ARTICLE **Communicated by Adam Charles**

A Distributed Framework for the Construction of Transport Maps

Diego A. Mesa*
diego.mesa@vanderbilt.edu
Department of Electrical Engineering and Computer Science and Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37205, U.S.A.

Justin Tantiongloc*
jctanti@gmail.com
Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Marcela Mendoza*
mendoza.marcelap@gmail.com
Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Sanggyun Kim
sgkim@stanford.edu
Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94304, U.S.A.

Todd P. Coleman
tpcoleman@ucsd.edu
Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, U.S.A.

The need to reason about uncertainty in large, complex, and multimodal data sets has become increasingly common across modern scientific environments. The ability to transform samples from one distribution P to another distribution Q enables the solution to many problems in machine learning (e.g., Bayesian inference, generative modeling) and has been actively pursued from theoretical, computational, and application perspectives across the fields of information theory, computer science, and biology. Performing such transformations in general still leads to computational difficulties, especially in high dimensions. Here, we consider the problem of computing such "measure transport maps" with efficient

*Diego Mesa, Justin Tantiongloc, and Marcela Mendoza contributed equally to this work.

Neural Computation **31, 613–652** (2019)
doi:10.1162/neco_a_01172

and parallelizable methods. Under the mild assumptions that P need not be known but can be sampled from and that the density of Q is known up to a proportionality constant, and that Q is log-concave, we provide in this work a convex optimization problem pertaining to relative entropy minimization. We show how an empirical minimization formulation and polynomial chaos map parameterization can allow for learning a transport map between P and Q with distributed and scalable methods. We also leverage findings from nonequilibrium thermodynamics to represent the transport map as a composition of simpler maps, each of which is learned sequentially with a transport cost regularized version of the aforementioned problem formulation. We provide examples of our framework within the context of Bayesian inference for the Boston housing data set and generative modeling for handwritten digit images from the MNIST data set.

1 Introduction

While scientific problems of interest continue to grow in size and complexity, managing uncertainty is increasingly paramount. As a result, the development and use of theoretical and numerical methods to reason in the face of uncertainty, in a manner that can accommodate large data sets, has been the focus of sustained research efforts in statistics, machine learning, information theory, and computer science. The ability to construct a mapping that transforms samples from one distribution P to another distribution Q enables the solution to many problems in machine learning.

One such problem is Bayesian inference, (Bernardo & Smith, 2001; Gelman et al., 2014; Sivia & Skilling, 2006), where a latent signal of interest is observed through noisy observations. Fully characterizing the posterior distribution is in general notoriously challenging due to the need to calculate the normalization constant pertaining to the posterior density. Traditionally, point estimation procedures are used, which obviate the need for this calculation, despite their inability to quantify uncertainty. Generating samples from the posterior distribution enables approximation of any conditional expectation, but this is typically performed with Markov chain Monte Carlo (MCMC) methods (Andrieu, De Freitas, Doucet, & Jordan, 2003; Geman & Geman, 1984; Gilks, 2005; Hastings, 1970; Liu, 2008) despite two drawbacks: (1) the convergence rates and mixing times of the Markov chain are generally unknown, thus leading to practical shortcomings like "sample burn-in" periods, and (2) the samples generated are necessarily correlated, lowering effective sample sizes and propagating errors throughout estimates (Robert & Casella, 2004). If we let P be the prior distribution and Q the posterior distribution for Bayesian inference, then an algorithm that can transform independent samples from P to Q, without knowledge of the normalization constant in the density of Q, enables calculation of any conditional expectation with fast convergence.

As another example, generative modeling problems entail observing a large data set with samples from an unknown distribution P (in high dimensions) and attempting to learn a representation or model so that new independent samples from P can be generated. Emerging approaches to generative modeling rely on the use of deep neural networks and include variational autoencoders (Kingma & Welling, 2013), generative adversarial networks (Goodfellow et al., 2014) and their derivatives (Li, Swersky, & Zemel, 2015), and autoregressive neural networks (Larochelle & Murray, 2011). These models have led to impressive results in a number of applications, but their tractability and theory are still not fully developed. If P can be transformed into a known and well-structured distribution Q (e.g., a multivariate standard gaussian), then the inverse of the transformation can be used to transform new independent samples from Q into new samples from P.

While these issues relate to the functional attractiveness of the ability to characterize and sample from nontrivial distributions, there is also the issue of computational efficiency. There continues to be an ongoing upward trend of the availability of distributed and hardware-accelerated computational resources. As such, it would be especially valuable to develop solutions to these problems that are not only satisfactory in a functional sense but are also capable of taking advantage of the ever-increasing scalability of parallelized computational capability.

1.1 Main Contribution. The main contribution of this work is to extend our previous results on finding transport maps to provide a more general transport-based push-forward theorem for pushing independent samples from a distribution P to independent samples from a distribution Q. Moreover, we show how given only independent samples from P, knowledge of Q up to a normalization constant, and under the traditionally mild assumption of the log-concavity of Q, it can be carried out in a distributed and scalable manner, leveraging the technique of alternating direction method of multipliers (ADMM) (Boyd, Parikh, Chu, Peleato, & Eckstein, 2011). We also leverage variational principles from nonequilibrium thermodynamics (Jordan, Kinderlehrer, & Otto, 1998a) to represent a transport map as an aggregate composition of simpler maps, each of which minimizes a relative entropy along with a transport-cost-based regularization term. Each map can be constructed with a complementary, ADMM-based formulation, resulting in the construction of a measure transport map smoothly and sequentially with applicability in high-dimensional settings.

Expanding on previous work on the real-world applicability of these general-purpose algorithms, we showcase the implementation of a Bayesian Lasso-based analysis of the Boston Housing data set (Harrison & Rubinfeld, 1978) and a high-dimensional example of using transport maps for generative modeling for the MNIST handwritten digits data set (LeCun, Bottou, Bengio, & Haffner, 1998).

1.2 Previous Work. A methodology for finding transport maps based on ideas from optimal transport within the context of Bayesian inference was first proposed in El Moselhy and Marzouk (2012) and expanded on in conjunction with more traditional MCMC-based sampling schemes in Marzouk, Moselhy, Parno, and Spantini (2016), Parno and Marzouk (2014), Parno, Moselhy, and Marzouk (2016), and Spantini, Bigoni, and Marzouk (2016).

Our previous work used ideas from optimal transport theory to generalize the posterior matching scheme, a mutual-information maximizing scheme for feedback signaling of a message point in arbitrary dimension (Ma & Coleman, 2011, 2014; Tantiongloc et al., 2017). Building on this, we considered a relative entropy minimization formulation, as compared to what was developed in El Moselhy and Marzouk (2012), and showed that for the class of log-concave distributions, this is a convex problem (Kim, Ma, Mesa, & Coleman, 2013). We also previously described a distributed framework (Mesa, Kim, & Coleman, 2015) that we expand on here.

In the more traditional optimal transportation literature, convex optimization has been used to varying success in specialized cases (Papadakis, Peyré, & Oudet, 2014), as well as gradient-based optimization methods (Benamou, Carlier, Cuturi, Nenna, & Peyré, 2015; Benamou, Carlier, & Laborde, 2015; Rezende & Mohamed, 2015). The use of stochastic optimization techniques in optimal transport is also of current interest (Genevay, Cuturi, Peyré, & Bach, 2016). In contrast, our work in this article presents a specific distributed framework where extensions to stochastic updating have been previously developed in a general case. Incorporating them into this framework remains to be explored.

In addition, there is much recent interest in the efficient and robust calculation of Wasserstein barycenters (center of mass) across partial empirical distributions calculated over batches of samples (Claici, Chien, & Solomon, 2018; Cuturi & Doucet, 2014). Wasserstein barycenters have also been applied to Bayesian inference (Srivastava, Li, & Dunson, 2015). While related, our work focuses instead on calculating the full empirical distribution through various efficient parameterizations that we discussed.

Building on much of this, there is growing interest in specific applications of these transport problems in various areas (Arjovsky, Chintala, & Bottou, 2017; Tolstikhin, Bousquet, Gelly, & Schoelkopf, 2017). These derived transport problems are proving to be a fruitful alternative approach and are the subject of intense research. The framework we present is general purpose and could benefit many of the derived transport problems.

Excellent introductory and references to the field can be found in Santambrogio (2015) and Villani (2008).

The rest of this article is organized as follows. In section 2, we provide some necessary definitions and background information; in section 3, we describe the distributed general push-forward framework and provide several details on its construction and use; in section 4, we formulate a specialized version of the objective specifically tailored for sequential composition;

in section 5, we discuss applications and examples of our framework; and we provide concluding remarks in section 6.

2 Preliminaries

In this section we make some preliminary definitions and provide background information for the rest of this article.

2.1 Definitions and Assumptions. Assume the space for sampling is given by $\mathsf{W} \subset \mathbb{R}^D$, a convex subset of D-dimensional Euclidean space. Define the space of all probability measures on W (endowed with the Borel sigma-algebra) as $\mathsf{P}(\mathsf{W})$. If $P \in \mathsf{P}(\mathsf{W})$ admits a density with respect to the Lebesgue measure, we denote it as p.

Assumption 1. We assume that $P, Q \in \mathsf{P}(\mathsf{W})$ admit densities p, q with respect to the Lebesgue measure.

This work is fundamentally concerned with trying to find an appropriate push-forward between two probability measures, P and Q:

Definition 1 *(push-forward). Given $P, Q \in \mathsf{P}(\mathsf{W})$ we say that a map $S : \mathsf{W} \to \mathsf{W}$ pushes forward P to Q (denoted as $S_{\#}P = Q$) if a random variable X with distribution P results in $Y \triangleq S(X)$ having distribution Q.*

Of interest to us is the class of invertible and "smooth" push-forwards:

Definition 2 *(diffeomorphism). A mapping S is a diffeomorphism on W if it is invertible, and both S and S^{-1} are differentiable. Let $\mathcal{D}$ be the space of all diffeomorphisms on W.*

A subclass of these are those that are "orientation preserving":

Definition 3 *(monotonic diffeomorphism). A mapping $S \in \mathcal{D}$ is orientation preserving, or monotonic, if its Jacobian is positive-definite:*

$$J_S(u) \succeq 0, \quad \forall u \in \mathsf{W}.$$

Let $\mathcal{D}_+ \subset \mathcal{D}$ be the set of all monotonic diffeomorphisms on W.

The Jacobian $J_S(u)$ can be thought of as how the map "warps" space to facilitate the desired mapping. Any monotonic diffeomorphism necessarily satisfies the following Jacobian equation:

Lemma 1 *(monotonic Jacobian equation). Let $P, Q \in \mathsf{P}(\mathsf{W})$, and assume they have densities p and q. Any map $S \in \mathcal{D}_+$ for which $S\#P = Q$ satisfies the following Jacobian equation:*

$$p(u) = q(S(u))\, det(J_S(u)) \quad \forall u \in \mathsf{W}. \tag{2.1}$$

We now concern ourselves with two different notions of "distance" between probability measures.

Definition 4 *(KL divergence). Let $P, Q \in \mathsf{P}(\mathsf{W})$, and assume they have densities p and q. The Kullback-Leibler (KL) divergence, or relative entropy, between P and Q is given by*

$$D(P\|Q) = \mathbb{E}_P\left[log\,\frac{p(X)}{q(X)}\right].$$

The KL divergence is nonnegative and is zero if and only if $p(u) = q(u)$ for all u.

Definition 5 *(Wasserstein distance). For $P, Q \in \mathsf{P}(\mathsf{W})$ with densities p and q, the Wasserstein distance of order two between P and Q can be described as*

$$d(P, Q)^2 \triangleq inf\left\{\mathbb{E}_{P_{X,Y}}[\|X - Y\|^2] : X \sim P, Y \sim Q\right\}. \tag{2.2}$$

The following theorem will be useful throughout:

Theorem 2.7 *(Brenier, 1987; Villani, 2003). Under assumption 1, $d(P, Q)$ can be equivalently expressed as*

$$d(P, Q)^2 \triangleq inf\left\{\mathbb{E}_P[\|X - S(X)\|^2] : S_{\#}P = Q\right\}, \tag{2.3}$$

and there is a unique minimizer S^*, which satisfies $S^* \in \mathcal{D}_+$.

Note that this implies the following corollary:

Corollary 1. *For any P, Q satisfying assumption 1, there exists a $S \in \mathcal{D}_+$ for which $S_{\#}P = Q$, or equivalently, for which equation 2.1 holds.*

3 KL Divergence-Based Push-Forward

In this section, we present the distributed push-forward framework that relies on our previously published relative entropy-based formulation of the measure transport problem and discuss several issues related to its construction.

3.1 General Push-Forward. According to lemma 1, a monotonic diffeomorphism pushing P to Q will necessarily satisfy the Jacobian equation 2.1. Note that although we think of a map S as pushing from P *to* Q, we have written equation 2.1 so that p appears by itself on the left-hand side, while S is being acted on by q on the right-hand side. This notation is suggestive of the following interpretation: If we think of the destination density q as an anchor point, then for any arbitrary mapping $S \in \mathcal{D}_+$, we can describe an induced density for $\tilde{p}(u; S)$ according to equation 2.1 as

$$\tilde{p}(u; S) = q(S(u))\det(J_S(u)) \quad \text{for all } u \in \mathsf{W}. \tag{3.1}$$

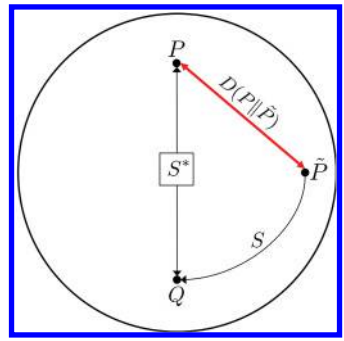

Figure 1: General push-forward. Probability measures P, $\tilde{P}(\cdot; S)$, and Q are represented as points in $\mathsf{P}(\mathsf{W})$. When Q is assumed to be constant, an arbitrary map $S \in \mathcal{D}_+$ can be thought of as inducing a distribution $\tilde{P}(\cdot; S)$. Thus, S pushes $\tilde{P}(\cdot; S)$ to Q (the solid black line labeled S in the figure). The problem of interest is to then find the S that minimizes the distance between the true P, and $\tilde{P}(\cdot; S)$. The optimal map S^*, represented by the center line, pushes P to Q.

With this notation, we can interpret $(\tilde{p}(u; S) : S \in \mathcal{D}_+)$ as a parametric family of densities, and for any fixed $S \in \mathcal{D}_+$, $\tilde{p}(u; S)$ is a density that integrates to 1. We note that by construction, any $S \in \mathcal{D}_+$ necessarily pushes $\tilde{P}(\cdot; S)$ to Q: $S_{\#}\tilde{P}(\cdot; S) = Q$. We can then cast the transport problem as finding the mapping $S \in \mathcal{D}_+$ that minimizes the relative entropy between P and the induced $\tilde{P}$:

$$S^* = \arg\min_{S \in \mathcal{D}_+} \; D(P \| \tilde{P}(\cdot; S)). \tag{3.2}$$

This perspective is represented visually in Figure 1.

If we again make another natural assumption,

Assumption 2. P admits a density p such that:

$$\mathbb{E}\left[\left|\log p(X)\right|\right] < \infty,$$

we can expand equation 3.2 and combine with equation 3.1 to write

$$\begin{aligned} S^* &= \arg\min_{S \in \mathcal{D}_+} \; D(P \| \tilde{P}(\cdot; S)) \\ &= \arg\min_{S \in \mathcal{D}_+} \; \mathbb{E}_P\left[\log \frac{p(X)}{\tilde{p}(X; S)}\right] \\ &= \arg\min_{S \in \mathcal{D}_+} \; -h(p) - \mathbb{E}_P\left[\log \tilde{p}(X; S)\right] \end{aligned} \tag{3.3}$$

$$= \arg\min_{S \in \mathcal{D}_+} \; -\mathbb{E}_P \left[\log \tilde{p}(X; S) \right] \tag{3.4}$$

$$= \arg\min_{S \in \mathcal{D}_+} \; -\mathbb{E}_P \left[\log q(S(X)) + \log \det J_S(X) \right], \tag{3.5}$$

where in equation 3.3, $h(p)$ is the Shannon differential entropy of p, which is fixed with respect to S; equation 3.4 is by assumption 2 and Jensen's inequality implying that $\left|h(p)\right| < \infty$ and the nonnegativity of KL divergence; and equation 3.5 is by combining with equation 3.1.

We now make another assumption for which we can guarantee efficient methods to solve for equation 3.2:

Assumption 3. The density q is log-concave.

We can now state the main theorem of this section (Kim, Mesa, Ma, & Coleman, 2015, Mesa et al., 2015):

Theorem 2 *(general push-forward). Under assumptions 1 to 3,*

$$\min_{S \in \mathcal{D}_+} \; D(P \| \tilde{P}(\cdot; S)) \tag{GP}$$

is a convex optimization problem.

Proof. For any $S, \tilde{S} \in \mathcal{D}_+$, we have that $J_S, J_{\tilde{S}} \succeq 0$. For any $\lambda \in [0, 1]$, we have that $\tilde{S}_\lambda \triangleq \lambda S + (1 - \lambda)\tilde{S}$ and $J_{\tilde{S}_\lambda} = \lambda J_S + (1 - \lambda) J_{\tilde{S}} \succeq 0$. Since $\log\det$ is strictly concave over the space of positive definite matrices (Boyd & Vandenberghe, 2004) and by assumption $\log q(\cdot)$ is concave, we have that $-\mathbb{E}_P \left[\log \tilde{p}(X; S) \right]$ is a convex function of S on $\mathcal{D}_+$. The existence of $S^* \in \mathcal{D}_+$ for which $D\big(P \| \tilde{P}(\cdot; S^*)\big) = 0$ is given by corollary 1. □

An important remark on this theorem:

Remark 1. Theorem 2 does not place any structural assumptions on P. It need not be log-concave, for example.

Beginning with equation 3.5, we see that problem GP can then be solved through the use of a Monte Carlo approximation of the expectation, and we arrive at the following sample-based version of the formulation,

$$S^* = \arg\min_{S \in \mathcal{D}_+} \; \frac{1}{N} \sum_{i=1}^{N} \left[-\log q(S(X_i)) - \log\det(J_S(X_i)) \right] \tag{3.6}$$

where $X_i \sim p(X)$.

3.2 Consensus Formulation. The stochastic optimization problem in equation 3.6 takes the general form of

$$\min_{S} \quad \sum_{i=1}^{N} f_i(S).$$

From this perspective, S can be thought of as a complicating variable. That is, this optimization problem would be entirely separable across the sum were it not for S. This can be instantiated as a global consensus problem,

$$\begin{aligned} \min_{S} \quad & \sum_{i=1}^{N} f_i(S_i) \\ \text{s.t.} \quad & S_i - S = 0, \end{aligned}$$

where the optimization is now separable across the summation, but we must achieve global consensus over S. With this in mind, we can now write a global consensus version of equation 3.6

$$\begin{aligned} \min_{S_i \in \mathcal{D}_+} \quad & -\frac{1}{N} \sum_{i=1}^{N} \log q(S_i(X_i)) + \log \det(J_{S_i}(X_i)) \\ \text{s.t.} \quad & S_i = S, \quad i =, 1 \ldots, N. \end{aligned} \tag{3.7}$$

In this problem, we can think of each (batch of) sample as independently inducing some random $\tilde{P}_i$ through a function S_i. The method we will propose can then be thought of as iteratively reducing the distance between each $\tilde{P}_i$ and the true P by reducing the distance between each S_i.

This problem is still over an infinite-dimensional space of functions $S \in \mathcal{D}_+$, however.

3.3 Transport Map Parameterization. To address the infinite-dimensional space of functions mentioned above, as in Kim et al. (2013, 2015), Marzouk et al. (2016) and Mesa et al. (2015), we parameterize the transport map over a space of multivariate polynomial basis functions formed as the product of D-many univariate polynomials of varying degree. That is, given some $\mathbf{x} = (x_1, \ldots, x_a, \ldots, x_D) \in \mathsf{W} \subset \mathbb{R}^D$, we form a basis function $\phi_{\mathbf{j}}(\mathbf{x})$ of multi-index degree $\mathbf{j} = (j_1, \ldots, j_a, \ldots, j_D) \in \mathcal{J}$ using univariate polynomials ψ_{j_a} of degree j_a as

$$\phi_{\mathbf{j}}(\mathbf{x}) = \prod_{a=1}^{D} \psi_{j_a}(x_a).$$

This allows us to represent one component of $S \in \mathcal{D}_+$ as a weighted linear combination of basis functions with weights $w_{d,\mathbf{j}}$ as

$$S^d(\mathbf{x}) = \sum_{\mathbf{j}\in\mathcal{J}} w_{d,\mathbf{j}}\ \phi_{\mathbf{j}}(\mathbf{x}),$$

where $\mathcal{J}$ is a set of multi-indices in the representation specifying the order of the polynomials in the associated expansion, and d denotes the dth component of the mapping. In order to make this problem finite-dimensional, we must truncate the expansion to some fixed maximum-order O:

$$\mathcal{J} = \left\{ \mathbf{j} \in \mathbb{N}^D : \sum_{i=1}^{D} j_i \leq O \right\}$$

We can now approximate any nonlinear function $S \in \mathcal{D}_+$ as

$$S(\mathbf{x}) = W\Phi(\mathbf{x}),$$

where $K \triangleq |\mathcal{J}|$ the size of the index set, $\Phi(\mathbf{x}) = [\phi_{\mathbf{j}_1}(\mathbf{x}), \ldots \phi_{\mathbf{j}_K}(\mathbf{x})]^T$, and $W \in \mathbb{R}^{D\times K}$ is a matrix of weights.

In order to avoid confusion and in the spirit of consensus ADMM as shown in Boyd et al. (2011), we introduce a consensus variable $B \triangleq W$. With this, we can now give a finite-dimensional version of equation 3.7 as

$$\begin{aligned} &\min_{W_i\in\mathbb{R}^{D\times K}} \ -\frac{1}{N}\sum_{i=1}^{N}\left[\log q(W_i\Phi(X_i)) + \log\det(W_i J_\Phi(X_i))\right] \\ &\text{s.t.} \quad W_i = B, \quad W_i J_\Phi(X_i) \succeq 0, \quad i = 1, \ldots, N, \end{aligned} \tag{3.8}$$

with:

$$\begin{aligned} W_i &= [w_1, \ldots, w_K] \quad D \times K, \\ \Phi(\cdot) &= \left[\phi_{\mathbf{j}_1}(\cdot), \ldots, \phi_{\mathbf{j}_K}(\cdot)\right]^T \quad K \times 1, \\ J_\Phi(\cdot) &= \left[\frac{\partial \phi_{\mathbf{j}_i}}{\partial x_j}(\cdot)\right]_{i,j} \quad K \times D, \end{aligned}$$

where we have made explicit the implicit constraint that $\det(J_S) \geq 0$ by ensuring that $BJ_\Phi \succeq 0$. We now provide two important remarks:

Remark 2. In principle, any basis of polynomials whose finite-dimensional approximations are sufficiently dense over W will suffice. In applications where P is assumed known, the basis functions are chosen to be orthogonal with respect to the reference measure P:

$$\int_{\mathsf{W}} \phi_{\mathbf{j}}(\mathbf{x})\ \phi_{\mathbf{i}}(\mathbf{x})\ p(x)dx = \mathbb{1}_{\mathbf{i}=\mathbf{j}}.$$

Within the context of Bayesian inference, for instance, this greatly simplifies computing conditional expectations, corresponding conditional moments, and so forth (Schoutens, 2000).

Remark 3. When it is important to ensure that the approximation satisfies the properties of a diffeomorphism, we can project $S(\mathbf{x})$ onto $\mathcal{D}_+$ with solving a quadratic optimization problem, as discussed in section 6.

We also note that the polynomial representation we have presented is chosen to best approximate a transport map, independent of a specific application or representation of the data (e.g., Fourier, wavelet). As noted in remark 2, in principle any dense basis will suffice.

3.4 Distributed Push-Forward with Consensus ADMM. In this section, we reformulate equation 3.8 within the framework of the ADMM, and provide our main result, corollary 2.

3.4.1 Distributed Algorithm. Using ADMM, we can reformulate equation 3.8 as a global consensus problem to accommodate a parallelizable implementation. For notational clarity, we write $\Phi_i \triangleq \Phi(X_i)$ and $J_i \triangleq J_\Phi(X_i)$. We then introduce the following auxiliary variables:

$$B\Phi_i \triangleq p_i, \quad BJ_i \triangleq Z_i.$$

We can now write equation 3.7 as

$$\begin{aligned}
\min_{\{W,Z,p\}_i,B} \quad & \frac{1}{N}\sum_{i=1}^{N} -\log q(p_i) - \log\det Z_i + \frac{1}{2}\rho\|W_i - B\|_2^2 \\
& + \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\rho\|B\Phi_i - p_i\|_2^2 + \frac{1}{2}\rho\|BJ_i - Z_i\|_2^2 \\
\text{s.t.} \quad & B\Phi_i = p_i : \qquad \gamma_i \quad (D\times 1) \\
& BJ_i = Z_i : \qquad \lambda_i \quad (D\times D) \\
& W_i - B = 0 : \qquad \alpha_i \quad (D\times K) \\
& Z_i \succeq 0,
\end{aligned}$$

where in the feasible set, we have denoted the Lagrange multiplier that will be associated with each constraint to the right.

Although coordinate descent algorithms solve for one variable at a time while fixing the others and can be extremely efficient, they are not always guaranteed to find the globally optimal solution (Wright, 2015). Using the consensus formulation of ADMM, we consider a problem formulation with the same global optimum that contains quadratic penalties associated with

equality constraints in the objective function and constraints still imposed. The consensus formulation has the key property that its Lagrangian, termed the augmented Lagrangian (Boyd et al., 2011), can be globally minimized with coordinated descent algorithms for any $\rho > 0$. Note that when $\rho = 0$, the augmented Lagrangian is equivalent to the standard (unaugmented) Lagrangian associated with equation 3.8.

We can now raise the constraints to form the fully penalized augmented Lagrangian as

$$\begin{aligned} L_\rho(W, Z, p, B; \gamma, \lambda, \alpha) = \frac{1}{N}\sum_{i=1}^{N} &- \log q(p_i) - \log\det Z_i \\ &+ \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\rho\|W_i - B\|_2^2 + \frac{1}{2}\rho\|B\Phi_i - p_i\|_2^2 \\ &+ \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\rho\|BJ_i - Z_i\|_2^2 + \gamma_i^T(p_i - B\Phi_i) \\ &+ \frac{1}{N}\sum_{i=1}^{N} \mathbf{tr}\left(\lambda_i^T(Z_i - BJ_i)\right) + \mathbf{tr}\left(\alpha_i^T(W_i - B)\right). \end{aligned}$$

The key property we leverage from the ADMM framework is the ability to minimize this Lagrangian across each optimization variable sequentially, using only the most recently updated estimates. After simplification (details are in the appendix), the final ADMM update equations for the remaining variables are

$$B^{k+1} = \mathcal{B}_i \cdot \mathcal{B}_s, \tag{3.9a}$$

$$W_i^{k+1} = -\frac{1}{\rho}\alpha_i^k + B^{k+1}, \tag{3.9b}$$

$$Z_i^{k+1} = Q\tilde{Z}_i Q^T, \tag{3.9c}$$

$$\gamma_i^{k+1} = \gamma_i^k + \rho(p_i^{k+1} - B^{k+1}\Phi_i), \tag{3.9d}$$

$$\lambda_i^{k+1} = \lambda_i^k + \rho(Z_i^{k+1} - B^{k+1}J_i), \tag{3.9e}$$

$$\alpha_i^{k+1} = \alpha_i^k + \rho(W_i^{k+1} - B^{k+1}), \tag{3.9f}$$

$$p_i^{k+1} = \arg\min_{p_i} -\log q(p_i) + \mathrm{pen}(p_i). \tag{3.9g}$$

We look first at the consensus variable B^{k+1}. We can separate its update into two pieces: a static component $\mathcal{B}_s$ and an iterative component $\mathcal{B}_i$:

$$\mathcal{B}_i = \frac{1}{N}\sum_{i=1}^{N}\left[\rho\left(W_i^k + p_i^k\Phi_i^T + Z_i^k J_i^T\right) + \gamma_i^k\Phi_i^T + \lambda_i^k J_i^T + \alpha_i^k\right], \tag{3.10a}$$

$$\mathcal{B}_s = \left[\rho\left(I + \frac{1}{N}\sum_{i=1}^{N}\Phi_i\Phi_i^T + J_i J_i^T\right)\right]^{-1}. \tag{3.10b}$$

The consensus variable can then be thought of as averaging the effect of all other auxiliary variables and forming the current best estimate for consensus among the distributed computational nodes.

The p-update is the only remaining minimization step that cannot necessarily be solved in closed form, as it completely contains the structure of the q density. In its penalization, all other optimization variables are fixed:

$$\text{pen}(p_i) = \frac{1}{2}\rho\|B^{k+1}\Phi_i - p_i\|_2^2 + \gamma_i^{kT}(p_i - B^{k+1}\Phi_i).$$

The formulation of equation 3.9 has the following desirable properties:

- Equations 3.9a to 3.9f admit closed-form solutions. In particular, equations 3.9b and 3.9d to 3.9f are simple arithmetic updates.
- Equation 3.9g is a penalized d-dimensional-vector convex optimization problem that entirely captures the structure of Q. In particular, any changes to the problem specifying a different structure of Q will be entirely confined in this update. Furthermore, algorithm designers can use any optimization procedure or library of their choosing to perform this update.

With this, we can now give an efficient, distributed version of the general push-forward theorem:

Corollary 2 *(distributed push-forward). Under assumptions 1 and 3,*

$$\begin{aligned}\min_{W_i\in\mathbb{R}^{d\times K}}\quad & -\frac{1}{N}\sum_{i=1}^{N} log\, q(W_i\Phi_i) + log\, det(W_i J_i)\\ s.t.\quad & W_i = W, \quad WJ_i \succeq 0 \quad i = 1,\ldots,N\end{aligned} \tag{3.11}$$

is a convex optimization problem.

Remark 4. ADMM convergence's properties are robust to inaccuracies in the initial stages of the iterative solving process (Boyd et al., 2011). In addition, several key concentration results provide very strong bounds for averages of random samples from log-concave distributions, showing that the approximation is indeed robust (Bobkov & Madiman, 2011, theorems 1.1 and 1.2).

The above framework, under natural assumptions, facilitates the efficient, distributed, and scalable calculation of an optimal map that pushes forward some P to some Q.

3.5 Structure of the Transport Map. An important consideration in ensuring the construction of transport maps is efficient is their underlying structure. In section 3.3, we described a parameterization of the transport map through the multi-index set $\mathcal{J}$—the indices of polynomial orders involved in the expansion. However, this parameterization tends to be unfeasible to use in high dimension or with high-order polynomials due to the exponential rate at which the number of polynomials increases with respect to these two properties.

Marzouk et al. (2016) discussed two less expressive but more computationally feasible map structures that can be used to generate the transport map. We briefly reproduce them here, along with some useful properties. (For more specific details and examples of multi-index sets pertaining to each mode for implementation purposes, see section 6.)

The first alternative to the map pertaining to the fully expressive mapping is the Knothe-Rosenblatt map (Bonnotte, 2013), which our group also previously used within the context of generating transport maps for optimal message point feedback communication (Ma & Coleman, 2011). Here, each component of the output, S^d, is only a function of the first d components of the input, resulting in a mapping that is lower-triangular. Both the Knothe-Rosenblatt and dense mapping described above perform the transport from one density to another, but with different geometric transformations. An example of these differences can be found in Figures 3 and 4 in Ma and Coleman (2011).

A Knothe-Rosenblatt arrangement gives the following multi-index set (note that the index set is now subscripted according to dimension of the data denoting the dependence on data component):

$$\mathcal{J}_d^{KR} = \left\{ \mathbf{j} \in \mathbb{N}^D : \sum_{i=1}^{D} j_i \leq O \wedge j_i = 0, \forall i > d \right\}, d = 1, \ldots, D.$$

An especially useful property of this parameterization is the following identity for the Jacobian of the map:

$$\log \det(J_S(X_i)) = \sum_{d=1}^{D} \log \partial_d S^d(X_i), \tag{3.12}$$

where $\partial_d S^d(X_i)$ represents the partial derivative of the dth component of the mapping with respect to the dth component of the data, evaluated at X_i.

Furthermore, the positive-definiteness of the Jacobian can equivalently be enforced for a lower-triangular mapping by ensuring the following:

$$\partial_d S^d > 0, \quad 1 \leq d \leq D. \tag{3.13}$$

We can then write a Knothe-Rosenblatt special-case version of equation 3.13 as

$$\begin{aligned} &\min_{S_i \in \mathcal{D}_+^{KR}} -\frac{1}{N}\sum_{i=1}^{N} \log q(S_i(X_i)) + \sum_{d=1}^{D} \log \partial_d S_i^d(X_i) \\ &\text{s.t.} \quad S_i = S, \quad i = 1, \ldots, N \end{aligned} \tag{3.14}$$

Indeed, we use this to our advantage in section 4.

Finally, in the event that the Knothe-Rosenblatt mapping also proves to have model complexity that is too high, an even less expressive mapping is a Knothe-Rosenblatt mapping that ignores all multivariate polynomials that involve more than one data component of the input at a time, resulting in the following multi-index set:

$$\mathcal{J}_d^{KRSV} = \left\{ \mathbf{j} \in \mathbb{N}^D : \sum_{i=1}^{D} j_i \leq O \wedge j_i j_l = 0, \forall i \neq l \wedge j_i = 0, \forall i > d \right\},$$
$$d = 1, \ldots, D.$$

Although less expressive and less precise than the total order Knothe-Rosenblatt map, these maps can often still perform at an acceptable level of accuracy with respect to many problems.

3.6 Algorithm for Inverse Mapping with Knothe-Rosenblatt Transport. It may be desirable to compute the inverse mapping of a given sample from Q, that is, $S^{-1}(X)$, $X \sim Q$. When the forward mapping S is constrained to have Knothe-Rosenblatt structure and a polynomial basis is used to parameterize the mapping, the process of inverting a sample from Q reduces to solving a sequential series of polynomial root-finding problems (Marzouk et al., 2016). We give a more detailed implementation-based explanation of this process alongside a discussion of implementation details for the Knothe-Rosenblatt maps in section 6.

4 Sequential Composition of Optimal Transportation Maps

In this section, we introduce a scheme for using many individually computed maps in sequential composition to achieve an overall effect of a single large mapping from P to Q. By using a sequence of maps to transform P to Q instead of a single one-shot map, one can theoretically rely on models

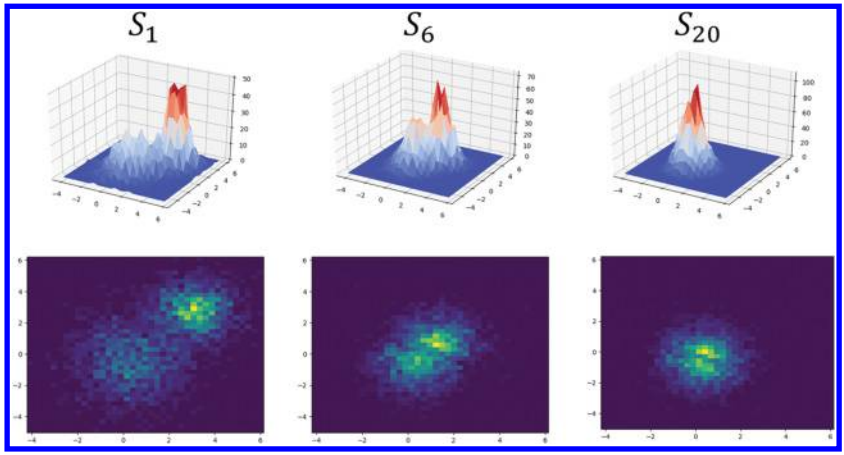

Figure 2: A visual representation of the effect a sequential composition has over the density of a set of samples shown at intermediary stages of the mapping sequence. P is a two-dimensional bimodal distribution, and Q is standard gaussian.

of lower complexity to represent each map in the sequence; although each map is, on its own, "weak" in the sense of its ability to induce large changes in the distribution space, the combined action of many such maps together can potentially transform samples as desired. This is especially attractive for model structures that increase exponentially in complexity with problem size, such as the dense polynomial chaos structure discussed in the previous section. This sequential composition process is visually represented in Figure 2.

Moving forward, we first take a brief look at a nonequilibrium thermodynamics interpretation of this methodology to further justify the use of such a scheme, and then derive a slightly different ADMM problem to implement it.

4.1 Nonequilibrium Thermodynamics and Sequential Evolution of Distributions. One approach to interpreting sequential composition of maps is to borrow ideas from statistical physics, where we can interpret q as the equilibrium density (ρ_∞) of particles in a system, which at time 0 is out of equilibrium with density P (also termed ρ_0). Since q is an equilibrium density, it can be written as a Gibbs distribution (with temperature equal to 1 for simplicity): $q(u) \equiv \rho_\infty(u) = Z^{-1} \exp(-\Psi(u))$. For instance, if Q pertains to a standard gaussian, then $\Psi(u) = \frac{1}{2}u^2$. Assuming the particles obey the Langevin equation, it is well known that the evolution of the particle density as a function of time ($\rho_t : t \geq 0$) obeys the Fokker-Planck

equation. Jordan, Kinderlehrer, and Otto (1998b) showed that the trajectory of $(\rho_t : t \geq 0)$ can be interpreted from variational principles. Specifically,

Theorem 3 *(Jordan et al., 1998b, theorem 5.1). Define $\rho_0 = p$ and $\rho_\infty = q$ and assume that $D(\rho_0 \| \rho_\infty) < \infty$. For any $h > 0$, consider the following minimization problem:*

$$A(\rho) \triangleq \frac{1}{2} d(\rho_{k-1}, \rho)^2 + hD(\rho \| \rho_\infty) \tag{4.1}$$

$$\rho_k \triangleq \underset{\rho \in P(W)}{arg\ min}\, A(\rho). \tag{4.2}$$

Then as $h \downarrow 0$, the piecewise constant interpolation, which equals ρ_k for $t \in [kh, (k+1)h)$, converges weakly in $L_1(\mathbb{R}^D)$ to $(\rho_t : t \geq 0)$, the solution to the Fokker-Planck equation.

The log-concave structure of q we have exploited previously also has implications for exponential convergence to equilibrium with this statistical physics perspective:

Theorem 4.2 *(Bakry & Émery, 1985). If q is uniform log-concave, namely,*

$$\nabla^2 \Psi(u) \succeq \lambda I_D,$$

for some $\lambda > 0$ with I_D the $D \times D$ identity matrix, then

$$D(\rho_t \| \rho_\infty) \leq e^{-2\lambda t} D(\rho_0 \| \rho_\infty)\,.$$

Note that if q is the density of a standard gaussian, this inequality holds with $\lambda = 1$.

4.2 Sequential Construction of Transport Maps. We now note that for any $h > 0$, equation 4.2 encodes a sequence $(\rho_k : k \geq 0)$ of densities that evolve toward $\rho_\infty \equiv q$. For notational conciseness in this section, we will be using the subscript on S to denote the position of the map in a sequence of maps. As such, from corollary 1, there exists an $S_1 \in \mathcal{D}_+$ for which $S_1 \# \rho_0 = \rho_1$, and, more generally, for any $k \geq 0$, there exists an $S_k \in \mathcal{D}_+$ for which $S_k \# \rho_{k-1} = \rho_k$.

Lemma 2. *Define $B : \mathcal{D}_+ \to \mathbb{R}$ as*

$$B(S) \triangleq \frac{1}{2} \mathbb{E}_{\rho_{k-1}} \left[\|X - S(X)\|^2 \right] + hD(\rho_{k-1} \| \tilde{p}(\cdot; S))$$

$$S_k \triangleq \underset{S \in \mathcal{D}_+}{arg\ min}\, B(S). \tag{4.3}$$

Then $A(\rho_k) = B(S_k)$ and $S_k \# \rho_{k-1} = \rho_k$.

Proof. From the definition of $\tilde{p}_{S,Q}$ in equation 3.1 and the invariance of relative entropy under an invertible transformation, any $S \in \mathcal{D}_+$ satisfies

$$D(\rho_{k-1}\|\tilde{p}(\cdot; S)) = D\big(\rho_{k-1}\|S^{-1}\#\rho_\infty\big) = D(S\#\rho_{k-1}\|\rho_\infty)\,.$$

As such, moving forward with the proof, we exploit how $B(S) = \tilde{B}(S)$ where

$$\tilde{B}(S) \triangleq \frac{1}{2}\mathbb{E}_{\rho_{k-1}}\left[\|X - S(X)\|^2\right] + hD(S\#\rho_{k-1}\|\rho_\infty)\,.$$

From theorem 1, $d(\rho_{k-1}, S\#\rho_{k-1}) \leq \mathbb{E}_{\rho_{k-1}}\left[(X - S(X))^2\right]$ for any $S \in \mathcal{D}_+$. Also, since the relative entropy terms of $\tilde{B}(S)$ and $A(S\#\rho_{k-1})$ are equal, it follows that $\tilde{B}(S) \geq A(S\#\rho_{k-1})$ for any $S \in \mathcal{D}_+$. Moreover, from corollary 1, we have that there exists an $S_k \in \mathcal{D}_+$ for which $S_k\#\rho_{k-1} = \rho_k$ and

$$\mathbb{E}_{\rho_{k-1}}\left[\|X - S_k(X)\|^2\right] = d(\rho_{k-1}, \rho_k)^2.$$

Thus, $\tilde{B}(S) = A(S\#\rho_{k-1})$. □

As such, a natural composition of maps underlies how a sample from $P \equiv \rho_0$ gives rise to a sample from ρ_k:

$$\rho_k = S_k\#\rho_{k-1} = S_k \circ S_{k-1}\#\rho_{k-2} = S_k \circ \cdots \circ S_1\#\rho_0. \tag{4.4}$$

Moreover, since as $h \downarrow 0$, $\rho_k \simeq \rho_{k-1}$ and so S_k approaches the identity map. Thus, for small $h > 0$, each S_k should be estimated with reasonable accuracy using lower-order maps. That is, S can be described as the composition of T maps as

$$S(x) = S_T \circ \ldots \circ S_2 \circ S_1(x) \tag{4.5}$$

for all $x \in \mathbb{R}^d$, such that each S_i is of relatively low order in the polynomial chaos expansion.

Note that $B(S)$ as written above involves a sum of expectations with respect to ρ_{k-1}. Since our scheme operates sequentially, we have already estimated $S_1, S_2, \ldots, S_{k-1}$ and can generate approximate independent and identically distributed (i.i.d.) samples from ρ_{k-1} by first generating $(X_i : i \geq 1)$ i.i.d. from $\rho_0 \equiv p$ and constructing

$$Z_i = S_{k-1} \circ \cdots \circ S_1(X_i), \quad i \geq 1.$$

We will demonstrate efficient ways to solve the below-convex-optimization problem, which replaces the expectation with respect to ρ_{k-1} instead with the empirical expectation with respect to $(Z_i : i = 1, \ldots, N)$:

$$\min_{S\in\mathcal{D}_+} \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\|Z_i - S(Z_i)\|^2 - h\log\tilde{p}(Z_i;S)\right].$$

To reiterate, we consider a distribution ρ_{k-1} formed by the sequential composition of previous mappings as

$$\rho_{k-1} = S_{k-1}^* \circ \cdots \circ S_1^* \# \rho_0,$$

where $\rho_0 \equiv p$. We then try to find a map S_k^* that pushes ρ_{k-1} forward closer to $\rho_\infty \equiv Q$. Each S_k is solved by the optimization problem 4.3, which we term SOT. As the number of compositions T in equation 4.5 increases, ρ_T approaches ρ_∞. When q is uniform log-concave, this greedy, sequential approach still guarantees exponential convergence.

In the context of Knothe-Rosenblatt maps, for every map in the sequence, we can solve the following optimization problem (in the following equation, we will be dropping the subscript k that indicates the sequential map index, as the formulation is not dependent on position in the map sequence, and we will once again be replacing the subscript with i to indicate the distributed variables for the consensus problem instead):

$$\min_{S_i\in\mathcal{D}_+^{KR}} \theta||S_i(X_i) - X_i||_2^2 - \frac{1}{N}\sum_{i=1}^{N}\log q(S_i(X_i)) + \sum_{d=1}^{D}\log\partial_d S^d(X_i) \quad (4.6)$$
$$\text{s.t.} \quad S_i = S, \quad \forall 0 \le i \le N,$$

where $\theta = h^{-1}$ can be interpreted as an inverse "step-size" parameter.

Though each map in the sequence must be calculated sequentially after the previous one, each mapping can still be calculated in the distributed framework described. This implies that at each round, one could adaptively decide the parameters for the next-round's solution.

4.3 ADMM Formulation for Learning Sequential Maps. We now showcase an ADMM formulation for the optimal transportation-based objective function, similar in spirit to that of equation 3.9.

We first introduce the following conventions:

- Φ_i^d represents the partial derivative of Φ_i taken with respect to the dth component. Therefore, $B\Phi_i^d = \partial_d S(X_i)$, and $\partial_d S^d(X_i)$ is the dth component of $B\Phi_i^d$.
- $\mathbf{1}_d$ represents a one-hot vector of length D with the one in the dth position.

We can then introduce a finite-dimensional representation of the transport map, as well as auxiliary variables and a consensus variable to equation

4.6, and rewrite the problem as

$$\begin{aligned}
&\min_{\{W,p\}_i,\{Y,Z\}_i^d,B} \theta||B\Phi_i - x_i||_2^2 + \frac{1}{N}\sum_{i=1}^{N} -\log q(p_i) \\
&\quad + \frac{1}{2}\rho||W_i - B||_2^2 + \frac{1}{2}||B\Phi_i - p_i||_2^2 \\
&\quad + \sum_{d=1}^{D} -\log Z_i^d + \frac{1}{2}\rho(Y_i^d\mathbf{1}_d - Z_i^d)^2 + \frac{1}{2}\rho||B\Phi_i^d - Y_i^d||_2^2 \\
&\text{s.t} \quad B\Phi_i = p_i \qquad \gamma_i \quad (D\times 1) \\
&\quad W_i - B = 0 \qquad \alpha_i \quad (D\times K) \\
&\quad Y_i^d\mathbf{1}_d = Z_i^d \qquad \beta_i^d \quad (1\times 1) \\
&\quad B\Phi_i^d = Y_i^d \qquad \lambda_i^d \quad (D\times 1) \\
&\quad Z_i^d > 0,
\end{aligned} \tag{4.7}$$

where we have once again denoted the corresponding Lagrange multipliers to the right of each constraint. The superscript d notation represents the fact that in this formulation, in addition to having separable variables for each data sample, some variables are now unique to an index over dimension as well. For example, there are DN-many Z variables that must be solved for. We can now raise the constraints to form the fully penalized Lagrangian as

$$\begin{aligned}
&L_{\rho,\theta}(W, Z, Y, p, B; \gamma, \alpha, \beta, \lambda) \\
&\quad = \theta||B\Phi_i - x_i||_2^2 + \frac{1}{N}\sum_{i=1}^{N} -\log q(p_i) \\
&\qquad + \frac{1}{2}\rho||W_i - B||_2^2 + \frac{1}{2}\rho||B\Phi_i - p_i||_2^2 \\
&\qquad + \gamma_i^T(p_i - B\Phi_i) + \mathbf{tr}(\alpha_i^T(F_i - B)) \\
&\qquad + \sum_{d=1}^{D} -\log Z_i^d + \frac{1}{2}\rho(Y_i^d\mathbf{1}_d - Z_i^d)^2 + \frac{1}{2}\rho||B\Phi_i^d - Y_i^d||_2^2 \\
&\qquad + \beta_i^d(Z_i^d - Y_i^d\mathbf{1}_d) + \lambda_i^{dT}(Y_i^d - B\Phi_i^d).
\end{aligned} \tag{4.8}$$

The final ADMM update equations for each variable are once again all closed form, with the exception of the optimization over p_i. (For the sake of brevity, we refer the readers to section A.6 for the exact update equations.)

One notable difference between this formulation and that of section 3.4.1 is that the update for Z_i^d has been simplified from requiring an eigenvalue

decomposition to requiring a simple scalar computation, thus significantly reducing computation time, especially in higher dimensions.

4.4 Scaling Parallelization with GPU Hardware. Given the parallelized formulations above, we implemented our algorithm using the Nvidia CUDA API to get as much performance as possible out of our formulation, and to maximize the problem sizes we could reasonably handle, while keeping computation time as short as possible. To test the algorithm's parallelizability, we ran our implementation on a single Nvidia GTX 1080ti GPU, as well as on a single p3.16xlarge instance available on Amazon Web Services, which itself contains eight on-board Tesla V100 GPUs.

For this test, we have sampled synthetic data from a bimodal P distribution specified as a combination of two gaussian distributions, for a wide range of problem dimensions, specifically $D = 5, 10, 20, 50, 100, 150, 200$, and a constant number of samples from P set to $N = 1000$. We then find a transport pushing P to $Q = \mathcal{N}(\mathbf{0}, \mathbf{I})$, composed of a sequence of 10 individual Knothe-Rosenblatt maps with no mixed multivariate terms. We then monitor the convergence of dual variables for proper termination of the algorithm.

Figure 3 shows the result of this analysis. The one GPU curve corresponds to performance using the single GTX 1080ti, and the AWS curve corresponds to the performance using the eight-GPU system on Amazon Web Services. The trending of the curves shows that as expected, as problem dimension increases, a multi-GPU system will continue to maintain reasonable computation times, at least with respect to a single-GPU system; however, fewer GPU's will begin to accumulate increasingly high computational costs. In addition, the parallelizability of our algorithm also has a subtle benefit of helping with memory-usage issues. Since we can distribute samples across multiple devices, we can also subsequently distribute all corresponding ADMM variables as well. Indeed, the single GTX 1080ti ran out of on-board memory roughly around $D = 230$, whereas the eight-GPU system can go well beyond that.

5 Applications

The framework we have presented is general purpose, and works to push forward a distribution P to a log-concave distribution Q. We now discuss some interesting applications, namely, Bayesian inference and a generative model, and show results with real-world data sets.

5.1 Bayesian Inference. A very important instantiation of this framework comes when we consider $P \equiv P_X$ to represent a prior distribution and $Q \equiv P_{X|Y=y}$ to be a Bayesian posterior,

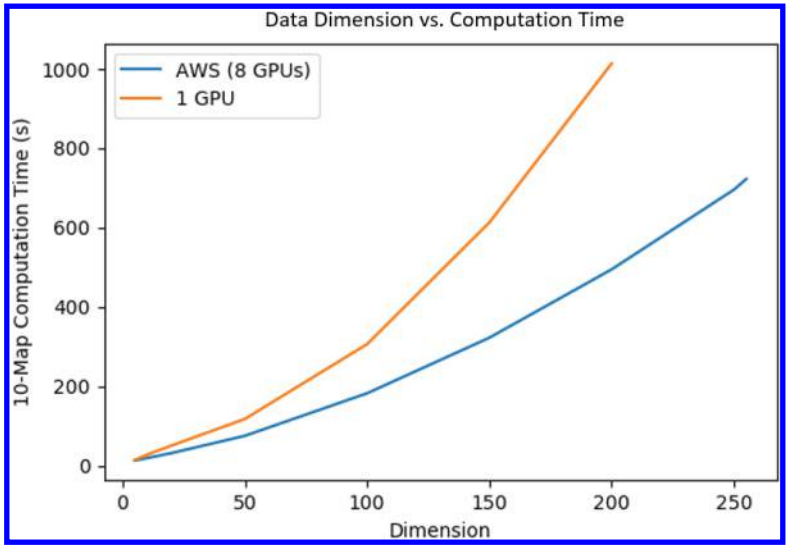

Figure 3: A comparison of using a single-GPU system versus an eight-GPU system to compute maps in increasingly high dimension. The trending of the two plots clearly shows the more reasonable growth in computation time of the eight-GPU system relative to the single-GPU system, as the samples from P are distributed among the multiple devices.

$$f_{X|Y=y}(x) = \frac{f_{Y|X}(y|x)f_X(x)}{\beta_y},$$

where β_y is a constant that does not vary with x, given by

$$\beta_y = \int_{v\in\mathcal{X}} f_{Y|X}(y|v)f_X(v)dv.$$

Using equation 2.1 and combining with Bayes' rule above, we can write

$$\begin{aligned} f_X(x) &= f_{X|Y=y}(S^*_{(y)}(x))\det\left(J_{S^*_{(y)}(x)}\right) \\ &= \frac{f_{Y|X}(y|S^*_{(y)}(x))f_X\left(S^*_{(y)}(x)\right)}{\beta_y}\det\left(J_{S^*_{(y)}(x)}\right), \end{aligned}$$

where the notation $S^*_{(y)}(x)$ indicates that the optimal map is found with respect to observations y. We note that since $q(u) = \frac{f_X(u)f_{Y|X}(y|u)}{\beta_y}$, log-concavity of q is equivalent to log-concavity of the prior density $f_X(u)$ and log-concavity of the likelihood density $f_{Y|X}(y|u)$ in u—the same criterion for

a maximum a-posteriori (MAP) estimation procedure to be convex. Thus, corollary 2 extends to the special case of Bayesian inference: we can generate i.i.d. samples from the posterior distribution by solving a convex optimization problem in a distributed fashion.

Due to the unique way the ADMM steps were structured, this special case only requires specifying a particular instance of equation 3.12g:

$$p_i^* = \arg\min_{p_i} - \log \underbrace{f_{Y|X}(y|p_i)}_{\text{likelihood}} - \log \underbrace{f_X(p_i)}_{\text{prior}} + \text{pen}(p_i).$$

Remark 5. This specific case establishes an important property. If the prior is chosen so that it is easy to sample from, and the prior and likelihood are both log-concave, then a deterministic function S can be efficiently computed that takes i.i.d. samples from the prior distribution and results in i.i.d. samples from the posterior distribution. The assumption of log-concavity is also typically used in large-scale point estimates, though this framework goes beyond point estimates and generates i.i.d. samples from the posterior.

As an instantiation of this framework, we consider a Bayesian estimation of regression parameters $x_1, \ldots, x_d$ in the model $y = \mu \mathbf{1_n} + \Phi x + \epsilon$, where y is the n-dimensional vector of responses, μ is the overall mean, Φ is an $n \times d$ regressor matrix, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise vector. The Lasso solution,

$$x^* = \arg\min_{x \in \mathbb{R}^d} ||y - \Phi x||_2^2 + \lambda ||x||_1, \tag{5.1}$$

for some $\lambda \geq 0$ induces sparsity in the latent coefficients. The solution to equation 5.1 can be seen as a posterior mode estimate when the regression parameters are distributed accordingly to a Laplacian prior:

$$p(x; \lambda) = \prod_{i=1}^{d} \frac{\lambda}{2} e^{-\lambda |x_i|}. \tag{5.2}$$

A number of Bayesian Lasso Gibbs samplers, which are Markov chain Monte Carlo algorithms, are used as standard methods by which to sample from the posterior associated with problem 5.1 (Park & Casella, 2008; Hans, 2009).

We study the accuracy and modularity of our measure transport methodology through a Bayesian Lasso analysis of the Boston Housing data set, first analyzed by Harrison and Rubinfeld (1978), which is a common data set used when comparing regression problems. We compare our results to those obtained from utilizing a corresponding Gibbs sampler. The Boston Housing data set consists of 13 independent predictors of the median value of owner-occupied homes and 506 cases. We are interested in which

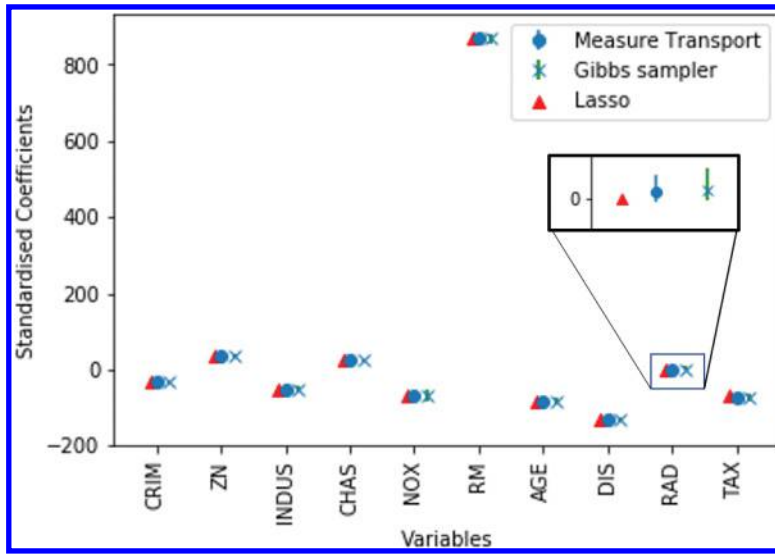

Figure 4: Posterior median Bayesian Lasso estimates and corresponding credible intervals for the 10 first variables of the Boston Housing data set. Median estimates were obtained with samples from a Gibbs sampler and a measure transport map. Lasso estimates are shown for comparison.

combination of these 13 variables best predicts the median value of homes observed in y and if we can eliminate variables that do not contribute much to prediction. The Lasso gives an automatic way for feature selection by forcing the coefficients of the predictors represented by x^* to be zero. The Bayesian Lasso solution allows for uncertainty quantification of feature selection, as we can obtain credible intervals corresponding to the coefficients of the estimates.

We used a Gibbs sampler as presented in Hans (2009) where the variance variable σ^2 is nonrandom. We used 3000 samples of burn-in and sampled 10,000 samples from the posterior distribution with a fixed λ chosen by minimizing the Bayes information criterion (BIC; Zou, Hastie, & Tibshirani, 2007). We compared that to sampling from a generated transport map with the same λ. We used $N = 2000$ samples from a Laplace prior to learn a fourth-order transport map of interest. In this case, we used a one-shot, dense map structure as described in section 3.

We note that the modularity of our problem allows for sampling from the posterior distribution of the Bayesian Lasso by only specifying the optimization problem of equation 3.12g to correspond to the likelihood and prior.

Figure 4 shows the posterior median estimates and the corresponding 95% credible intervals for the marginal distributions of the first 10 variables

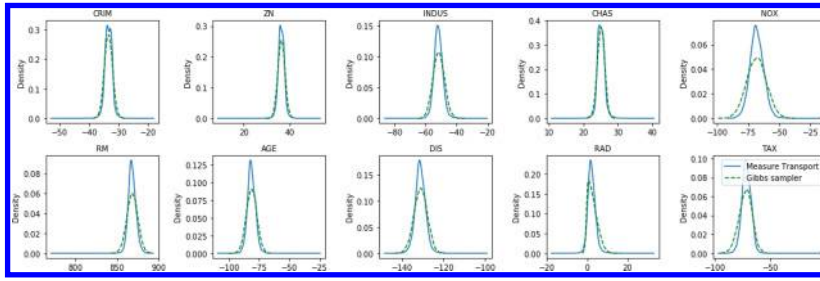

Figure 5: Kernel density estimate comparisons of marginal posteriors for the Boston Housing data set.

of the Boston housing data set. The Lasso estimates are shown for comparison. Figure 5 shows the kernel density estimates for these variables constructed with 10,000 samples of either the Bayesian Lasso Gibbs sampler or the measure transported samples. The density estimates of both methods are similar, verifying the accuracy of our methodology.

5.2 High-Dimensional Maps Using the MNIST Data Set. The parallelizability of our formulation of the optimal transport-based mapping for sequential transport maps also allows us to efficiently compute maps for relatively high-dimensional data. As a demonstration of this, we used the MNIST handwritten digits data set (LeCun et al., 1998) as a subject of experimentation.

Similar to the density estimation case, we assume that samples from each class of MNIST data are drawn from some P_{digit}, where *digit* denotes the MNIST written digit associated with that distribution. We then attempt to construct a (sequential) mapping, $S_{(digit)}$, that pushes P_{digit} to a reference distribution, $Q = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Again, similar to before, the selection of the Q density to be a standard gaussian is expressly for the purpose of analytical simplicity; Q can theoretically be anything we like, so it benefits us during the generative step to select Q such that it is easy to sample from. Each image in MNIST is a 28×28 pixel image; therefore, after flattening each image into a vector of data, our maps operate in $D = 784$. We then solve for each map $S_{(digit)}$ for every handwritten digit class in the MNIST set.

We can then treat the inverse map as a generative model. With the maps $S_{(digit)}$ in hand, we can theoretically draw samples from Q and push these samples through the inverse map, $S^{-1}_{(digit)}$, resulting in randomly generated samples from P_{digit}.

Figure 6 shows the result of this process using a sequential composition of 15 maps, with maximum order of the basis of each sequential map being set to 2 and each sequential map using the Knothe-Rosenblatt basis with no mixed multivariate terms from section 3.5. Our results show that even in high dimension and even while using a relatively weak polynomial basis

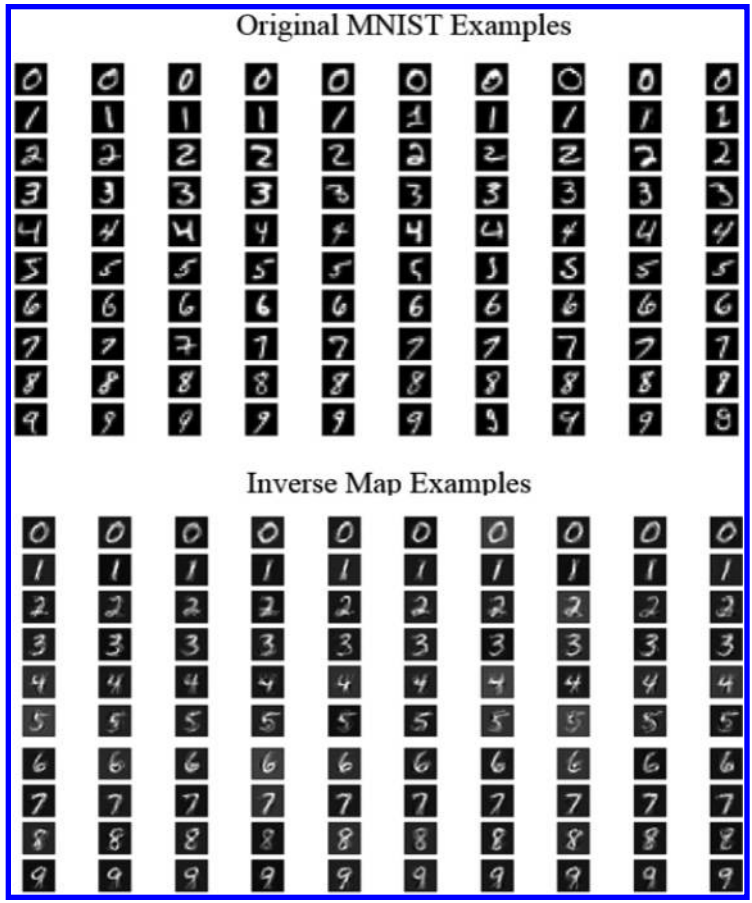

Figure 6: A comparison of original MNIST data samples versus random samples drawn using the inverse map. The left-most 10 columns of images pertain to randomly selected data examples from the original MNIST set, and the right-most 10 columns of images are randomly generated by the inverse map, $S^{-1}_{(digit)}(X)$, $X \sim Q$. Each mapping for this example was a sequential composition of 15 maps of maximum order 2, using the Knothe-Rosenblatt mapping with no mixed terms.

per sequential map, the resulting transport maps can effectively generate approximate samples from P_{digit} in this way.

6 Discussion

In this work, we have proposed a general-purpose framework for pushing independent samples from one distribution P to independent samples from another distribution Q through the efficient and distributed construction of

transport maps, with only independent samples from P, and knowledge of Q up to a normalization constant. We showed that when the target distribution Q is log-concave, this problem is convex. Using ADMM, we instantiated two finite-dimensional problems for finding both one-shot and sequential transport maps and provided distributed algorithms for carrying out the underlying optimization problems. As our framework is distributed by nature, we can continue to take advantage of the ever-increasing availability and evolution of distributed computational resources to speed up computation, with few to no changes to our formulation whatsoever.

We applied our framework to a Bayesian Lasso problem, that, while it requires the prior and likelihood to be log-concave, is no different from existing frameworks that carry out efficient point estimates in that regard; however, by contrast, our framework does succeed in efficiently generating independent samples from the actual target distribution Q. We emphasize that the class of log-concave distributions is quite large and widely used in various applications (Bagnoli & Bergstrom, 2005), and that this is the same convexity condition required for Bayesian point (MAP) estimation using many modern techniques. As such, we have shown that from the perspective of convexity, we can go from point estimation to fully Bayesian estimation without requiring significantly more.

Finally, we applied our framework to a high-dimensional problem of approximating a generative model for the MNIST data set and provided a qualitatively striking demonstration of how well the construction of sequential transport maps can give rise to such a model. The connection and comparison of this method to other generative models, especially deep learning-based methods such as generative adversarial networks (Goodfellow et al., 2014) and variational autoencoders (Kingma & Welling, 2013), remains to be explored and is the subject of future work. We believe that this alternate form of generative model, one based on calculating a transport map that is parameterized over the space of polynomial basis functions orthogonal to the distribution of the data, stands in contrast to the black-box nature of neural networks. Moreover, although certain works have explored the invertibility of deep neural networks (Lipton & Tripathi, 2017; Gilbert, Zhang, Lee, Zhang, & Lee, 2017), in general a single output of a neural network might map to multiple latent vectors. Our transport maps, chosen over the space of diffeomorphisms, remain necessarily invertible, and indeed this property is exploited in the generation of samples. One can surmise that this invertibility leads to more tractability of the generative model. The general connection to optimal transport and deep generative models is a subject of recent interest and has fostered pertinent work in the literature (Genevay, Peyré, & Cuturi, 2017; Salimans, Zhang, Radford, & Metaxas, 2018).

We also stress that ADMM and other related large-scale optimization methods have many existing refinements (Azadi & Sra, 2014; Jordan, Kinderlehrer, & Otto, 1997; Jordan et al., 1998b; Zhong & Kwok, 2014) from which this framework would immediately benefit. Future work could

explore these refinements and applications as approximations to nonconvex problems.

Although we have established the convexity of these schemes, further work needs to be done characterizing the fundamental limits of sample complexity of this approach and can help guide how these architectures may possibly be soundly implemented. Optimizing architectures for hardware optimization and understanding performance-energy-complexity trade-offs will allow for wider exploration of these methods within the context of emerging applications.

Appendix

We provide some additional details on several aspects of the main text.

A.1 Derivation of Dense ADMM Formulation. We show a more complete derivation of the ADMM formulation from section 3.4.1. ADMM yields the following sequential updates to the penalized Lagrangian:

$$B^{k+1} = \arg\min_{B} L_\rho(W^k, Z^k, p^k, B; \gamma^k, \lambda^k, \alpha^k) \tag{A.1a}$$

$$W^{k+1} = \arg\min_{W} L_\rho(W, Z^k, p^k, B^{k+1}; \gamma^k, \lambda^k, \alpha^k) \tag{A.1b}$$

$$Z^{k+1} = \arg\min_{Z \succ 0} L_\rho(W^{k+1}, Z, p^k, B^{k+1}; \gamma^k, \lambda^k, \alpha^k) \tag{A.1c}$$

$$p^{k+1} = \arg\min_{p} L(W^{k+1}, Z^{k+1}, p, B^{k+1}; \gamma^k, \lambda^k, \alpha^k) \tag{A.1d}$$

$$\gamma_i^{k+1} = \gamma_i^k + \rho(p_i^{k+1} - B^{k+1}\Phi_i) \quad 1 \le i \le n \tag{A.1e}$$

$$\lambda_i^{k+1} = \lambda_i^k + \rho(Z_i^{k+1} - B^{k+1}J_i) \quad 1 \le i \le n \tag{A.1f}$$

$$\alpha_i^{k+1} = \alpha_i^k + \rho(W_i^{k+1} - B^{k+1}) \quad 1 \le i \le n \tag{A.1g}$$

The closed-form solutions to equations A.1a to A.1c are given as follows. First, as for equation A.1a, the cost function $C(B^{k+1})$ is given by

$$\begin{aligned} C(B^{k+1}) = {} & \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\rho\|W_i^k - B\|_F^2 + \frac{1}{2}\rho\|B\Phi_i - p_i^k\|_2^2 \\ & + \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\rho\|BJ_i - Z_i^k\|_F^2 + \gamma_i^{kT}(p_i^k - B\Phi_i) \\ & + \frac{1}{N}\sum_{i=1}^{N} \mathbf{tr}\left(\lambda_i^{kT}(Z_i^k - BJ_i)\right) + \mathbf{tr}\left(\alpha_i^{kT}(W_i^k - B)\right). \end{aligned} \tag{A.2}$$

The first-order derivative of the equation A.2 in terms of B^{k+1} is expressed as

$$\begin{aligned}\frac{\partial C(B^{k+1})}{\partial B^{k+1}} = &\frac{1}{N}\sum_{i=1}^{N}\rho(B - W_i^k) + \rho(B\Phi_i - p_i^k)\Phi_i^T \\ &+ \frac{1}{N}\sum_{i=1}^{N}\rho(BJ_i - Z_i^k)J_i^T - \gamma_i^k\Phi_i^T \\ &+ \frac{1}{N}\sum_{i=1}^{N} -\lambda_i^k J_i - \alpha_i^{kT}. \end{aligned} \tag{A.3}$$

By setting the equation A.3 to zero and expressing it in terms of B, we get

$$\begin{aligned}&B\left[\rho\left(I + \frac{1}{N}\sum_{i=1}^{N}\Phi_i\Phi_i^T + J_iJ_i^T\right)\right] \\ &\quad = \frac{1}{N}\sum_{i=1}^{N}\left[\rho\left(W_i^k + p_i^k\Phi_i^T + Z_i^kJ_i^T\right) + \gamma_i^k\Phi_i^T + \lambda_i^kJ_i^T + \alpha_i^k\right]. \end{aligned} \tag{A.4}$$

If we define

$$L \triangleq \left[\rho\left(I + \frac{1}{N}\sum_{i=1}^{N}\Phi_i\Phi_i^T + J_iJ_i^T\right)\right] \tag{A.5}$$

and

$$M \triangleq \frac{1}{N}\sum_{i=1}^{N}\left[\rho\left(W_i^k + p_i^k\Phi_i^T + Z_i^kJ_i^T\right) + \gamma_i^k\Phi_i^T + \lambda_i^kJ_i^T + \alpha_i^k\right], \tag{A.6}$$

then we have

$$B^{k+1} = M \cdot L^{-1}. \tag{A.7}$$

Second, as for equation A.1b, the cost function $C(W_i^{k+1})$ is given by

$$C(W_i^{k+1}) = \frac{1}{2}\rho\|W_i - B^{k+1}\|_2^2 + \mathbf{tr}\left(\alpha_i^{kT}(W_i - B^{k+1})\right). \tag{A.8}$$

The first-order derivative of equation A.8 in terms of W_i^{k+1} is expressed as

$$\frac{\partial C(W_i^{k+1})}{\partial W_i^{k+1}} = \rho(W_i - B^{k+1}) + \alpha_i^k. \tag{A.9}$$

Thus,

$$W_i^{k+1} = -\frac{1}{\rho}\alpha_i^k + B^{k+1}. \tag{A.10}$$

Finally, as for equation A.10, following the steps in (Boyd et al., 2011), the first-order optimality condition using equation A.1c is expressed as

$$-Z_i^{-1} + \rho(Z_i - B^{k+1}J_i) + \lambda_i^k = 0. \tag{A.11}$$

Rewriting this, we get

$$\rho Z_i - Z_i^{-1} = \rho B^{k+1}J_i - \lambda_i^k. \tag{A.12}$$

First, take the orthogonal eigenvalue decomposition of the right-hand side,

$$\rho B^{k+1}J_i - \lambda_i^k = Q\Lambda Q^T, \tag{A.13}$$

where $\Lambda = \mathbf{diag}(\nu_1, \ldots, \nu_d)$, and $Q^TQ = QQ^T = I$. Multiplying equation A.12 by Q^T on the left and by Q on the right gives

$$\rho\tilde{Z}_i - \tilde{Z}_i^{-1} = \Lambda, \tag{A.14}$$

where $\tilde{Z}_i = Q^TZ_iQ$. A diagonal solution of this equation is given by

$$\tilde{Z}_{i,(jj)} = \frac{\nu_j + \sqrt{\nu_j^2 + 4\rho}}{2\rho}, \tag{A.15}$$

and the final solution is given as

$$Z_i^{k+1} = Q\tilde{Z}_iQ^T. \tag{A.16}$$

A.2 Derivation of Knothe-Rosenblatt ADMM Formulation and Final Updates. In similar fashion, here we outline the derivation of the ADMM formulation from section 4.3.

First, we note that the closed-form updates for W_i and p_i are identical to for the original formulation. Here we show the derivation only for the remainder of updates. In what follows, ADMM iteration superscripts, k, are

now enclosed in parentheses so as not to confuse them with the d superscript indexing over dimension:

The cost function $C(B^{(k+1)})$ is given by

$$\begin{aligned} C(B^{(k+1)}) = \frac{1}{N}\sum_{i=1}^{N} & \frac{1}{2}\rho||W_i^{(k)} - B||_2^2 + \theta||B\Phi_i - X_i||_2^2 \\ & + \frac{1}{2}\rho||B\Phi_i - p_i^{(k)}||_2^2 + \gamma_i^{(k)T}(p_i^{(k)} - B\Phi_i) + \\ & + \mathbf{tr}(\alpha_i^{(k)T}(W_i^{(k)} - B)) \\ & + \sum_{d=1}^{D} \frac{1}{2}\rho||B\Phi_i^d - Y_i^{d(k)}||_2^2 + \lambda_i^{d(k)T}(Y_i^{d(k)} - B\Phi_i^d). \end{aligned} \tag{A.17}$$

Taking the first-order derivative of equation A.17 and setting to 0, we arrive at the following expression:

$$\begin{aligned} & B\left[\rho(\mathbf{I} + \frac{1}{N}\sum_{i=1}^{N}\Phi_i\Phi_i^T + \frac{2\theta}{\rho}\Phi_i\Phi_i^T + \sum_{d=1}^{D}\Phi_i^d\Phi_i^{dT})\right] \\ & = \frac{1}{N}\sum_{i=1}^{N}\rho W_i^{(k)} + \rho p_i^{(k)}\Phi_i^T + 2\theta X_i\Phi_i^T + \gamma_i^{(k)}\Phi_i^T + \alpha_i^{(k)T} \\ & \quad + \sum_{d=1}^{D}\rho Y_i^{d(k)}\Phi_i^{dT} + \lambda_i^{d(k)}\Phi_i^{dT}. \end{aligned} \tag{A.18}$$

If we define

$$\mathcal{B}_s \triangleq \rho\left(\mathbf{I} + \frac{1}{N}\sum_{i=1}^{N}\Phi_i\Phi_i^T + \frac{2\theta}{\rho}\Phi_i\Phi_i^T + \sum_{d=1}^{D}\Phi_i^d\Phi_i^{dT}\right) \tag{A.19}$$

and

$$\begin{aligned} \mathcal{B}_i \triangleq & \frac{1}{N}\sum_{i=1}^{N}\rho W_i^{(k)} + \rho p_i^{(k)}\Phi_i^T + 2\theta X_i\Phi_i^T + \gamma_i^{(k)}\Phi_i^T + \alpha_i^{(k)T} \\ & + \sum_{d=1}^{D}\rho Y_i^{d(k)}\Phi_i^{dT} + \lambda_i^{d(k)}\Phi_i^{dT}, \end{aligned} \tag{A.20}$$

we have

$$B^{(k+1)} = \mathcal{B}_i \cdot \mathcal{B}_s^{-1}. \tag{A.21}$$

The loss function associated with Z_i^d for a given i and d is

$$C(Z_i^{d(k+1)}) = -\log Z_i^d + \frac{1}{2}\rho(Y_i^{d(k)}\mathbf{1}_d - Z_i^d)^2 + \beta_i^{d(k)}(Z_i^d - Y_i^{d(k)}\mathbf{1}_d).$$

Taking the derivative and setting to zero, we get the following quadratic expression:

$$\rho Z_i^{d2} + (\beta_i^{d(k)} - \rho Y_i^{d(k)}\mathbf{1}_d)Z_i^d - 1 = 0. \tag{A.22}$$

As we would like $Z_i^{d(k+1)}$ to be greater than zero according to our constraints, we set the closed-form solution to the positive root of this quadratic equation:

$$Z_i^{d(k+1)} = \frac{\rho Y_i^{d(k)}\mathbf{1}_d - \beta_i^{d(k)} + \sqrt{(\rho Y_i^{d(k)}\mathbf{1}_d - \beta_i^{d(k)})^2 + 4\rho}}{2\rho}. \tag{A.23}$$

The loss function associated with Y_i^d for a given i and d is

$$\begin{aligned} C(Y_i^{d(k+1)}) = {} & \frac{1}{2}\rho(Y_i^d\mathbf{1}_d - Z_i^{d(k+1)})^2 + \frac{1}{2}\rho||B^{(k+1)}\Phi_i^d - Y_i^d||_2^2 \\ & + \beta_i^{d(k)}(Z_i^{d(k+1)} - Y_i^d\mathbf{1}_d) + \lambda_i^{d(k)T}(Y_i^d - B^{(k+1)}\Phi_i^d). \end{aligned} \tag{A.24}$$

Taking the derivative with respect to Y_i^d and setting to zero, we get

$$\begin{aligned} Y_i^{d(k+1)} = {} & (\rho Z_i^{d(k+1)}\mathbf{1}_d^T + \rho B^{(k+1)}\Phi_i^d + \beta_i^{d(k)}\mathbf{1}_d^T - \lambda_i^{d(k)T}) \\ & \cdot (\rho\mathbf{1}_d\mathbf{1}_d^T + \rho\mathbf{I})^{-1}. \end{aligned} \tag{A.25}$$

Finally, our complete set of updates is:

$$B^{(k+1)} = \mathcal{B}_i \cdot \mathcal{B}_s, \tag{A.26a}$$

$$W_i^{(k+1)} = -\frac{1}{\rho}\alpha_i^{(k)} + B^{(k+1)}, \tag{A.26b}$$

$$Z_i^{d(k+1)} = \frac{\rho Y_i^{d(k)}\mathbf{1}_d - \beta_i^{d(k)} + \sqrt{(\rho Y_i^{d(k)}\mathbf{1}_d - \beta_i^{d(k)})^2 + 4\rho}}{2\rho}, \tag{A.26c}$$

$$\begin{aligned} Y_i^{d(k+1)} = {} & (\rho Z_i^{d(k+1)}\mathbf{1}_d^T + \rho B^{(k+1)}\Phi_i^d + \beta_i^{d(k)}\mathbf{1}_d^T - \lambda_i^{d(k)T}) \\ & \cdot (\rho\mathbf{1}_d\mathbf{1}_d^T + \rho\mathbf{I})^{-1}, \end{aligned} \tag{A.26d}$$

$$\gamma_i^{(k+1)} = \gamma_i^{(k)} + \rho(p_i^{(k+1)} - B^{(k+1)}\Phi_i), \tag{A.26e}$$

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} + \rho(W_i^{(k+1)} - B^{(k+1)}), \tag{A.26f}$$

$$\lambda_i^{d(k+1)} = \lambda_i^{d(k)} + \rho(Y_i^{d(k+1)} - B^{(k+1)}\Phi_i^d), \tag{A.26g}$$

$$\beta_i^{d(k+1)} = \beta_i^{d(k)} + \rho(Z_i^{d(k+1)} - Y_i^{d(k+1)}\mathbf{1}_d), \tag{A.26h}$$

$$p_i^{(k+1)} = \arg\min_{p_i} -\log q(p_i) + \text{pen}(p_i), \tag{A.26i}$$

where the p_i update can once again be performed using any number of appropriate optimization techniques.

A.3 Transport Map Multi-Indices Details. In this section, we give a few concrete examples of the various multi-index-sets presented in section 3.5 for clarification in practical use cases, as well as for actual implementation purposes.

In the case of a dense map, recall the index set:

$$\mathcal{J}^D = \left\{ \mathbf{j} \in \mathbb{N}^d : \sum_{i=1}^{d} j_i \leq O \right\}.$$

For example, in the case where $D = O = 3$, the resulting index set will have the following form:

$$\mathcal{J}^D = \begin{bmatrix} 0&1&2&3&0&0&0&1&1&2&0&0&0&1&1&2&0&0&1&0 \\ 0&0&0&0&1&2&3&1&2&1&0&1&2&0&1&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&2&2&2&3 \end{bmatrix},$$

where every jth column is one D-long multi-index for a single multivariate polynomial basis term, $\phi_{\mathbf{j}}$.

The size of this set $K \triangleq |\mathcal{J}^D|$ for any given maximum polynomial order O is

$$K = \binom{D+O}{O}.$$

In the case of the total order Knothe-Rosenblatt map, the index set is

$$\mathcal{J}_d^{KR} = \left\{ \mathbf{j} \in \mathbb{N}^d : \sum_{i=1}^{d} j_i \leq O \wedge j_i = 0, \forall i > d \right\}, d = 1, \ldots, D.$$

In this case, the size of the set $K_d \triangleq |\mathcal{J}_d^{KR}|$ becomes dependent on the component of the mapping.

Revisiting our previous example with $D = O = 3$, we have

$$\begin{aligned}
\mathcal{J}_1^{KR} &= \left\{ \mathbf{j} \in \mathbb{N}^3 : \sum_{i=1}^{3} j_i \leq O \wedge j_2 = j_3 = 0 \right\} \\
&= \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \\
\mathcal{J}_2^{KR} &= \left\{ \mathbf{j} \in \mathbb{N}^3 : \sum_{i=1}^{3} j_i \leq O \wedge j_3 = 0 \right\} \\
&= \begin{bmatrix} 0 & 1 & 2 & 3 & 0 & 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 3 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\
\mathcal{J}_3^{KR} &= \left\{ \mathbf{j} \in \mathbb{N}^3 : \sum_{i=1}^{3} j_i \leq O \right\} \\
&= \begin{bmatrix} 0 & 1 & 2 & 3 & 0 & 0 & 0 & 1 & 1 & 2 & 0 & 0 & 0 & 1 & 1 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 3 & 1 & 2 & 1 & 0 & 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 3 \end{bmatrix}.
\end{aligned}$$

In contrast to a dense mapping, this construction yields a weight matrix that has

$$|\mathcal{J}_d^{KR}| = \binom{d+O}{O} \tag{A.27}$$

many nonzero weights per row d, for a total of

$$\sum_{d=1}^{D} \binom{d+O}{O} \tag{A.28}$$

nonzero weights. In terms of implementation, note that we can enforce a lower-triangular structure of the mapping simply by constructing Φ according to the full index set ordering of $\mathcal{J}_D^{KR}$, and constraining the coefficient matrix W to have zeros embedded with the following structure:

Definition 1 *(lower-triangular weight matrix). A weight matrix $W \in \mathbb{R}^{D \times K}$ corresponds to a lower-triangular transport map if it can be expressed as*

$$W = \begin{bmatrix} \boldsymbol{w}_1^T & 0 & 0 & 0 & 0 & 0 & 0 \\ & \ldots & \boldsymbol{w}_d^T & \ldots & 0 & 0 & 0 \\ & & \ldots & \boldsymbol{w}_D^T & \ldots & & \end{bmatrix},$$

where each $\boldsymbol{w}_d$ is a vector in $\mathbb{R}^{|\mathcal{J}_d^{KR}|}$.

When constructed as such, $W\Phi_i = S(X_i)$, where S is a Knothe-Rosenblatt map.

In the case of the single univariate Knothe-Rosenblatt map, the index set becomes the following subset of $\mathcal{J}^{KR}$, again dependent on the component d:

$$\mathcal{J}_d^{KRSV} = \left\{ \mathbf{j} \in \mathbb{N}^d : \sum_{i=1}^{d} j_i \leq O \wedge j_i j_l = 0, \forall i \neq l \wedge j_i = 0, \forall i > d \right\},$$
$$d = 1, \ldots, D.$$

Revisiting our previous example with $D = O = 3$, we have the following multi-index sets:

$$\begin{aligned}
\mathcal{J}_1^{KRSV} &= \left\{ \mathbf{j} \in \mathbb{N}^3 : \sum_{i=1}^{3} j_i \leq O \wedge j_2 = j_3 = 0 \wedge j_i j_l = 0, \forall i \neq l \right\} \\
&= \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \\
\mathcal{J}_2^{KRSV} &= \left\{ \mathbf{j} \in \mathbb{N}^3 : \sum_{i=1}^{3} j_i \leq O \wedge j_3 = 0 \wedge j_i j_l = 0, \forall i \neq l \right\} \\
&= \begin{bmatrix} 0 & 1 & 2 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \\
\mathcal{J}_3^{KRSV} &= \left\{ \mathbf{j} \in \mathbb{N}^3 : \sum_{i=1}^{3} j_i \leq O \wedge j_i j_l = 0, \forall i \neq l \right\} \\
&= \begin{bmatrix} 0 & 1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 \end{bmatrix}.
\end{aligned}$$

Here, all multivariate polynomial basis terms that are a product of mixed univariate polynomial terms are eliminated from the basis, resulting in a weight matrix that has

$$|\mathcal{J}_d^{KRSV}| = dO + 1 \tag{A.29}$$

many nonzero weights per row d, for a total of

$$\sum_{d=1}^{D} dO + 1 \tag{A.30}$$

nonzero weights. In terms of implementation, the 0-embedding strategy from the total order Knothe-Rosenblatt mapping still applies as long as the complete index set is constructed as $\mathcal{J}_D^{KRSV}$.

A.4 Ensuring Diffeomorphism Properties of Parameterized Maps. For any $\tilde{S} \in \mathcal{D}_+$ parameterized as in section 3.3,

$$\tilde{S}_K(x) = W\Phi(x). \tag{A.31}$$

We must ensure that $WJ_\Phi(x)$ is positive-definite for all $\mathbf{x} \in \mathsf{W}$. Here, we define an additional optimization problem to ensure this property. We begin with the Euclidean projection or the proximal operator of the indicator function of $\mathcal{D}_+$:

$$S_W(x) = \underset{m(x)=W\Phi(\mathbf{x}):J_\Phi(\mathbf{x})\geq 0}{\arg\min} ||m(x) - W\Phi(\mathbf{x})||^2. \tag{A.32}$$

As such, S_W retains the properties of a diffeomorphism.

A.5 Inverse Map Details. Computing the inverse map also becomes straightforward given the above methodology of representing B and Φ_i.

We begin by first showing the Knothe-Rosenblatt property of the map in the complete forward-map equation assuming we are using our polynomial basis representation for a given X_i:

$$\underbrace{\begin{bmatrix} b_{11} & b_{12} & \dots & b_{1(K_1)} & \dots & 0 & 0 & 0 \\ b_{21} & b_{22} & \dots & \dots & b_{2(K_2)} & \dots & 0 & 0 \\ \vdots & & & & & & & \\ b_{D1} & b_{D2} & \dots & \dots & \dots & \dots & \dots & b_{D(K_D)} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} \Phi(X_i^1) \\ \Phi(X_i^1, X_i^2) \\ \vdots \\ \Phi(X_i^1, \dots, X_i^D) \end{bmatrix}}_{\Phi_i}$$

$$= \begin{bmatrix} S(X_i^1) \\ S(X_i^2) \\ \vdots \\ S(X_i^D) \end{bmatrix}, \tag{A.33}$$

where X_i^d represents the dth component of the ith sample.

Here, to fulfill our KR assumption, we assume that Φ_i is a column vector of the polynomial bases evaluated at X_i, ordered according to how many components of X_i the bases are a function of. That is, if $K_d = |\mathcal{J}_d^{KR}|$, then $\Phi(X_i^1)$ are the first K_1 basis functions that are only a function of X_1, $\Phi(X_i^1, X_i^2)$ are the $K_2 - K_1$ basis functions that are only a function of X_1 and X_2, and so on. As such, as only the first K_d elements of every dth row of B are (potentially) nonzero, the map should have the appropriate Knothe-Rosenblatt structure by construction.

In the case where we want to invert a sample $S(X_i)$, this defines a system of equations that can be solved row by row for each component of the solution, $S(X_i^d)$, in the form of a polynomial root-finding problem for each row. For example, we first solve for X_i^1, the solution of which we can call X_i^{1*} by finding the (single variable) root of

$$\begin{bmatrix} b_{11} & b_{12} & \dots & b_{1(K_1)} \end{bmatrix} \begin{bmatrix} | \\ \Phi(X_i^1) \\ | \end{bmatrix} = S(X_i^1). \tag{A.34}$$

Subsequently, we can solve for X_i^2 plugging X_i^{1*} into the second equation:

$$\begin{bmatrix} b_{21} & b_{22} & \dots & \dots & b_{2(K_2)} \end{bmatrix} \begin{bmatrix} | \\ \Phi(X_i^{1*}) \\ | \\ \Phi(X_i^{1*}, X_i^2) \\ | \end{bmatrix} = S(X_i^2), \tag{A.35}$$

and so on. Note that this results in D-many single variable root-finding problems per sample to invert, and the order of the polynomial that must be solved for will be equal to the order of the polynomial chosen to represent the basis.

Acknowledgments

We thank the Amazon Web Services Cloud Credits for Research Program for their partial and continued funding of this project with respect to cloud computing resources needed to push the current incarnation of the algorithm to the fullest. We also thank Gabe Schamberg and Alexis Allegra for their useful discussions and comments. Finally, we thank the anonymous reviewers, whose comments greatly improved the content and presentation of this material.

References

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Mach. Learning*, *50*(1–2), 5–43.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein GAN*. arXiv:1701.07875.

Azadi, S., & Sra, S. (2014). Towards an optimal stochastic alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, *32*, 620–628.

Bagnoli, M., & Bergstrom, T. (2005). Log-concave probability and its applications. *Economic Theory*, *26*(2), 445–469.

Bakry, D., & Émery, M. (1985). Diffusions hypercontractives. In J. Azema & M. Yor (Eds.), *Séminaire de probabilités xix 1983/84* (pp. 177–206). Berlin: Springer.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., & Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, *37*(2), A1111–A1138. doi:10.1137/141000439

Benamou, J.-D., Carlier, G., & Laborde, M. (2015). An augmented Lagrangian approach to Wasserstein gradient flows and applications. *ESAIM*, *54*, 1–17.

Bernardo, J. M., & Smith, A. F. (2001). *Bayesian theory*. Bristol, U.K.: IOP Publishing.

Bobkov, S., & Madiman, M. (2011). Concentration of the information in data with log-concave distributions. *Annals of Probability*, *39*, 1528–1543.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. *SIAM Journal on Mathematical Analysis*, *45*(1), 64–87. doi:10.1137/120874850

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, *3*(1), 1–122.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Brenier, Y. (1987). Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, *305*, 805–808.

Claici, S., Chien, E., & Solomon, J. (2018). *Stochastic Wasserstein barycenters*. arXiv:1802.05757.

Cuturi, M., & Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *Proceedings of the International Conference on Machine Learning* (pp. 685–693).

El Moselhy, T. A., & Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics*, *231*(23), 7815–7850. doi:10.1016/j.jcp.2012.07.022

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *bayesian data analysis* (Vol. 2). Boca Raton, FL: CRC Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, *6*, 721–741.

Genevay, A., Cuturi, M., Peyré, G., & Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems*, *29* (pp. 3440–3448). Red Hook, NY: Curran.

Genevay, A., Peyré, G., & Cuturi, M. (2017). *GAN and VAE from an optimal transport point of view*. arXiv:1706.01807.

Gilbert, A. C., Zhang, Y., Lee, K., Zhang, Y., & Lee, H. (2017). *Towards understanding the invertibility of convolutional neural networks*. arXiv:1705.08664.

Gilks, W. R. (2005). *Markov chain Monte Carlo*. New York: Wiley.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, (pp. 2672–2680). Red Hook, NY: Curran.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, *96*(4), 835–845.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Jordan, R., Kinderlehrer, D., & Otto, F. (1997). Free energy and the Fokker-Planck equation. *Physica D*, *107*, 265–271.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998a). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, *29*(1), 1–17.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998b). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, *29*(1), 1–17. doi:10.1137/S0036141096303359

Kim, S., Ma, R., Mesa, D., & Coleman, T. P. (2013). Efficient Bayesian inference methods via convex optimization and optimal transport. In *Proceedings of the 2013 IEEE International Symposium on Information Theory*. Piscataway, NJ: IEEE.

Kim, S., Mesa, D., Ma, R., & Coleman, T. P. (2015). *Tractable fully Bayesian inference via convex optimization and optimal transport theory*. arXiv:1509.08582.

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes*. arXiv:1312.6114.

Larochelle, H., & Murray, I. (2011). The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, *15*, 29–37.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, Y., Swersky, K., & Zemel, R. (2015). Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 1718–1727).

Lipton, Z. C., & Tripathi, S. (2017). *Precise recovery of latent vectors from generative adversarial networks*. arXiv:1702.04782.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Berlin: Springer.

Ma, R., & Coleman, T. (2011). Generalizing the posterior matching scheme to higher dimensions via optimal transportation. In *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*. Piscataway, NJ: IEEE.

Ma, R., & Coleman, T. (2014). Necessary and sufficient conditions for reliability of posterior matching in arbitrary dimensions. In *Proceedings of the IEEE International Symposium on Information Theory*. Piscataway, NJ: IEEE.

Marzouk, Y. M., Moselhy, T., Parno, M., & Spantini, A. (2016). *An introduction to sampling via measure transport*. http://arxiv.org/abs/1602.05023

Mesa, D., Kim, S., & Coleman, T. (2015). A scalable framework to transform samples from one continuous distribution to another. In *Proceedings of the 2015 IEEE International Symposium on Information Theory* (pp. 676–680). Piscataway, NJ: IEEE.

Papadakis, N., Peyré, G., & Oudet, E. (2014). Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, *7*(1), 212–238. doi:10.1137/130920058

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Parno, M., & Marzouk, Y. M. (2014). *Transport map accelerated Markov chain Monte Carlo*. ArXiv. http://arxiv.org/abs/1412.5492

Parno, M., Moselhy, T., & Marzouk, Y. M. (2016). A multiscale strategy for Bayesian inference using transport maps. *SIAM/ASA Journal on Uncertainty Quantification*, *4*(1), 1160–1190. doi:10.1137/15M1032478

Rezende, D. J., & Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning* (pp. 1530–1538). http://arxiv.org/abs/1505.05770

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods* (Vol. 319). Citeseer.

Salimans, T., Zhang, H., Radford, A., & Metaxas, D. (2018). *Improving gans using optimal transport*. arXiv:1803.05573.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Basel: Birkäuser.

Schoutens, W. (2000). *Stochastic processes and orthogonal polynomials*. Berlin: Springer-Verlag.

Sivia, D., & Skilling, J. (2006). *Data analysis: A Bayesian tutorial*. New York: Oxford University Press.

Spantini, A., Bigoni, D., & Marzouk, Y. M. (2016). Variational inference via decomposable transports: Algorithms for Bayesian filtering and smoothing. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

Srivastava, S., Li, C., & Dunson, D. B. (2015). *Scalable Bayes via barycenter in Wasserstein space*. arXiv:1508.05880.

Tantiongloc, J., Mesa, D., Ma, R., Kim, S., Alzate, C. H., Camacho, J. J., . . . Coleman, T. P. (2017). An information and control framework for optimizing user-compliant human computer interfaces. *Proceedings of the IEEE*, *102*(2), 273–285.

Tolstikhin, I., Bousquet, O., Gelly, S., & Schoelkopf, B. (2017). *Wasserstein auto-encoders*. arXiv:1711.01558.

Villani, C. (2003). *Topics in optimal transportation*. Providence, RI: American Mathematical Society.

Villani, C. (2008). *Optimal transport: Old and new*. Berlin: Springer.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, *151*(1), 3–34.

Zhong, W., & Kwok, J. (2014). Fast stochastic alternating direction method of multipliers. *Journal of Machine Learning Research*, *32*, 46–54. http://arxiv.org/abs/1308.3558

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso. *Annals of Statistics*, *35*(5), 2173–2192.

Received January 24, 2018; accepted November 29, 2018.