Adaptive construction of measure transports for Bayesian inference

Daniele Bigoni
MIT
Cambridge, MA 02139
dabi@mit.edu

Alessio Spantini
MIT
Cambridge, MA 02139
spantini@mit.edu

Youssef Marzouk
MIT
Cambridge, MA 02139
ymarz@mit.edu

Abstract

Measure transport provides a useful tool for characterizing multivariate non-Gaussian target distributions arising in Bayesian inference. The transport approach seeks a parametric map that pushes forward a chosen reference distribution to the target/posterior distribution, through minimization of a certain Kullback–Leibler divergence. Among the distinguishing features of this approach is the availability of a tractable error estimator for posterior approximation, along with the idea that transport can be cast as an infinite-dimensional optimization problem whose variations can be evaluated in closed form. We use these ingredients to develop a method for *adaptively* constructing transport maps—balancing the complexity of the map representation, approximation error, and computational cost.

1 Statistical inference as a measure transportation problem

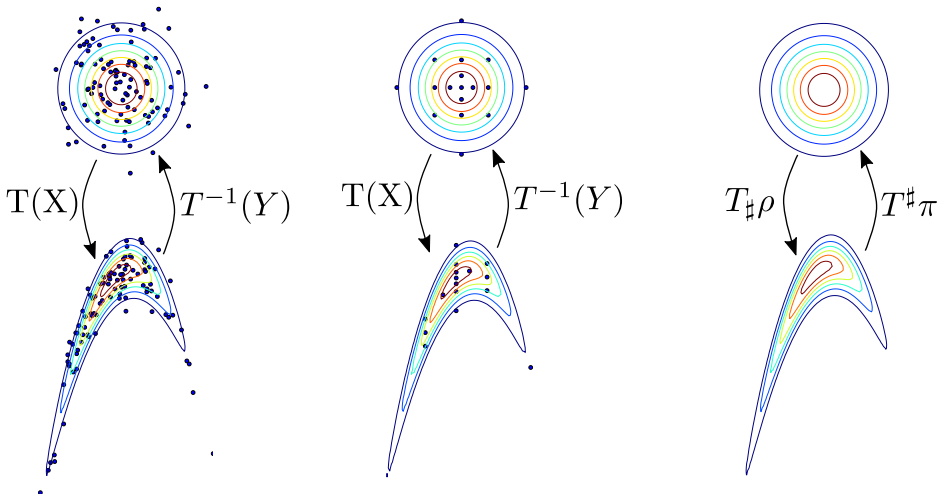
The solution of many statistical inference problems requires the evaluation of integrals $I[f] := \int f(\boldsymbol{x})P(d\boldsymbol{x})$ with respect to complex distributions P. These distributions occur, for example, as posteriors resulting from the application of Bayes' rule. A key challenge in this context is the creation of effective quadrature schemes for arbitrary posterior distributions. Here we use the term "quadrature" broadly to include Monte Carlo, quasi-Monte Carlo (QMC), and a variety of structured (e.g., sparse grid) numerical integration schemes.

We will frame this problem in the context of transportation of measures. Given the sample space $\mathbb{R}^d$ and the Borel σ-algebra $\sigma(\mathbb{R}^d)$, let $R : \sigma(\mathbb{R}^d) \to \mathbb{R}$ be a tractable distribution,[1] called the *reference*, and let $P : \sigma(\mathbb{R}^d) \to \mathbb{R}$ be the intractable distribution of interest, called the *target*. The transportation problem consists in finding the measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ such that $R(A) = P(T(A))$ for any $A \in \sigma(\mathbb{R}^d)$; in other words, T *pushes forward* R to P, which we write as $T_\sharp R = P$. In the classical treatments of optimal transportation by Monge and later Kantorovich [1, 2], the map is required to minimize a cost that reflects transportation effort. Such a cost is not intrinsic to most Bayesian inference problems, however, and thus our solution of the transportation problem will not be required to fulfill this optimality condition. Rather than seeking optimal maps, we will seek maps that satisfy the coupling condition $T_\sharp R = P$ along with certain additional structure discussed in Section 2.

In the context of statistical inference, explicit availability of a transport map allows the accurate and cheap evaluation of integrals with respect to P, by noting that $I[f] = \int f(\boldsymbol{x})P(d\boldsymbol{x}) = \int f \circ T(\boldsymbol{x})R(d\boldsymbol{x})$. Furthermore, if $(\mathbf{x}_k, \mathbf{w}_k)_{k=1}^q$ is a quadrature rule for R, then $(T(\mathbf{x}_k), \mathbf{w}_k)_{k=1}^q$ is a quadrature rule for P that can be used to approximate $I[f]$. Figures 1a and 1b illustrate the notions of transforming quadratures and mapping between distributions.

[1] Here tractability refers to the ready availability of quadrature rules for R. A quadrature rule is a collection of points and weights defining the approximation $\int gR(d\boldsymbol{x}) \approx \sum g(\mathbf{x}_k)\mathbf{w}_k$. Here we restrict our attention to $R = \mathcal{N}(\mathbf{0}, \mathbf{I})$, but we stress the fact that R is a degree of freedom.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.



(a) Monte Carlo sampling (b) Sparse grid quadrature (c) Density transformation

Figure 1: Using measure transport to characterize intractable distributions. Figures (a) and (b) show the mapping of Monte Carlo points and a sparse grid rule, respectively, from a standard Gaussian distribution to a complex target. Figure (c) shows the associated density transformations: pushing forward the reference (3) and pulling back the target (4).

2 Construction of transport maps

In order to make the identification of a transport T computationally tractable, we will restrict our attention to the set $\mathcal{T}_\triangle$ of lower triangular monotone increasing maps, known as the Knothe–Rosenblatt rearrangements [3, 4]. While this restriction greatly reduces the size of the search space for T, it does not rule out the existence of a solution. In fact, for any two absolutely continuous measures R and P, there exists a unique $T \in \mathcal{T}_\triangle$ such that $R(A) = P(T(A))$ for any $A \in \sigma(\mathbb{R}^d)$. Maps in $\mathcal{T}_\triangle$ take the form

$$T(\boldsymbol{x}) = \begin{bmatrix} T_1(x_1) \\ T_2(x_1, x_2) \\ \vdots \\ T_d(x_1, \ldots, x_d) \end{bmatrix}, \tag{1}$$

and satisfy a monotonicity condition corresponding to $\partial_i T_i > 0$, $i = 1 \ldots d$. This condition can be enforced by setting

$$T_i(x_1, \ldots, x_i) = c_i(x_1, \ldots, x_{i-1}) + \int_0^{x_i} \exp(h_i(x_1, \ldots, x_{i-1}, t))dt \,, \tag{2}$$

for functions $c_i : \mathbb{R}^{i-1} \to \mathbb{R}$ and $h_i : \mathbb{R}^i \to \mathbb{R}$ [5, 6].

As it is common in statistical inference, we will assume that R and P are absolutely continuous with respect to the Lebesgue measure and denote their densities by ρ and π. This implies that finding a map T that pushes forward R to P corresponds to satisfying $T_\sharp\rho = \pi$ almost everywhere (or equivalently $\rho = T^\sharp\pi$ a.e.), where

$$T_\sharp\rho(\boldsymbol{x}) := \rho \circ T^{-1}(\boldsymbol{x}) \det(\nabla T^{-1}(\boldsymbol{x})) \text{ and} \tag{3}$$

$$T^\sharp\pi(\boldsymbol{x}) := \pi \circ T(\boldsymbol{x}) \det(\nabla T(\boldsymbol{x})) \,, \tag{4}$$

are the pushforward of ρ through T and the pullback of π through T, respectively.[2]

The problem is then formalized as a minimization problem in terms of the Kullback-Leibler divergence from π to $T_\sharp\rho$ [7, 8]:

$$T^\star = \arg\min_{T \in \mathcal{T}_\triangle} \mathcal{D}_{KL}(T_\sharp\rho \| \pi) \,. \tag{5}$$

[2]For convenience, we apply the pushforward and pullback notations to both measures and densities.

In practice the map T needs to be parameterized. We will denote a finite-dimensional parameterization by $T[\mathbf{a}]$ for some coefficients $\mathbf{a} \in \mathbb{R}^n$. In the same fashion we will denote the parameterizations of T_i, c_i, and h_i by $T_i[\mathbf{a}_i]$, $c_i[\mathbf{a}_i^c]$, and $h_i[\mathbf{a}_i^h]$, respectively. This parameterization will define the n-dimensional subspace $\mathcal{T}_\triangle^n \subset \mathcal{T}_\triangle$ and lead to the minimization problem:

$$\mathbf{a}_n^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \mathcal{D}_{KL}(T[\mathbf{a}]_\sharp \rho \| \pi) \; . \tag{6}$$

Problem (6) is a stochastic optimization problem. We solve it using a sample average approximation approach [9]. Note that $\mathcal{D}_{KL}(T[\mathbf{a}]_\sharp\rho\|\pi) = \mathcal{D}_{KL}(\rho\|T[\mathbf{a}]^\sharp\pi)$. Then the resulting expectation with respect to R can be approximated by a Monte Carlo estimator,[3] leading to the deterministic problem

$$\mathbf{a}_n^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^n} \underbrace{\frac{1}{q}\sum_{k=1}^{q} -\log T[\mathbf{a}]^\sharp\bar{\pi}(\mathbf{x}_k)}_{\mathcal{J}_n^q(T[\mathbf{a}])} \; , \tag{7}$$

for a sample $(\mathbf{x}_i)_{i=1}^q$ with $\mathbf{x}_i \overset{\text{iid}}{\sim} R$. Above, we have also replaced the target density π with its (potentially) unnormalized counterpart $\bar{\pi}$, without affecting the minimizer. Problem (7) can now be solved with any nonlinear optimization method of choice. The availability of information such as the gradient and, optionally, the Hessian of $\bar{\pi}$ enables the use of high-order (e.g., Newton or quasi-Newton) optimization methods which guarantee fast convergence. Furthermore, the objective in (7) allows for embarrassingly parallel implementations.

One very useful property of the method is that it allows for an estimation of the posterior approximation error.[4] In particular, as the approximation improves, $\mathbb{V}_\rho[\log \frac{\rho}{T^\sharp\bar{\pi}}] \to 0$ at the same asymptotic rate at which $\mathcal{D}_{KL}(T_\sharp\rho\|\pi) \to \mathcal{D}_{KL}(T_\sharp^\star\rho\|\pi)$, where $T^\star \in \mathcal{T}_\triangle$ is the global minimizer of (5). Hence, despite the objective of (7) not being zero at optimality due to the unknown integration constant of $\bar{\pi}$, the *variance diagnostic* $\mathbb{V}_\rho[\log \frac{\rho}{T^\sharp\bar{\pi}}]$ can be used to assess the quality of the approximation [7].

3 Adaptive construction of transport maps

Problem (5) is defined over the infinite-dimensional space $\mathcal{T}_\triangle$. Finding a good parameterization for T corresponds to finding the finite-dimensional subspace $\mathcal{T}_\triangle^n \subset \mathcal{T}_\triangle$ that leads to the smallest error (i.e., smallest KL divergence at optimality) for a given dimension n. The quality of this subspace is dictated by the sources of low-dimensional structure in the problem. Here we develop an adaptive strategy that leverages infinite-dimensional information to drive the enrichment of the map parameterization, focusing on the smoothness and marginal independence properties of π. Other sources of low dimensionality, such as conditional independence and low-rank structure, are analyzed in [10].

We first note that if π encodes some degree of marginal independence between components of $\mathbf{X}$, the parameterization of a map component T_i in (1) will involve only a subset of its inputs. In particular, let $\mathbf{X} \sim P$ and (A, B) be a partition of $\{1, \dots, i\}$. If $\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_B$ for every $i \in A$, then T_i is a $|A|$–dimensional function of the variables $\boldsymbol{x}_A$. This property suggests that one could start with a very sparse approximation of T (e.g., diagonal) and enrich it through the following adaptive procedure.

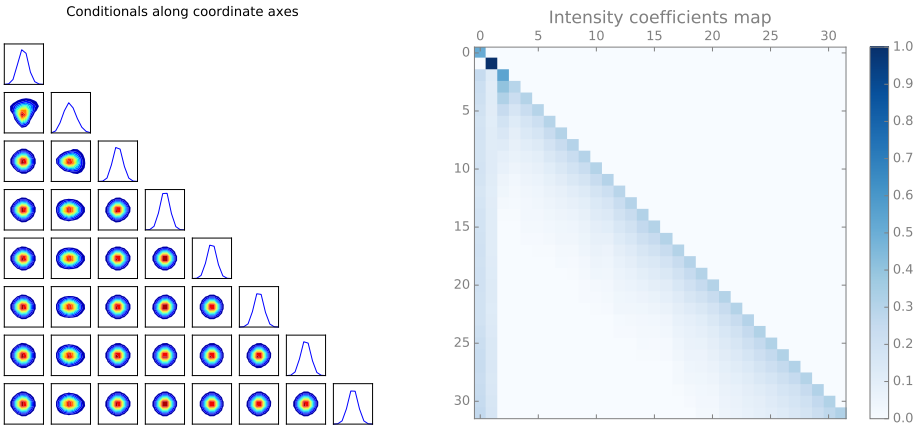
Let us consider problem (7) of finding $\mathbf{a}_n^\star$ such that $T[\mathbf{a}_n^\star] \in \mathcal{T}_\triangle^n \subset \mathcal{T}_\triangle$ minimizes $\mathcal{J}_n^q(T[\mathbf{a}])$. For simplicity, we consider only target distributions P with finite variance and therefore restrict our attention to square integrable maps, i.e., $\mathcal{T}_\triangle \subset \mathcal{H} := L^2_\rho$ and $\mathcal{T}_\triangle^n \subset \mathcal{H}^n \subset \mathcal{H}$, where $\mathcal{H}^n$ is an n-dimensional subspace of the Hilbert space $\mathcal{H}$. We know that, at optimality, $\nabla_{\mathbf{a}}\mathcal{J}_n^q(T[\mathbf{a}_n^\star]) = 0$. However the gradient identifying the first variation[5] of $\mathcal{J}_n^q(T[\mathbf{a}_n^\star])$,

$$\nabla\mathcal{J}_n^q(T[\mathbf{a}_n^\star]) = (\nabla_{\boldsymbol{x}}T)^{-1}\left(\nabla_{\boldsymbol{x}}\log\frac{\rho}{T^\sharp\bar{\pi}}\right) \; , \tag{8}$$

[3] One could instead approximate this expectation with QMC, sparse grid quadrature, or any other scheme.

[4] This is a non-trivial task in Bayesian inference, due to the unavailability of the integration constant of π.

[5] Recall that in the Hilbert space $\mathcal{H}$, the first variation given by $\delta\mathcal{J}_n^q(T[\mathbf{a}_n^\star])(R) = \lim_{\varepsilon\to 0}\frac{\mathcal{J}_n^q(T[\mathbf{a}_n^\star]+\varepsilon R)-\mathcal{J}_n^q(T[\mathbf{a}_n^\star])}{\varepsilon}$ is identified by the vector $\nabla\mathcal{J}_n^q(T[\mathbf{a}_n^\star]) \in \mathcal{H}$ such that $\delta\mathcal{J}_n^q(T[\mathbf{a}_n^\star])(R) = \langle\nabla\mathcal{J}_n^q(T[\mathbf{a}_n^\star]), R\rangle$. The Riesz representation theorem guarantees that this vector exists and is unique.



(a) Iteration 13 - $T^\sharp\pi$

(b) Iteration 13 - $\nabla_{\boldsymbol{x}} T$

Figure 2: Quality (left) and sparsity pattern (right) of the adaptively constructed transport map for the stochastic volatility problem, after 13 iterations.

is *itself* a map that will be different from zero, unless $T[\mathbf{a}_n^\star]$ represents a global optimum in the infinite-dimensional space $\mathcal{T}_\triangle$. This gradient can be evaluated in closed form (8), and identifies a direction in $\mathcal{H}$ such that there exists an ε-neighborhood where the objective can be improved, i.e., $\mathcal{J}_n^q(T[\mathbf{a}_n^\star] + \varepsilon \nabla \mathcal{J}_n^q(T[\mathbf{a}_n^\star])) < \mathcal{J}_n^q(T[\mathbf{a}_n^\star])$. Since $\nabla \mathcal{J}_n^q(T[\mathbf{a}_n^\star]) \in \mathcal{H}$ is a map, we can approximate it by

$$\mathbf{a}_m^\star = \arg\min_{\mathbf{a}\in\mathbb{R}^m} \|\nabla \mathcal{J}_n^q(T[\mathbf{a}_n^\star]) - U[\mathbf{a}]\|_{L_\rho^2}^2 \ , \tag{9}$$

where $U[\mathbf{a}] \in \mathcal{H}^m$ and $\mathcal{H}^n \subset \mathcal{H}^m$. Analyzing the magnitude of the coefficients $\mathbf{a}$, we can develop a heuristic to choose the new parameterization coefficients to be used in the new approximation space $\mathcal{T}_\triangle^l$, with $n \leq l \leq m$. Note that if we approximate the L_ρ^2 norm in (9) using the same quadrature rule used to evaluate (6), no additional evaluation of π is required for the computation of the gradient (8).

This adaptive procedure is terminated when the variance diagnostic $\mathbb{V}_\rho[\log \rho \,/\, T[\mathbf{a}_l^\star]^\sharp \bar{\pi}]$ reaches a user-defined tolerance. If the approximation is not satisfactory, exact samples from the target distribution can be drawn by applying importance sampling or MCMC [11] to the more amenable "Gaussianized" density $T[\mathbf{a}_n^\star]^\sharp \bar{\pi} \approx \rho$ [12, 13].

4 Numerical example

We are currently applying the adaptive map construction to a range of inference problems; here we show a simple example. Consider inference of the time-dependent volatility of an asset, given observations of its return at certain times. We use [14] an auto-regressive AR(1) process to model the log-volatility X_t of the asset at time t:

$$X_{t+1} = \mu + \phi(X_t - \mu) + \eta_t \ , \quad \eta_t \sim \mathcal{N}(0,1) \ , \quad X_1 \sim \mathcal{N}\left(0, 1/\left(1-\phi^2\right)\right) \ , \tag{10}$$

where the hyperparameters μ and ϕ are endowed with priors $\mu \sim \mathcal{N}(0, \sigma_\mu^2)$ and $\frac{\phi+1}{2} \sim \text{Beta}(10, 1)$. The observed return Y_t follows the price evolution model suggested by Black and Scholes [15]:

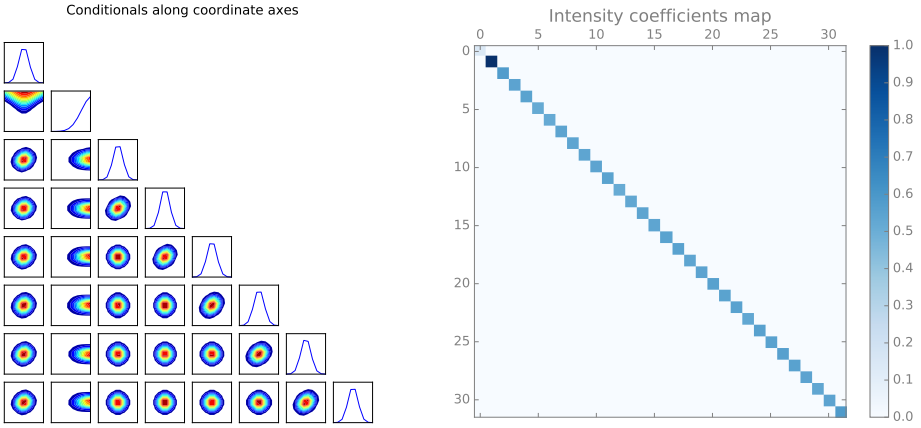
$$Y_t = \varepsilon_t \exp(X_t/2) \ , \quad \varepsilon_t \sim \mathcal{N}(0,1) \ . \tag{11}$$

We characterize the full Bayesian posterior $\pi \sim \mu, \phi, \mathbf{X}_{1:N} | \mathbf{Y}_{1:N}$ of the parameters and states up to time $N = 30$. To illustrate the capability of the method, Figure 2a shows slices (i.e., two-dimensional conditionals) of the pullback density $T[\mathbf{a}^\star]^\sharp \pi$ at the 13th step of the adaptation scheme. Since $T[\mathbf{a}^\star]^\sharp \pi \approx \rho$, these slices should resemble those of the standard Gaussian density if the map is accurate. Figure 2b shows the magnitudes of elements of $\nabla_{\boldsymbol{x}} T[\mathbf{a}^\star]$ at the same step, suggesting the sparsity pattern of $T[\mathbf{a}^\star]$. The number of coefficients identified by the adaptive scheme at step 13 is ≈ 700. The variance diagnostic $\mathbb{V}_\rho\left[\log \rho / T[\mathbf{a}^\star]^\sharp \bar{\pi}\right]$ is steadily decreased via the adaptive procedure, improving by one order of magnitude overall and suggesting a good overall agreement between $T[\mathbf{a}^\star]^\sharp \pi$ and ρ. Additional results are shown in Appendix A.

Acknowledgments

This work was supported by the US Department of Energy, Office of Advanced Scientific Computing (ASCR), under grant numbers DE-SC0010518 and DE-SC0009297.

A Additional numerical results for the example in Section 4



(a) Iteration 1 - $T^\sharp \pi$ (b) Iteration 1 - $\nabla_{\boldsymbol{x}} T$

Figure 3: Quality (left) and sparsity pattern (right) of the initial transport map used to start the adaptive procedure for the stochastic volatility problem. The initial map is chosen to be diagonal.

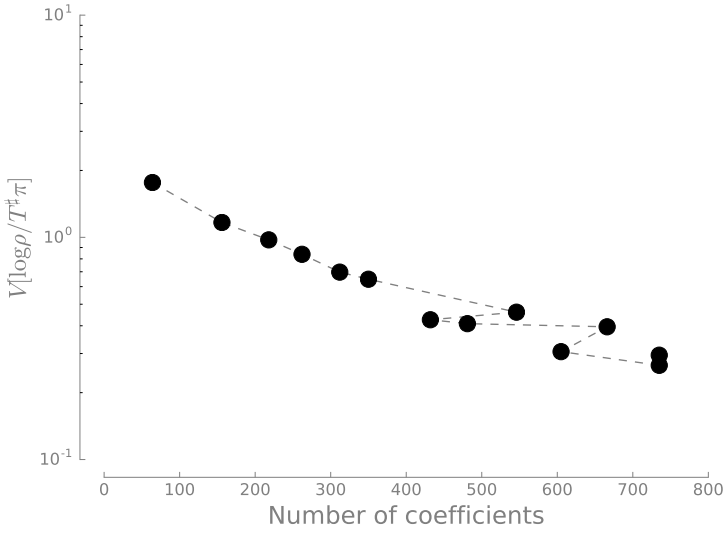


Figure 4: Decay of the variance diagnostic $\mathbb{V}_\rho \left[\log \frac{\rho}{T[\mathbf{a}_n^\star]^\sharp \bar{\pi}} \right]$ as coefficients are adaptively added and/or removed from the approximation $T[\mathbf{a}]$. The removal of coefficients is driven by an estimate of their sensitivities to the sample used for the Monte Carlo approximation in (7). For comparison, the variance diagnostic associated with the *Laplace approximation* of the posterior is approximately 10.

References

[1] G. Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*. 1781, pp. 666–704.

[2] L. V. Kantorovich. “On the Translocation of Masses”. In: *Journal of Mathematical Sciences* 133.4 (Mar. 2006), pp. 1381–1382. ISSN: 1072-3374. DOI: 10.1007/s10958-006-0049-2.

[3] Murray Rosenblatt. “Remarks on a Multivariate Transformation”. In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 470–472. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729394.

[4] Herbert Knothe. *Contributions to the Theory of Convex Bodies*. 1957. DOI: 10.1307/mmj/1028990175.

[5] J. O. Ramsay. “Estimating smooth monotone functions”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2 (1998), pp. 365–375. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00130.

[6] Daniele Bigoni, Alessio Spantini, and Youssef Marzouk. “On the computation of monotone transports”. In: *Preprint* (2016).

[7] Tarek A. El Moselhy and Youssef M. Marzouk. “Bayesian inference with optimal maps”. In: *Journal of Computational Physics* 231.23 (Oct. 2012), pp. 7815–7850. ISSN: 00219991. DOI: 10.1016/j.jcp.2012.07.022.

[8] Youssef Marzouk et al. “Sampling via Measure Transport: An Introduction”. In: *Handbook of Uncertainty Quantification*. Ed. by Roger G. Ghanem, David Higdon, and Houman Owhadi. Cham: Springer International Publishing, 2016, pp. 1–41. DOI: 10.1007/978-3-319-11259-6_23-1.

[9] Alexander Shapiro. “Sample Average Approximation”. In: *Encyclopedia of Operations Research and Management Science*. 3. Boston, MA: Springer US, 2013, pp. 1350–1355. ISBN: 978-1-4419-1137-7. DOI: 10.1007/978-1-4419-1153-7_1154.

[10] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. “Measure transport, variational inference, and low-dimensional maps”. In: *Preprint* (2016).

[11] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Ed. by Intergovernmental Panel on Climate Change. Vol. 1. Springer Texts in Statistics. New York, NY: Springer New York, 2004, pp. 1–30. ISBN: 978-1-4419-1939-7. DOI: 10.1007/978-1-4757-4145-2.

[12] Valero Laparra, Gustavo Camps-Valls, and Jesus Malo. “Iterative gaussianization: From ICA to random rotations”. In: *IEEE Transactions on Neural Networks* 22.4 (2011), pp. 537–549. ISSN: 10459227. DOI: 10.1109/TNN.2011.2106511.

[13] Matthew Parno and Youssef Marzouk. “Transport map accelerated Markov chain Monte Carlo”. In: *submitted* (2015).

[14] Sangjoon Kim, Neil Shephard, and Siddhartha Chib. “Stochastic volatility: likelihood inference and comparison with ARCH models”. In: *The Review of Economic Studies* 65.December 1994 (1998), pp. 361–393. ISSN: 0034-6527. DOI: 10.1111/1467-937X.00050.

[15] Fischer Black and Myron Scholes. “The Pricing of Options and Corporate Liabilities.” In: *Journal of Political Economy* 81.3 (1973), p. 637. ISSN: 00223808. DOI: 10.1086/260062.