Measure Transport with Kernel Stein Discrepancy

Matthew A. Fisher[1] **Tui H. Nolan**[2,3] **Matthew M. Graham**[1,4]
Dennis Prangle[1] **Chris. J. Oates**[1,4]
Newcastle University[1] Cornell University[2] University of Technology Sydney[3] Alan Turing Institute[4]

Abstract

Measure transport underpins several recent algorithms for posterior approximation in the Bayesian context, wherein a transport map is sought to minimise the Kullback–Leibler divergence (KLD) from the posterior to the approximation. The KLD is a strong mode of convergence, requiring absolute continuity of measures and placing restrictions on which transport maps can be permitted. Here we propose to minimise a kernel Stein discrepancy (KSD) instead, requiring only that the set of transport maps is dense in an L^2 sense and demonstrating how this condition can be validated. The consistency of the associated posterior approximation is established and empirical results suggest that KSD is a competitive and more flexible alternative to KLD for measure transport.

1 Introduction

A popular and constructive approach to approximation of complicated distributions is to learn a transformation from a simpler reference distribution. Within machine learning, neural networks are often used to provide flexible families of transformations which can be optimised by stochastic gradient descent on a suitable objective, with *variational autoencoders* (Kingma and Welling, 2013; Rezende et al., 2014), *generative adversarial networks* (Goodfellow et al., 2014), *generative moment matching networks* (Li et al., 2015; Dziugaite et al., 2015) and *normalizing flows* (Rezende and Mohamed, 2015; Kingma et al., 2016; Dinh et al., 2016; Papamakarios et al., 2019; Kobyzev et al., 2020) all fitting in this framework. The principal application for such generative models is *distribution estimation*; samples are provided from the target distribution and the task is to fit a distribution to these samples. Parallel developments within applied mathematics view the transformation as a *transport map* performing *measure transport* (Marzouk et al., 2016; Parno and Marzouk, 2018). The principal application for measure transport is *posterior approximation*; an un-normalised density function defines the complicated distribution and the task is to approximate it. In this paper we study posterior approximation, noting that the flexible transformations developed in the machine learning literature can also be applied to this task.

Measure transport provides a powerful computational tool for Bayesian inference in settings that can be challenging for standard approaches, such as Markov chain Monte Carlo (MCMC) or mean field variational inference. For example, even sophisticated MCMC methods can fail when a posterior is concentrated around a sub-manifold of the parameter space (Livingstone and Zanella, 2019; Au et al., 2020), while it can be relatively straight-forward to define a transport map whose image is the sub-manifold (Parno and Marzouk, 2018; Brehmer and Cranmer, 2020). Likewise, mean field variational inference methods can perform poorly in this context, since independence assumptions can be strongly violated (Blei et al., 2017).

Let $\mathcal{Y}$ be a measurable space equipped with a probability measure P, representing the posterior to be approximated. The task that we consider in this paper is to elicit a second measurable space $\mathcal{X}$, equipped with a probability measure Q, and a measurable function $T : \mathcal{X} \to \mathcal{Y}$, such that the push-forward $T_{\#}Q$ (i.e. the measure produced by applying T to samples from Q) approximates P, in a sense to be specified. It is further desired that Q should be a "simple" distribution that is easily sampled. In contrast to the literature on normalising flows, it is *not* stipulated that T should be a bijection, since we wish to allow for situations where $\mathcal{X}$ and $\mathcal{Y}$ have different cardinalities or where P is supported on a sub-manifold.

A natural starting point is a notion of *discrepancy*

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

$\mathcal{D}(P_1, P_2)$ between two probability measures, P_1 and P_2, on $\mathcal{Y}$, with the property that $\mathcal{D}(P_1, P_2) = 0$ if and only if P_1 and P_2 are equal. Then one selects a measurable space $\mathcal{X}$ and associated probability measure Q and seeks a solution to

$$\arg\min_{T \in \mathcal{T}} \mathcal{D}(P, T_{\#}Q), \tag{1}$$

over a suitable set $\mathcal{T}$ of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. A popular choice of $\mathcal{D}$ is the Kullback-Leibler divergence (KLD), giving rise to *variational inference* (Blei et al., 2017; Agrawal et al., 2020), but other discrepancies can be considered (Ranganath et al., 2016). The problem in (1) can be augmented to include also the selection of $\mathcal{X}$ and Q, if desired.

The solution of (1) provides an approximation to P whose quality will depend on the set $\mathcal{T}$ and the discrepancy $\mathcal{D}$. This motivates us to consider the choice of $\mathcal{T}$ and $\mathcal{D}$, taking into account considerations that go beyond computational tractability. For example, a desirable property would be that, for a sequence of probability measures $(P_n)_{n\in\mathbb{N}}$, if $\mathcal{D}(P, P_n) \to 0$ then $P_n \to P$ in some suitable sense. For $\mathcal{D} = \mathcal{D}_{\mathrm{KL}}$, the KLD[1], it holds that $\mathcal{D}_{\mathrm{KL}}(P, P_n) \to 0$ implies P_n converges to P in *total variation*, from Pinsker's inequality (Tsybakov, 2009). This is a strong mode of convergence, requiring absolute continuity of measures that may be difficult to ensure when the posterior is concentrated near to a sub-manifold. Accordingly, the use of KLD for measure transport places strong and potentially impractical restrictions on which maps T are permitted (e.g. Marzouk et al., 2016; Parno and Marzouk, 2018, required that T is a diffeomorphism with $\det\nabla T > 0$ on $\mathcal{X}$). This motivates us in this paper to consider the use of an alternative discrepancy $\mathcal{D}$, corresponding to a weaker mode of convergence, for posterior approximation using measure transport. The advantage of discrepancy measures inducing weaker modes of convergence has also motivated recent developments in generative adversarial networks (Arjovsky et al., 2017).

Our contributions are as follows:

- We propose kernel Stein discrepancy (KSD) as an alternative to KLD for posterior approximation using measure transport, showing that KSD renders (1) tractable for standard stochastic optimisation methods (Proposition 1).
- Using properties of KSD we are able to establish consistency under explicit and verifiable assumptions on P, Q and $\mathcal{T}$ (Theorem 2).
- Our theoretical assumptions are weak – we do not even require T to be a bijection – and are verified for a particular class of neural network (Proposition 3). In particular, we do not require Q and P to be defined on the same space, allowing quite flexible mappings T to be constructed.
- Empirical results support KSD as a competitive alternative to KLD for measure transport.

Earlier work on this topic appears limited to Hu et al. (2018), who trained a neural network with KSD. Here we consider general transport maps and we establish consistency of the method, which these earlier authors did not. We note also that gradient flows provide an alternative (implicit) approach to measure transport (Liu and Wang, 2016).

Outline: The remainder of the paper is structure as follows: Section 2 introduces measure transport using KSD, Section 3 contains theoretical analysis for this new method, Section 4 presents a detailed empirical assessment and Section 5 contains a discussion of our main findings.

2 Methods

This section introduces measure transport using KSD. In Section 2.1 and Section 2.2 we recall mathematical definitions from measure transport and Hilbert spaces, respectively; in Section 2.3 we recall the definition and properties of KSD; in Section 2.4 we formally define our proposed method, and in Section 2.5 we present some parametric families $\mathcal{T}$ that can be employed.

Notation: The set of probability measures on a measurable space $\mathcal{X}$ is denoted $\mathcal{P}(\mathcal{X})$ and a point mass at $x \in \mathcal{X}$ is denoted $\delta(x) \in \mathcal{P}(\mathcal{X})$. For $P \in \mathcal{P}(\mathcal{X})$ let $L^q(P) \coloneqq \{f : \mathcal{X} \to \mathbb{R} : \int f^q \mathrm{d}P < \infty\}$. For $P \in \mathcal{P}(\mathbb{R}^d)$ and $(P_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R}^d)$, let $P_n \Rightarrow P$ denote weak convergence of the sequence of measures $(P_n)_{n\in\mathbb{N}}$ to P. The Euclidean norm on $\mathbb{R}^n$ is denoted $\|\cdot\|$. Partial derivatives are denoted ∂_x. For a function $f : \mathbb{R}^n \to \mathbb{R}$ the gradient is defined as $[\nabla f]_i = \partial_{x_i} f$. For a function $f = (f_1, \ldots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$, the divergence is defined as $\nabla \cdot f = \sum_{i=1}^n \partial_{x_i} f_i$.

Our main results in this paper concern the Euclidean space $\mathbb{R}^d$, but in some parts of the paper, such as Section 2.1, it is possible to state definitions at a greater level of generality at no additional effort - in such situations we do so.

[1] We use the notation $\mathcal{D}_{\mathrm{KL}}(P, Q) \coloneqq \mathrm{KL}(Q||P)$.

2.1 Measure Transport

A *Borel space* $\mathcal{X}$ is a topological space equipped with its Borel σ-algebra, denoted $\Sigma_{\mathcal{X}}$. Throughout this paper we restrict attention to Borel spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $Q \in \mathcal{P}(\mathcal{X})$ and $P \in \mathcal{P}(\mathcal{Y})$. In the parlance of measure transport, Q is the *reference* and P the *target*. Let $T : \mathcal{X} \to \mathcal{Y}$ be a measurable function and define the *pushforward* of Q through T as the probability measure $T_{\#}Q \in \mathcal{P}(\mathcal{Y})$ that assigns mass $(T_{\#}Q)(A) = Q(T^{-1}(A))$ to each $A \in \Sigma_{\mathcal{Y}}$. Here $T^{-1}(A) = \{x \in \mathcal{X} : T(x) \in A\}$ denotes the pre-image of A under T. Such a function T is called a *transport map* from Q to P if $T_{\#}Q = P$.

Faced with a complicated distribution P, if one can express P using a transport map T and a distribution Q that can be sampled, then samples from P can be generated by applying T to samples from Q. This idea underpins elementary methods for numerical simulation of random variables (Devroye, 2013). However, in posterior approximation it will not typically be straightforward to identify a transport map and at best one can seek an *approximate* transport map, for which $T_{\#}Q$ approximates P in some sense to be specified. In this paper we seek approximations in the sense of KSD, which is formally introduced in Section 2.3 and requires concepts in Section 2.2, next.

2.2 Hilbert Spaces

A Hilbert space $\mathcal{H}$ is a complete inner product space; in this paper we use subscripts, such as $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, to denote the associated inner product. Given two Hilbert spaces $\mathcal{G}$, $\mathcal{H}$, the *Cartesian product* $\mathcal{G} \times \mathcal{H}$ is again a Hilbert space equipped with the inner product $\langle (g_1, h_1), (g_2, h_2) \rangle_{\mathcal{G}\times\mathcal{H}} \coloneqq \langle g_1, g_2 \rangle_{\mathcal{G}} + \langle h_1, h_2 \rangle_{\mathcal{H}}$. In what follows we let $\mathcal{B}(\mathcal{H}) \coloneqq \{h \in \mathcal{H} : \langle h, h \rangle_{\mathcal{H}} \leq 1\}$ denote the unit ball in a Hilbert space $\mathcal{H}$.

From the Moore–Aronszajn theorem (Aronszajn, 1950), any symmetric positive definite function $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ defines a unique *reproducing kernel Hilbert space* of real-valued functions on $\mathcal{Y}$, denoted $\mathcal{H}_k$ and with inner-product denoted $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$. Indeed, $\mathcal{H}_k$ is a Hilbert space characterised by the properties (i) $k(\cdot, y) \in \mathcal{H}_k$ for all $y \in \mathcal{Y}$, (ii) $\langle h, k(\cdot, y) \rangle_{\mathcal{H}_k} = h(y)$ for all $h \in \mathcal{H}_k$, $y \in \mathcal{Y}$. Reproducing kernels are central to KSD, as described next.

2.3 Kernel Stein Discrepancy

Stein discrepancies were introduced in Gorham and Mackey (2015) to provide a notion of discrepancy that is computable in the Bayesian statistical context. In this paper we focus on so-called *kernel Stein discrepancy* (KSD; Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017) since this has lower computational overhead compared to the original proposal of Gorham and Mackey (2015).

The construction of KSD relies on Stein's method (Stein, 1972) where, for a possibly complicated probability measure $P \in \mathcal{P}(\mathcal{Y})$ of interest, one identifies a *Stein set* $\mathcal{F}$ and a *Stein operator* $\mathcal{A}_P$, such that $\mathcal{A}_P$ acts on elements $f \in \mathcal{F}$ to return functions $\mathcal{A}_P f : \mathcal{Y} \to \mathbb{R}$ with the property that

$$P' = P \quad \text{iff} \quad \mathbb{E}_{Y \sim P'}[(\mathcal{A}_P f)(Y)] = 0 \; \forall f \in \mathcal{F} \tag{2}$$

for all $P' \in \mathcal{P}(\mathcal{Y})$. A *Stein discrepancy* uses the extent to which (2) is violated to quantify the discrepancy between P' and P:

$$\mathcal{D}_{\mathrm{S}}(P, P') \coloneqq \sup_{f \in \mathcal{F}} |\mathbb{E}_{Y \sim P'}[(\mathcal{A}_P f)(Y)]|$$

Note that $\mathcal{D}_{\mathrm{S}}$ is not symmetric in its arguments. For $\mathcal{Y} = \mathbb{R}^d$ and suitably regular P, which admits a positive and differentiable density function p, Liu et al. (2016); Chwialkowski et al. (2016) showed that one may take $\mathcal{F}$ to be a set of smooth vector fields $f : \mathbb{R}^d \to \mathbb{R}^d$ and $\mathcal{A}_P$ to be a carefully chosen differential operator on $\mathbb{R}^d$. More precisely, and letting $s_p \coloneqq \nabla \log p$, we have Theorem 1 below, which is due to Gorham and Mackey (2017, Theorem 7):

Definition 1 (Eberle (2015))**.** *A probability measure $P \in \mathcal{P}(\mathbb{R}^d)$ is called* distantly dissipative *if $\liminf_{r\to\infty} \kappa(r) > 0$, where*

$$\kappa(r) \coloneqq -r^{-2} \inf_{\|x-y\|=r} \langle s_p(x) - s_p(y), x - y \rangle.$$

Theorem 1. *Suppose that $P \in \mathcal{P}(\mathbb{R}^d)$ is distantly dissipative. For some $c > 0$, $\ell > 0$ and $\beta \in (-1, 0)$, let*

$$\mathcal{F} \coloneqq \mathcal{B}\left(\prod_{i=1}^d \mathcal{H}_k\right), \quad k(x, y) \coloneqq (c^2 + \|\tfrac{x-y}{\ell}\|^2)^{\beta} \tag{3}$$

$$\mathcal{A}_P f \coloneqq f \cdot \nabla \log p + \nabla \cdot f. \tag{4}$$

Then (2) holds. Moreover, if $\mathcal{D}_{\mathrm{S}}(P, P_n) \to 0$, then $P_n \Rightarrow P$.

The kernel k appearing in (3) is called the *inverse multi-quadric* kernel. It is known that the elements of $\mathcal{H}_k$ are smooth functions, which justifies the application of the differential operator. The last part of Theorem 1 clarifies why KSD is useful; convergence in KSD controls the standard notion of weak convergence of measures to P.

KSD, in contrast to KLD, is well-defined when the approximating measure P' and the target P differ in their support. Moreover, in some situations KSD can be exactly computed: from Liu et al. (2016, Theorem 3.6) or equivalently Chwialkowski et al. (2016, Theorem 2.1),

$$\mathcal{D}_{\mathrm{S}}(P, P') = \sqrt{\mathbb{E}_{Y,Y' \sim P'}[u_p(Y, Y')]} \tag{5}$$

$$u_p(y, y') \coloneqq s_p(y)^\top k(y, y') s_p(y') + s_p(y)^\top \nabla_{y'} k(y, y') + \nabla_y k(y, y')^\top s_p(y') + \nabla_y \cdot \nabla_{y'} k(y, y'). \quad (6)$$

It follows that KSD can be exactly computed whenever P' has a finite support and s_p can be evaluated on this support:

$$\mathcal{D}_{\mathrm{S}}\left(P, \tfrac{1}{n}\textstyle\sum_{i=1}^n \delta(y_i)\right) = \sqrt{\tfrac{1}{n^2}\textstyle\sum_{i,j=1}^n u_p(y_i, y_j)}. \quad (7)$$

Computation of (7) can proceed with p available up to an unknown normalisation constant, facilitating application in the Bayesian context. Now we are in a position to present our proposed method.

2.4 Measure Transport with KSD

Our proposed method for posterior approximation is simply stated at a high level; we attempt to solve (1) with $\mathcal{D} = \mathcal{D}_{\mathrm{S}}$ and over a set $\mathcal{T}$ of candidate functions $T^\theta : \mathcal{X} \to \mathcal{Y}$ indexed by a finite-dimensional parameter $\theta \in \Theta$. That is, we aim to solve

$$\operatorname*{arg\,min}_{\theta \in \Theta} \mathcal{D}_{\mathrm{S}}(P, T^\theta_\# Q). \quad (8)$$

Discussion of the choice of $\mathcal{T}$ is deferred until Section 2.5. Compared to previous approaches to measure transport using KLD (Rezende and Mohamed, 2015; Kingma et al., 2016; Marzouk et al., 2016; Parno and Marzouk, 2018), KSD is arguably more computationally and theoretically tractable; the computational aspects will now be described.

The solution of (8) is equivalent to minimisation of the function $F(\theta) \coloneqq \mathcal{D}_{\mathrm{S}}(P, T^\theta_\# Q)^2$ over $\theta \in \Theta$. In order to employ state-of-the-art algorithms for stochastic optimisation, an unbiased estimator for the gradient $\nabla_\theta F(\theta)$ is required. A naive starting point would be to differentiate the expression for the KSD of an empirical measure in (7), however the resulting *V-statistic* is biased. Under weak conditions, we establish instead the following unbiased estimator (a *U-statistic*) for the gradient:

Proposition 1. *Let $\Theta \subseteq \mathbb{R}^p$ be an open set. Assume that $\forall \theta \in \Theta$*

(A1) $(x, x') \mapsto u_p(T^\theta(x), T^\theta(x'))$ *is measurable;*

(A2) $\mathbb{E}_{X,X' \sim Q}\left[|u_p(T^\theta(X), T^\theta(X'))|\right] < \infty$*;*

(A3) $\mathbb{E}_{X,X' \sim Q}\left[\|\nabla_\theta u_p(T^\theta(X), T^\theta(X'))\|\right] < \infty$*;*

and that $\forall x, x' \in \mathcal{X}$,

(A4) $\theta \mapsto \nabla_\theta u_p(T^\theta(x), T^\theta(x'))$ *is continuous.*

Then $\forall \theta \in \Theta$

$$\nabla_\theta F(\theta) = \mathbb{E}\Big[\tfrac{1}{n(n-1)} \sum_{i \neq j} \nabla_\theta u_p(T^\theta(x_i), T^\theta(x_j))\Big],$$

where the expectation is taken with respect to independent samples $x_1, \ldots, x_n \sim Q$.

All proofs are contained in Appendix A. The assumptions on u_p amount to assumptions on T, p and k, by virtue of (6). It is not difficult to find explicit assumptions on T, p and k that imply (A1-4), but these may be stronger than required and we prefer to present the most general result.

Armed with an unbiased estimator of the gradient, we can employ a stochastic optimisation approach, such as stochastic gradient descent (SGD; Robbins and Monro, 1951) or adaptive moment estimation (Adam; Kingma and Ba, 2015). See Kushner and Yin (2003); Ruder (2016). For the results reported in the main text we used Adam, with θ initialised as described in Appendix C.1, but other choices were investigated (see Appendix C.2).

2.5 Parametric Transport Maps

In this section we describe some existing classes of transport map $T : \mathcal{X} \to \mathcal{Y}$ that are compatible with KSD measure transport. From Proposition 1 we see that measure transport using KSD does not impose strong assumptions on the transport map. Indeed, compared to KLD (Rezende and Mohamed, 2015; Kingma et al., 2016; Marzouk et al., 2016; Parno and Marzouk, 2018) we do not require that T is a diffeomorphism (T need not even be continuous, nor a bijection), making our framework considerably more general. This additional flexibility may allow measure to be transported more efficiently, using simpler maps. That being said, if one wishes to compute the density of $T_\# Q$ (in addition to sampling from $T_\# Q$), then a diffeomorphism, along with the usual change-of-variables formula, should be used.

Triangular Maps: Rosenblatt (1952) and Knothe et al. (1957) observed that, for $P, Q \in \mathcal{P}(\mathbb{R}^d)$ admitting densities, a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ can without loss of generality be sought in the *triangular form*

$$T(x) = (T_1(x_1), T_2(x_1, x_2), \ldots, T_d(x_1, \ldots, x_d)), \quad (9)$$

where each $T_i : \mathbb{R}^i \to \mathbb{R}$ and $x = (x_1, \ldots, x_d)$ (Bogachev et al., 2005, Lemma 2.1). The triangular form was used in Marzouk et al. (2016); Parno and Marzouk (2018), since the Jacobian determinant, that is required when using KLD (but not KSD), can exploit the fact that ∇T is triangular to maintain linear complexity in d.

Maps from Measure Transport: In the context of a triangular map $T = (T_1, \ldots, T_d)$, Marzouk et al. (2016) and Parno and Marzouk (2018) considered several parametric models for the components T_i, including polynomials, radial basis functions and monotone parameterisations of the form

$$T_i(x_1, \ldots, x_i) = f_i(x_1, \ldots, x_{i-1}) + \int_0^{x_i} \exp(g_i(x_1, \ldots, x_{i-1}, y))\, \mathrm{d}y,$$

for functions $f_i : \mathbb{R}^{i-1} \to \mathbb{R}$ and $g_i : \mathbb{R}^i \to \mathbb{R}$. The monotone parameterisation ensures that $\det \nabla T > 0$ on $\mathbb{R}^d$, which facilitates computation of the density of $T_\# P$, as required for KLD[2].

Maps from Normalising Flows: The principal application of normalising flows is density estimation (Papamakarios et al., 2019; Kobyzev et al., 2020), but the parametric families of transport map used in this literature can also be used for posterior approximation (Rezende and Mohamed, 2015). A normalising flow is required to be a diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ with the property that the density of $T_\# Q$ can be computed. A popular choice that exploits the triangular form (9) is an *autoregressive flow* $T_i(x) = \tau(c_i(x_1, \ldots, x_{i-1}), x_i)$, where τ is a monotonic transformation of x_i parameterised by c_i, e.g. an affine transformation $T_i(x) = \alpha_i x_i + \beta_i$ where c_i outputs $\alpha_i \neq 0$ and β_i. For instance, Kingma et al. (2016) proposed *inverse autoregressive flows* (IAF), taking $T(x) = \mu + \exp(\sigma) \odot x$. Here $\odot$ is elementwise multiplication and μ and σ are vectors output by an autoregressive neural network: one designed so that μ_i, σ_i depend on x only through x_j for $j < i$. In Huang et al. (2018), τ was the output of a monotonic neural network and the resulting flow was called a *neural autoregressive flow* (NAF). Compositions of normalising flows can also be considered, of the form

$$T = T^{(n)} \circ \cdots \circ T^{(1)} \tag{10}$$

where each $T^{(i)}$ is itself a normalising flow e.g. a IAF. For instance, Dinh et al. (2014) proposed using *coupling layers* of the form $T^{(i)}(x) = (h(x_1, \ldots, x_r), x_{r+1}, \ldots, x_d)$, where $r < d$ and $h : \mathbb{R}^r \to \mathbb{R}^r$ is a bijection. These only update the first r components of x, so they are typically composed with permutations.

Regardless of the provenance of a transport map T, all free parameters of T are collectively denoted θ, and are to be estimated. The suitability of a parametric set of candidate maps in combination with KSD is studied both empirically in Section 4 and theoretically, next.

[2]For polynomials and radial basis functions, these authors only enforced $\det \nabla T > 0$ locally, introducing an additional approximation error in evaluation of KLD; such issues do not arise with KSD.

3 Theoretical Assessment

In Section 3.1 we affirm basic conditions on P and Q for a transport map to exist. In Section 3.2 we establish sufficient conditions for the consistency of our method and in Section 3.3 we consider a particular class of transport maps based on neural networks, to demonstrate how our conditions on the transport map can be explicitly validated.

3.1 Existence of an L^2 Transport Map

For a complete separable metric space $\mathcal{X}$, recall that the *Wasserstein space* of order $p \geq 1$ is defined by taking some $x_0 \in \mathcal{X}$ and

$$\mathcal{P}_p(\mathcal{X}) := \left\{ P \in \mathcal{P}(\mathcal{X}) : \int \mathrm{dist}(x, x_0)^p \, \mathrm{d}P(x) < \infty \right\},$$

where the definition is in fact independent of the choice of $x_0 \in \mathcal{X}$ (Villani, 2009, Definition 6.4). For existence of a transport map, we make the following assumptions on P and Q:

Assumption 1 (Assumptions on Q)**.** *The reference measure $Q \in \mathcal{P}(\mathcal{X})$, where $\mathcal{X}$ is a complete separable metric space, and $Q(\{x\}) = 0$ for all $x \in \mathcal{X}$.*

Assumption 2 (Assumptions on P)**.** *The target measure $P \in \mathcal{P}_2(\mathbb{R}^d)$ has a strictly positive density p on $\mathbb{R}^d$.*

These assumptions guarantee the existence of a transport map with L^2 regularity, as shown in the following result:

Proposition 2. *If Assumptions 1 and 2 hold, then there exists a transport map $T \in \prod_{i=1}^d L^2(Q)$ such that $T_\# Q = P$.*

Of course, such a transport map will not be unique in general.

3.2 Consistent Posterior Approximation

The setting for our theoretical analysis considers a sequence $(\mathcal{T}_n)_{n \in \mathbb{N}}$ of parametric classes of transport map, where intuitively $\mathcal{T}_n$ provides a more flexible class of map as n is increased. For example, $\mathcal{T}_n$ could represent the class of triangular maps comprising of nth order polynomials, or a class of normalising flows comprising of n layers in (10).

Assumption 3 (Assumptions on $\mathcal{T}_n$)**.** *There exists a subset $\mathfrak{T} \subseteq \prod_{i=1}^d L^2(Q)$ containing an element $T \in \mathfrak{T}$ for which $T_\# Q = P$. The sequence $(\mathcal{T}_n)_{n \in \mathbb{N}}$ satisfies $\mathcal{T}_n \subseteq \mathfrak{T}$ with $\mathcal{T}_n \subseteq \mathcal{T}_m$ for $n \leq m$ and $\mathcal{T}_\infty := \lim_{n \to \infty} \mathcal{T}_n$ is a dense set in $\mathfrak{T}$.*

Proposition 2 provides sufficient conditions for the set $\mathfrak{T}$ in Assumption 3 to exist; the additional content of

Assumption 3 ensures that $\mathcal{T}_\infty$ is rich enough to consistently approximate an exact transport map, in principle at least. Next, we state our consistency result:

Theorem 2. *Let Assumptions 1 to 3 hold. Further suppose that P is distantly dissipative, with $\nabla \log p$ Lipschitz and $\mathbb{E}_{X\sim P}[\|\nabla \log p(X)\|^2] < \infty$. Suppose that $T_n \in \mathcal{T}_n$ satisfies*

$$\mathcal{D}_{\mathrm{S}}(P, (T_n)_{\#}Q) - \inf_{T\in\mathcal{T}_n} \mathcal{D}_{\mathrm{S}}(P, T_{\#}Q) \overset{n\to\infty}{\to} 0, \qquad (11)$$

with $\mathcal{D}_{\mathrm{S}}$ defined in Theorem 1. Then $(T_n)_{\#}Q \Rightarrow P$.

The statement in (11) accommodates the reality that, although finding the global optimum $T \in \mathcal{T}_n$ will typically be impractical, one can realistically expect to find an element T_n that achieves an almost-as-low value of KSD, e.g. using a stochastic optimisation method. To our knowledge, no comparable consistency guarantees exist for measure transport using KLD.

3.3 Validating our Assumptions on $\mathcal{T}_n$

Recall that earlier work on measure transport placed strong restrictions on the set of maps $\mathcal{T}_n$, requiring each map to be a diffeomorphism with non-vanishing Jacobian determinant. In contrast, our assumptions on $\mathcal{T}_n$ are almost trivial; we do not require smoothness and there is not a bijection requirement. Our assumptions can be satisfied *in principle* whenever $\mathcal{X}$ is a complete separable metric space, since then $\prod_{i=1}^d L^2(Q)$ is separable (Cohn, 2013, Proposition 3.4.5) and admits an orthogonal basis $\{\phi_i\}_{i\in\mathbb{N}}$, so we may take $\mathcal{T}_n = \mathrm{span}\{\phi_1, \ldots, \phi_n\}$ for Assumption 3 to hold. In practice we are able to verify Assumption 3 for quite non-trivial classes of map $\mathcal{T}_n$. To demonstrate, one such example is presented next, similar to that considered in Lu and Lu (2020):

We consider deep neural networks with multi-layer perceptron architecture and ReLU activation functions Let $\mathcal{R}_{l,n}(\mathbb{R}^p \to \mathbb{R}^d)$ denote the set of such *ReLU neural networks* $f : \mathbb{R}^p \to \mathbb{R}^d$ with l layers and width at most n. See Definition 4 in Appendix A.4 for a formal definition.

Proposition 3. *Let Assumptions 1 and 2 hold. Let Q admit a positive, continuous and bounded density on $\mathcal{X} = \mathbb{R}^p$. Let $\mathcal{T}_n = \mathcal{R}_{l,n}(\mathbb{R}^p \to \mathbb{R}^d)$ with $l \coloneqq \lceil \log_2(p+1) \rceil$. Then Assumption 3 holds.*

The maps in Proposition 3 are not bijections, illustrating the greater flexibility of KSD compared to KLD for measure transport. This completes our theoretical discussion, and our attention now turns to empirical assessment.

4 Empirical Assessment

The purpose of this section is to investigate whether KSD is competitive with KLD for measure transport. Section 4.1 compares both approaches using a variety of transport maps and a synthetic test-bed. Then, in Sections 4.2 and 4.3 we consider more realistic posterior approximation problems arising from, respectively, a biochemical oxygen model and a parametric differential equation model.

In all experiments we used the kernel (3) with $c = 1$, $\ell = 0.1$, $\beta = -1/2$ (other choices were investigated in Appendix C.5), the stochastic optimiser Adam with batch size $n = 100$ and learning rate 0.001 (other choices were investigated in Appendix C.2), and the reference distribution Q was taken to be a standard Gaussian on $\mathbb{R}^p$ (other choices were considered in Appendix C.4). Code to reproduce these results is available at `https://github.com/MatthewAlexanderFisher/MTKSD`.

4.1 Synthetic Test-Bed

First we consider a set of synthetic examples that have previously been used to motivate measure transport as an alternative to MCMC. Three targets were considered; p_1 is a sinusoidal density, p_2 is a banana density and p_3 is multimodal; these are formally defined in Appendix B.2. Results for p_1 and p_3 are displayed in Figure 1. The convergence of the approximation to the target is shown for KSD and the corresponding approximation after 10^4 iterations of Adam is shown for KLD. Since, for both objectives, one iteration requires 10^2 evaluations of $\log p_i$ or its gradient, this represents a total of 10^6 calls to $\log p_i$ or its gradient. The corresponding approximation produced using an adaptive Hamiltonian Monte Carlo (HMC) algorithm (Hoffman and Gelman, 2014; Betancourt, 2017) is shown, where the HMC chains were terminated once 10^6 evaluations of $\log p_i$ or its gradient had been performed. Both p_1 and p_3 present challenges for HMC that, to some extent, can be overcome using measure transport.

The results in Figure 1 are for a fixed class of transport map, but now we report a systematic comparison of KSD and KLD. The majority of maps that we consider are diffeomorphic (in order that KLD can be used), implemented in Pyro (Bingham et al., 2018). Since KSD does not place such requirements on the transport map, we also report results for a (non-bijective) ReLU neural network. Our performance measure is an estimate of the Wasserstein-1 distance between the target and approximate distributions computed using 10^4 samples (see Appendix B.1 for details). Results are detailed in Table 1. Overall, there is no clear sense in which KSD out-performs KLD or *vice versa*; KSD

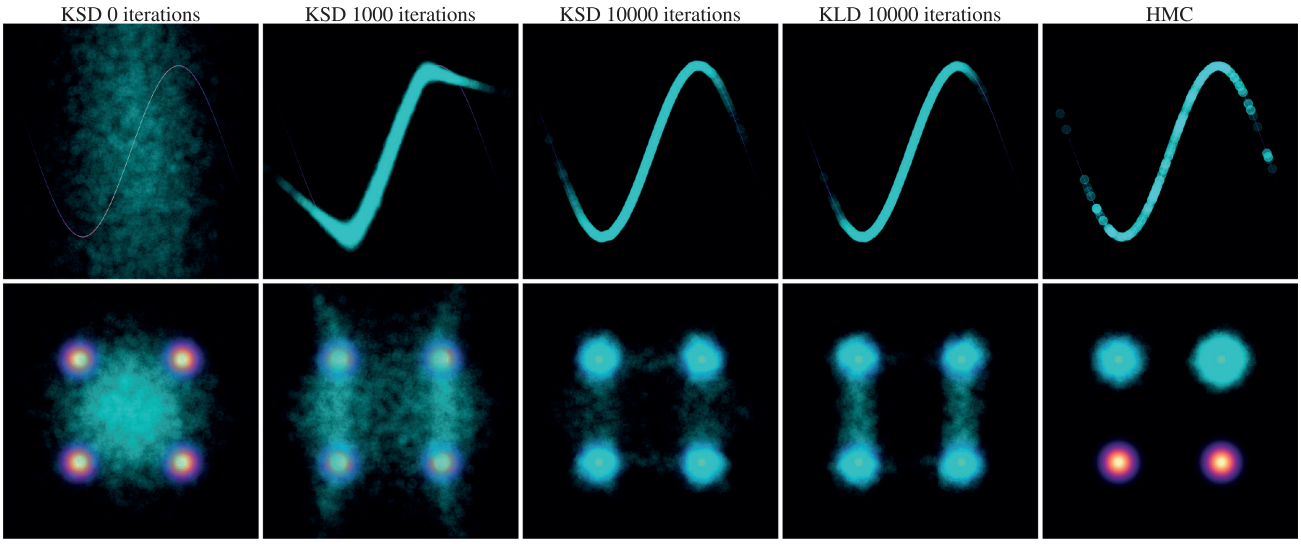

Figure 1: Measure transport with KSD, versus KLD and HMC. The top row reports results for approximation of a sinusoidal density using an inverse autoregressive flow, while the bottom row reports analogous results for a multimodal density and a neural autoregressive flow. The first three columns display convergence of the KSD-based method as the number of iterations of stochastic optimisation is increased. The remaining columns compare the output of the KLD-based method and HMC for an identical computational budget.

performed best on p_1, KLD performed best on p_2, and for p_3 the results were mixed. We conclude that these objectives offer similar performance for measure transport. However, KLD cannot be applied to the ReLU neural network (denoted N/A in Table 1) due to the strong constraints on the mapping that are required by KLD.

Two discussion points are now highlighted: First, it is known that certain normalising flows can capture multiple modes due to their flexibility, however others cannot (Huang et al., 2018). One solution is to consider a mixture of transport maps; i.e. $\sum_{i=1}^{d} w_i T_{\#}^{(i)} Q_i$ with reference distribution $Q_1 \times \cdots \times Q_d$ and mixing weights $w_i > 0$ satisfying $\sum_i w_i = 1$. This idea has been explored recently in Pires and Figueiredo (2020). In Table 1 we report results using mixtures of *inverse autoregressive flows* (IAF). As one might hope, these approximations were successful in finding each of the modes in p_3, but fared relatively worse for p_1 and p_2. Second, since in Adam we are using a Monte Carlo estimator of the gradient, it is natural to ask whether a quasi Monte Carlo estimator would offer an improvement (Wenzel et al., 2018). This was investigated and our results are reported in Appendix C.3.

4.2 Biochemical Oxygen Demand Model

Next we reproduce an experiment that was used to illustrate measure transport using KLD in Parno and Marzouk (2018). The task is parameter inference in a $d = 2$ dimensional oxygen demand model, of the form $B(t) = \alpha_1(1 - \exp(-\alpha_2 t))$, where $B(t)$ is the biochemical oxygen demand at time t, a measure of the consumption of oxygen in a given water column sample due to the decay of organic matter (Sullivan et al., 2010). The parameters to be inferred are $\alpha_1, \alpha_2 > 0$. Full details of the prior and the likelihood are contained in Appendix B.3.

For our experiment, we trained a *block neural autoregressive flow*[3] using $N = 30,000$ iterations of Adam. Results are presented in Figure 2. Unlike the synthetic experiments, we no longer have a closed form for the target P; however, this problem was amenable to MCMC and a long run of HMC (10^6 iterations, thinned by a factor of 100) provided a gold standard, allowing us to approximate the Wasserstein-1 distance from $T_{\#}Q$ to P as in Section 4.1. For the KSD-based method, we obtained a Wasserstein-1 distance of 0.069, while KLD achieved 0.015. Although the Wasserstein-1 distance for KSD is larger than that for KLD, both values are close to the *noise floor* for our approximation of the Wasserstein-1 distance; two independent runs of HMC (10^6 iterations, thinned by a factor of 100), differed in Wasserstein-1 distance by 0.022. We therefore conclude that KSD and KLD performed comparably on this task.

[3]This class of transport map was experimentally observed to outperform the other classes we considered.

		Sinusoidal		Banana		Multimodal	
Transport Map	N	KSD	KLD	KSD	KLD	KSD	KLD
IAF	10^4	**0.38**	0.52	0.20	**0.07**	**0.67**	1.1
IAF (stable)	10^4	**0.35**	0.39	0.16	**0.11**	**0.61**	0.62
NAF	10^4	**0.55**	0.64	0.39	**0.025**	**0.095**	0.11
SAF	10^4	**0.23**	0.58	0.20	**0.18**	**0.30**	0.48
B-NAF	10^4	**0.78**	1.2	0.70	**0.18**	1.0	**0.99**
Polynomial (cubic)	10^4	**0.40**	0.84	0.25	**0.059**	0.51	**0.43**
IAF mixture	3×10^4	1.29	**0.61**	0.19	**0.14**	0.037	**0.036**
ReLU network	5×10^4	**0.71**	N/A	**0.43**	N/A	**0.22**	N/A

Table 1: Results from the synthetic test-bed. The first column indicates which parametric class of transport map was used; full details for each class can be found in Appendix B.2. A map-dependent number of iterations of stochastic optimisation, N, are reported - this is to ensure that all optimisers approximately converged. The main table reports the (first) Wasserstein distance between the approximation $T_\# Q$ and the target P. Bold values indicate which of KSD or KLD performed best.

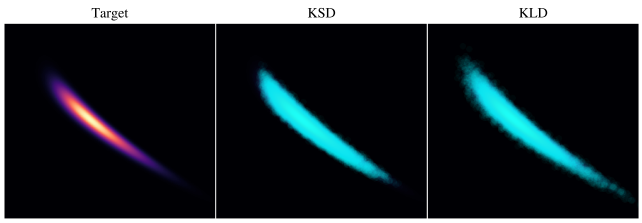

Figure 2: Results for the biochemical oxygen demand model. The leftmost panel is the target distribution, while the central and rightmost panels show samples generated from the output of the methods based, respectively, on KSD and KLD.

4.3 Generalised Lotka-Volterra Model

Our final experiment is a realistic inference problem involving a non-trivial likelihood. Following Parno and Marzouk (2018), we consider parameter inference for a generalised Lotka–Volterra model

$$\begin{aligned} \tfrac{\mathrm{d}p}{\mathrm{d}t}(t) &= rp(t)\big(1 - \tfrac{p(t)}{k}\big) - s\tfrac{p(t)q(t)}{a+p(t)}, \\ \tfrac{\mathrm{d}q}{\mathrm{d}t}(t) &= u\tfrac{p(t)q(t)}{a+p(t)} - vq(t), \end{aligned} \tag{12}$$

where $p(t), q(t) > 0$ are the predator and prey populations respectively at time t and r, k, s, u, a and v, along with the initial conditions $p(0) = p_0$ and $q(0) = q_0$, are parameters to be inferred. Together, these $d = 8$ parameters were inferred from a noisy dataset, with the prior and likelihood reported in Appendix B.4. This task is realistic and yet amenable to MCMC; the latter is an essential requirement to allow us to provide a gold standard against which to assess KSD and KLD, and we again used an extended run of HMC.

For this experiment, the B-NAF class and $N = 5 \cdot 10^4$ iterations of Adam were used. The gradients, required both for HMC and KSD measure transport, were computed using automatic differentiation through the numerical integrator used to solve (12), implemented in the `torchdiffeq` Python package (Chen et al., 2018).

For the KSD-based method, we obtained an approximate Wasserstein-1 distance from $T_\# Q$ to P of 0.130, while KLD achieved 0.110. The noise floor for our approximation of the Wasserstein-1 distance in this case was 0.107. We therefore conclude that KSD and KLD also performed comparably on this more challenging task.

5 Discussion

This paper proposed and studied measure transport using KSD, which can be seen as an instance of *operator variational inference* (Ranganath et al., 2016). Our findings suggest that KSD is a suitable variational objective for measure transport; we observed empirical performance comparable with that of KLD, yet only minimal and verifiable conditions on the map T were required.

There are three potential limitations of KSD compared to KLD: First, the parameters of the kernel must be specified, and a poor choice of kernel parameters can result in poor approximation; see Appendix C.5. It would be interesting to explore whether adversarial maximisation of KSD with respect to the kernel parameters, while minimising KSD over the choice of transport map, offers a solution (Grathwohl et al., 2020). Second, while only first order derivatives are required for KLD, gradient-based optimisation of KSD requires second order derivatives of p. In most automatic differentiation frameworks, and for most models, this is possible at little extra computational cost, but

sometimes this will present difficulties e.g. for models with differential equations involved. Third, it is known that score-based variational objectives can sometimes exhibit pathologies (Wenliang, 2020); some of these are illustrated in Appendix C.9.

Several recent works explored the possibility of combining measure transport with Monte Carlo (Salimans et al., 2015; Wolf et al., 2016; Hoffman, 2017; Caterini et al., 2018; Prangle, 2019; Thin et al., 2020) and it would also be interesting to consider the use of KSD in that context. Related, for both KSD and KLD there is freedom to select the space $\mathcal{X}$ and the reference distribution Q. This could also be handled within the optimisation framework, but further work would be needed to determine how these additional degrees of freedom should be parametrised.

Acknowledgments

MAF was supported by the EPSRC Centre for Doctoral Training in Cloud Computing for Big Data EP/L015358/1 at Newcastle University, UK. THN was supported by a Fulbright scholarship, an American Australian Association scholarship and a Roberta Sykes scholarship. MMG and CJO were supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. The authors thank Onur Teymur for helpful comments on the manuscript.

References

Agrawal, A., Sheldon, D., and Domke, J. (2020). Advances in Black-Box VI: Normalizing Flows, Importance Weighting, and Optimization.

Aliprantis, C. D. and Burkinshaw, O. (1998). *Principles of Real Analysis*. Academic Press.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 214–223.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Arora, R., Basu, A., Mianjy, P., and Mukherjee, A. (2018). Understanding deep neural networks with rectified linear units. In *Proceedings of the 6th International Conference on Learning Representations*.

Au, K. X., Graham, M. M., and Thiery, A. H. (2020). Manifold lifting: Scaling MCMC to the vanishing noise regime. *arXiv:2003.03950*.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434*.

Betancourt, M., Byrne, S., and Girolami, M. (2014). Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv:1411.6669*.

Billingsley, P. (1979). *Probability and Measure*. John Wiley and Sons.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2018). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(18):403.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Bogachev, V. I., Kolesnikov, A. V., and Medvedev, K. V. (2005). Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309–335.

Border, K. (2016). Differentiating an integral: Leibniz' rule. Technical report, Caltech Division of the Humanities and Social Sciences.

Brehmer, J. and Cranmer, K. (2020). Flows for simultaneous manifold learning and density estimation. *arXiv:2003.13913*.

Buchholz, A., Wenzel, F., and Mandt, S. (2018). Quasi-Monte Carlo variational inference.

Cao, N. D., Titov, I., and Aziz, W. (2019). Block neural autoregressive flow. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Caterini, A. L., Doucet, A., and Sejdinovic, D. (2018). Hamiltonian variational auto-encoder. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pages 8167–8177.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural ordinary differential equations. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on Machine Learning*.

Cohn, D. L. (2013). *Measure Theory*. Springer.

Devroye, L. (2013). *Non-Uniform Random Variable Generation*. Springer.

Dinh, L., Krueger, D., and Bengio, Y. (2014). NICE: Non-linear independent components estimation. *arXiv:1410.8516*.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real NVP. *arXiv:1605.08803*.

Dolatabadi, H. M., Erfani, S., and Leckie, C. (2020). Invertible generative modeling using linear rational splines. *arXiv:2001.05168*.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural spline flows. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 258–267.

Eberle, A. (2015). Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3-4):851–886.

Flamary, R. and Courty, N. (2017). POT: Python Optimal Transport library. `https://pythonot.github.io/`.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2018). Large sample analysis of the median heuristic.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). MADE: Masked autoencoder for distribution estimation. In *Proceedings of the 32nd International Conference on Machine Learning*.

Glynn, P. W. (1986). Stochastic approximation for Monte Carlo optimization. In *Proceedings of the 18th Winter Simulation Conference*, pages 356–365.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 28th Conference on Neural Information Processing Systems*, pages 2672–2680.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. In *Proceedings of the 29th Conference on Neural Information Processing Systems*.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*.

Graham, M. M. (2020). Mici: Python implementations of manifold MCMC methods. `https://github.com/matt-graham/mici`.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. (2020). Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *Proceedings of the 37th International Conference on Machine Learning*.

Hoffman, M. D. (2017). Learning deep latent Gaussian models with Markov chain Monte Carlo. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1510–1519.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hu, T., Chen, Z., Sun, H., Bai, J., Ye, M., and Cheng, G. (2018). Stein neural sampler. *arXiv:1810.03545*.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. In *Proceedings of the 35th International Conference on Machine Learning*.

Izmailov, P., Kirichenko, P., Finzi, M., and Wilson, A. G. (2020). Semi-supervised learning with normalizing flows. In *Proceedings of the 37th International Conference on Machine Learning*.

Kechris, A. (1995). *Classical Descriptive Set Theory*. Springer.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Knothe, H. et al. (1957). Contributions to the theory of convex bodies. *The Michigan Mathematical Journal*, 4(1):39–52.

Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. To appear.

Kushner, H. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer Science & Business Media.

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

Liu, Q., Lee, J. D., and Jordan, M. I. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Proceedings of the 30th Conference on Neural Information Processing Systems*.

Livingstone, S. and Zanella, G. (2019). On the robustness of gradient-based MCMC algorithms. *arXiv:1908.11812*.

Lu, Y. and Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. In *Proceedings of the 34th Conference on Neural Information Processing Systems*.

L'Ecuyer, P. (1995). Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41:738–747.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016). Sampling via Measure Transport: An Introduction. *Handbook of Uncertainty Quantification*, page 1–41.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *arXiv:1912.02762*.

Parno, M. D. and Marzouk, Y. M. (2018). Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682.

Pires, G. G. P. F. and Figueiredo, M. A. T. (2020). Variational mixture of normalizing flows. *arXiv:2009.00585*.

Prangle, D. (2019). Distilling importance sampling. *arXiv:1910.03632*.

Ranganath, R., Tran, D., Altosaar, J., and Blei, D. (2016). Operator variational inference. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pages 496–504.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and variational inference in deep latent Gaussian models. In *Proceedings of the 31st International Conference on Machine Learning*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Rockwood, L. L. (2015). *Introduction to Population Ecology*. Wiley-Blackwell.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv:1609.04747*.

Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain Monte Carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1218–1226.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 583–602, Berkeley, Calif. University of California Press.

Sullivan, A. B., Snyder, D. M., and Rounds, S. A. (2010). Controls on biochemical oxygen demand in the upper Klamath river, Oregon. *Chemical Geology*, 269(1):12 – 21.

Thin, A., Kotelevskii, N., Denain, J.-S., Grinsztajn, L., Durmus, A., Panov, M., and Moulines, E. (2020). MetFlow: A new efficient method for bridging the gap between Markov chain Monte Carlo and variational inference. *arXiv:2002.12253*.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer.

Villani, C. (2009). *Optimal Transport, Old and New*. Springer.

Wenliang, L. K. (2020). Blindness of score-based methods to isolated components and mixing proportions. *arXiv:2008.10087*.

Wenzel, F., Buchholz, A., and Mandt, S. (2018). Quasi-Monte Carlo flows. In *Proceedings of the 3rd Workshop on Bayesian Deep Learning*.

Wolf, C., Karl, M., and van der Smagt, P. (2016). Variational inference with Hamiltonian Monte Carlo. *arXiv:1609.08203*.