Chaotic State Space Models

There is been significant focus here on estimating parameters in simple SSMs, largely due to well understood characteristics of such a model and where the parameters would lie, as well as the ability to express certain phenomena in closed-form (e.g., the posterior after an observation or an exact transport map). However, in practice, we will seldom see parameters in such models, and instead will record observations of phenomena that are much more nonlinear. In particular, since numerical weather prediction was the motivation for this method, it would be nice to apply this to online methods for chaotic dynamics. However, in practice, this has shown to be difficult to use for several reasons. For example, eq. (4.1) shows how nonlinear we expect a transport map to be for perhaps the most trivial parameterized SSM. Therefore, we expect to need a map that has high frequency modes in its multiindex sets. However, due to the chaotic dynamics, if we don't sample from the exact posterior, we observe the dynamics amplifying errors in the trajectory by several orders of magnitude.

Bibliography

[1] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," en, *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960.

[2] S. Särkkä, *Bayesian filtering and smoothing* (Institute of Mathematical Statistics textbooks 3). Cambridge, U.K. ; New York: Cambridge University Press, 2013, OCLC: ocn840462877.

[3] J. Liu and M. West, "Combined Parameter and State Estimation in Simulation-Based Filtering," in *Sequential Monte Carlo Methods in Practice*, A. Doucet, N. Freitas, and N. Gordon, Eds., New York, NY: Springer New York, 2001, pp. 197–223.

[4] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On Particle Methods for Parameter Estimation in State-Space Models," *Statistical Science*, vol. 30, no. 3, Aug. 2015.

[5] C. Andrieu and G. O. Roberts, "The pseudo-marginal approach for efficient Monte Carlo computations," *The Annals of Statistics*, vol. 37, no. 2, Apr. 2009.

[6] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov Chain Monte Carlo Methods," en, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 72, no. 3, pp. 269–342, Jun. 2010.

[7] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, *SMC^2: An efficient algorithm for sequential analysis of state-space models*, arXiv:1101.1528 [stat], Jan. 2012.

[8] G. Evensen, "Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics," en, *Journal of Geophysical Research*, vol. 99, no. C5, p. 10 143, 1994.

[9] G. Burgers, P. Jan Van Leeuwen, and G. Evensen, "Analysis Scheme in the Ensemble Kalman Filter," en, *Monthly Weather Review*, vol. 126, no. 6, pp. 1719–1724, Jun. 1998.

[10] P. L. Houtekamer and F. Zhang, "Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation," en, *Monthly Weather Review*, vol. 144, no. 12, pp. 4489–4532, Dec. 2016.

[11] G. Evensen, *Data assimilation: the ensemble Kalman filter*. Berlin ; New York: Springer, 2007, OCLC: ocm74270320.

[12] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation: A Mathematical Introduction* (Texts in Applied Mathematics), en. Cham: Springer International Publishing, 2015, vol. 62.

[13] G. Evensen, F. C. Vossepoel, and P. J. van Leeuwen, *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem*, English. Springer Nature, 2022.

[14] A. Spantini, R. Baptista, and Y. Marzouk, *Coupling techniques for nonlinear ensemble filtering*, arXiv:1907.00389 [stat], Apr. 2022.

[15] A. Y. Sun, A. Morris, and S. Mohanty, "Comparison of deterministic ensemble Kalman filters for assimilating hydrogeological data," en, *Advances in Water Resources*, vol. 32, no. 2, pp. 280–292, Feb. 2009.

[16] H. Moradkhani, S. Sorooshian, H. V. Gupta, and P. R. Houser, "Dual state–parameter estimation of hydrological models using ensemble Kalman filter," en, *Advances in Water Resources*, vol. 28, no. 2, pp. 135–147, Feb. 2005.

[17] Y. Chen, D. Sanz-Alonso, and R. Willett, *Auto-differentiable Ensemble Kalman Filters*, arXiv:2107.07687 [cs, stat], Jul. 2021.

[18] Y. Zhao and T. Cui, *Tensor-based Methods for Sequential State and Parameter Estimation in State Space Models*, arXiv:2301.09891 [cs, math, stat], Jan. 2023.

[19] F. Daum, J. Huang, and A. Noushin, "New Theory and Numerical Results for Gromov's Method for Stochastic Particle Flow Filters," in *2018 21st International Conference on Information Fusion (FUSION)*, Cambridge: IEEE, Jul. 2018, pp. 108–115.

[20] Q. Liu and D. Wang, "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm," in *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc., 2016.

[21] K. B. Petersen, M. S. Pedersen, *et al.*, *The matrix cookbook*. Technical University of Denmark, Nov. 2012, vol. 7.

[22] C. Villani, *Optimal Transport* (Grundlehren der mathematischen Wissenschaften). Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 338.

[23] M. D. Parno and Y. M. Marzouk, "Transport Map Accelerated Markov Chain Monte Carlo," en, *SIAM/ASA Journal on Uncertainty Quantification*, vol. 6, no. 2, pp. 645–682, Jan. 2018.

[24] Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini, "An introduction to sampling via measure transport," in *Handbook of Uncertainty Quantification*, arXiv:1602.05023 [math, stat], Springer, 2016, pp. 1–41.

[25] R. Baptista, Y. Marzouk, and O. Zahm, *On the representation and learning of monotone triangular transport maps*, arXiv:2009.10303 [cs, math, stat] type: article, Jul. 2022.

[26] M. Ramgraber, R. Baptista, D. McLaughlin, and Y. Marzouk, *Ensemble transport smoothing – Part 2: Nonlinear updates*, arXiv:2210.17435 [stat], Oct. 2022.

[27] N. Kovachki, R. Baptista, B. Hosseini, and Y. Marzouk, *Conditional Sampling With Monotone GANs*, arXiv:2006.06755 [cs, stat], Feb. 2021.

[28] J. O. Ramsay, "Estimating Smooth Monotone Functions," en, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 60, no. 2, pp. 365–375, Jul. 1998.

[29] D. Bigoni, A. Spantini, and Y. Marzouk, "Adaptive construction of measure transports for bayesian inference," in *NIPS*, *Advances in Approximate Bayesian Inference*, Jan. 2016.

[30] M. K. Stoyanov, "Hierarchy-Direction Selective Approach for Locally Adaptive Sparse Grids," en, Oak Ridge National Laboratory, Tech. Rep. ORNL/TM-2013/384, 1097490, Sep. 2013, ORNL/TM–2013/384, 1 097 490.

[31] X. Ma and N. Zabaras, "An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations," en, *Journal of Computational Physics*, vol. 228, no. 8, pp. 3084–3113, May 2009.

[32] M. Griebel, "Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences," en, *Computing*, vol. 61, no. 2, pp. 151–179, Jun. 1998.

[33] M. Ramgraber, *Adaptive localization in nonlinear ensemble transport filtering*, Norway, May 2023.

[34] N. Wiener, "The Homogeneous Chaos," *American Journal of Mathematics*, vol. 60, no. 4, p. 897, Oct. 1938.

[35] D. Xiu and G. E. Karniadakis, "The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations," en, *SIAM Journal on Scientific Computing*, vol. 24, no. 2, pp. 619–644, Jan. 2002.

[36] É. Fouvry, E. Kowalski, and P. Michel, "Counting sheaves using spherical codes," arXiv, Tech. Rep. arXiv:1210.0851, Aug. 2013, arXiv:1210.0851 [math] type: article.

[37] G. Szegő, *Orthogonal polynomials* (Colloquium publications - American Mathematical Society v. 23), 4th ed. Providence: American Mathematical Society, Jan. 1939.

[38] M. Ramgraber, *Private Communication on Map Parameterizations*, Mar. 2023.

[39] S. Wang and Y. Marzouk, "On minimax density estimation via measure transport," *arXiv preprint arXiv:2207.10231*, 2022.

[40] J. Stoer, R. Bulirsch, R. Bartels, W. Gautschi, C. Witzgall, and J. Stoer, *Introduction to numerical analysis* (Texts in applied mathematics 12), eng, 3. ed., Softcover version of original hardcover edition 2002. New York, NY: Springer, Jan. 2002.

[41] W. Förstner and B. Moonen, "A Metric for Covariance Matrices," en, in *Geodesy-The Challenge of the 3rd Millennium*, E. W. Grafarend, F. W. Krumm, and V. S. Schwarze, Eds., Berlin, Heidelberg: Springer, 2003, pp. 299–309.

[42] C. H. Bishop, B. J. Etherton, and S. J. Majumdar, "Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects," en, *Monthly Weather Review*, vol. 129, no. 3, pp. 420–436, Mar. 2001.

[43] C. Tsitouras, "Runge–Kutta pairs of order 5(4) satisfying only the first column simplifying assumption," en, *Computers & Mathematics with Applications*, vol. 62, no. 2, pp. 770–775, Jul. 2011.