SLICED KERNELIZED STEIN DISCREPANCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Kernelized Stein discrepancy (KSD), though being extensively used in goodness-of-fit tests and model learning, suffers from the curse-of-dimensionality. We address this issue by proposing the *sliced Stein discrepancy* and its scalable and kernelized variants, which employ kernel-based test functions defined on the optimal one-dimensional projections. When applied to goodness-of-fit tests, extensive experiments show the proposed discrepancy significantly outperforms KSD and various baselines in high dimensions. For model learning, we show its advantages over existing Stein discrepancy baselines by training independent component analysis models with different discrepancies. We further propose a novel particle inference method called *sliced Stein variational gradient descent* (S-SVGD) which alleviates the mode-collapse issue of SVGD in training variational autoencoders.

1 INTRODUCTION

Discrepancy measures for quantifying differences between two probability distributions play key roles in statistics and machine learning. Among many existing discrepancy measures, Stein discrepancy (SD) is unique in that it only requires samples from one distribution and the score function (i.e. the gradient up to a multiplicative constant) from the other (Gorham & Mackey, 2015). SD, a special case of *integral probability metric* (IPM) (Sriperumbudur et al., 2009), requires finding an optimal test function within a given function family. This optimum is analytic when a *reproducing kernel Hilbert space* (RKHS) is used as the test function family, and the corresponding SD is named *kernelized Stein discrepancy* (KSD) (Liu et al., 2016; Chwialkowski et al., 2016). Variants of SDs have been widely used in both Goodness-of-fit (GOF) tests (Liu et al., 2016; Chwialkowski et al., 2016) and model learning (Liu & Feng, 2016; Grathwohl et al., 2020; Hu et al., 2018; Liu & Wang, 2016).

Although theoretically elegant, KSD, especially with RBF kernel, suffers from the "curse-of-dimensionality" issue, which leads to significant deterioration of test power in GOF tests (Chwialkowski et al., 2016; Huggins & Mackey, 2018) and mode collapse in particle inference (Zhuo et al., 2017; Wang et al., 2018). A few attempts have been made to address this problem, however, they either are limited to specific applications with strong assumptions (Zhuo et al., 2017; Chen & Ghattas, 2020; Wang et al., 2018) or require significant approximations (Singhal et al., 2019). As an alternative, in this work we present our solution to this issue by adopting the idea of "slicing". Here the key idea is to project the score function and test inputs onto multiple one dimensional slicing directions, resulting in a variant of SD that only requires to work with one-dimensional inputs for the test functions. Specifically, our contributions are as follows.

- We propose a novel theoretically validated family of discrepancies called *sliced Stein discrepancy* (SSD), along with its scalable variant called *max sliced kernelized Stein discrepancy* (maxSKSD) using kernel tricks and the *optimal test directions*.
- A GOF test is derived based on an unbiased estimator of maxSKSD with optimal test directions. MaxSKSD achieves superior performance on benchmark problems and *restricted Boltzmann machine* models (Liu et al., 2016; Huggins & Mackey, 2018).
- We evaluate the maxSKSD in model learning by two schemes. First, we train an independent component analysis (ICA) model in high dimensions by directly minimising maxSKSD, which results in faster convergence compared to baselines (Grathwohl et al., 2020). Further, we propose a particle inference algorithm based on maxSKSD called the *sliced Stein variational gradient descent* (S-SVGD) as a novel variant of the original SVGD (Liu & Wang, 2016). It alleviates the posterior collapse of SVGD when applied to training variational autoencoders (Kingma & Welling, 2013; Rezende et al., 2014).

2 BACKGROUND

2.1 KERNELIZED STEIN DISCREPANCY

For two probability distributions p and q supported on $\mathcal{X} \subseteq \mathbb{R}^D$ with continuous differentiable densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, we define the score $\boldsymbol{s}_p(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ and $\boldsymbol{s}_q(\boldsymbol{x})$ accordingly. For a test function $f : \mathcal{X} \to \mathbb{R}^D$, the Stein operator is defined as

$$\mathcal{A}_p f(\boldsymbol{x}) = \boldsymbol{s}_p(\boldsymbol{x})^T f(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}^T f(\boldsymbol{x}). \tag{1}$$

For a function $f_0 : \mathbb{R}^D \to \mathbb{R}$, the *Stein class* $\mathcal{F}_q$ of q is defined as the set of functions satisfying Stein's identity (Stein et al., 1972): $\mathbb{E}_q[\boldsymbol{s}_q(\boldsymbol{x}) f_0(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} f_0(\boldsymbol{x})] = \mathbf{0}$. This can be generalized to a vector function $f : \mathbb{R}^D \to \mathbb{R}^D$ by letting f_i belongs to the Stein class of q for each $i \in D$. Then the Stein discrepancy (Liu et al., 2016; Gorham & Mackey, 2015) is defined as

$$D(q, p) = \sup_{f \in \mathcal{F}_q} \mathbb{E}_q[\mathcal{A}_p f(\boldsymbol{x})] = \sup_{f \in \mathcal{F}_q} \mathbb{E}_q[(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x}))^T f(\boldsymbol{x})]. \tag{2}$$

When $\mathcal{F}_q$ is sufficiently rich, and q vanishes at the boundary of $\mathcal{X}$, the supremum is obtained at $f^*(\boldsymbol{x}) \propto \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_q(\boldsymbol{x})$ with some mild regularity conditions on f (Hu et al., 2018). Thus, the Stein discrepancy focuses on the score difference of p and q. *Kernelized Stein discrepancy* (KSD) (Liu et al., 2016; Chwialkowski et al., 2016) restricts the test functions to be in a D-dimensional RKHS $\mathcal{H}_D$ with kernel k to obtain an analytic form. By defining $u_p(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{s}_p(\boldsymbol{x})^T \boldsymbol{s}_p(\boldsymbol{x}') k(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{s}_p(\boldsymbol{x})^T \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{s}_p(\boldsymbol{x}')^T \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}') + \mathrm{Tr}(\nabla_{\boldsymbol{x}, \boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}'))$ the analytic form of KSD is:

$$D^2(q, p) = \left(\sup_{f \in \mathcal{H}_D, ||f||_{\mathcal{H}_D} \leq 1} \mathbb{E}_q[\mathcal{A}_p f(\boldsymbol{x})] \right)^2 = \mathbb{E}_{q(\boldsymbol{x}) q(\boldsymbol{x}')}[u_p(\boldsymbol{x}, \boldsymbol{x}')]. \tag{3}$$

2.2 STEIN VARIATIONAL GRADIENT DESCENT

Although SD and KSD can be directly minimized for variational inference (VI) (Ranganath et al., 2016; Liu & Feng, 2016; Feng et al., 2017), Liu & Wang (2016) alternatively proposed a novel particle inference algorithm called *Stein variational gradient descent* (SVGD). It applies a sequence of deterministic transformations to a set of points such that each of mappings maximally decreases the Kullback-Leibler (KL) divergence from the particles' underlying distribution q to the target p.

To be specific, we define the mapping $T(\boldsymbol{x}) : \mathbb{R}^D \to \mathbb{R}^D$ as $T(\boldsymbol{x}) = \boldsymbol{x} + \epsilon \boldsymbol{\phi}(\boldsymbol{x})$ where $\boldsymbol{\phi}$ characterises the perturbations. The result from Liu & Wang (2016) shows that the optimal perturbation inside the RKHS is exactly the optimal test function in KSD.

Lemma 1. *(Liu & Wang, 2016) Let $T(\boldsymbol{x}) = \boldsymbol{x} + \epsilon \boldsymbol{\phi}(\boldsymbol{x})$ and $q_{[T]}(\boldsymbol{z})$ be the density of $\boldsymbol{z} = T(\boldsymbol{x})$ when $\boldsymbol{x} \sim q(\boldsymbol{x})$. If the perturbation $\boldsymbol{\phi}$ is in the RKHS $\mathcal{H}_D$ and $||\boldsymbol{\phi}||_{\mathcal{H}_D} \leq D(q, p)$, then the steepest descent directions $\boldsymbol{\phi}^*_{q,p}$ is*

$$\boldsymbol{\phi}^*_{q,p}(\cdot) = \mathbb{E}_q[\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) k(\boldsymbol{x}, \cdot) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)] \tag{4}$$

and $\nabla_\epsilon KL[q_{[T]} || p]|_{\epsilon=0} = -D^2(q, p)$.

The first term in Eq.(4) is called drift, which drives the particles towards a mode of p. The second term controls the repulsive force, which spreads the particles around the mode. When particles stop moving, the KL decrease magnitude $\epsilon D^2(q, p)$ is 0, which means the KSD is zero and $p = q$ a.e.

3 SLICED KERNELIZED STEIN DISCREPANCY

We propose the *sliced Stein discrepancy* (SSD) and kernelized version named maxSKSD. Theoretically, we prove their correctness as discrepancy measures. Methodology-wise, we apply maxSKSD to GOF tests, and develop two ways for model learning.

3.1 SLICED STEIN DISCREPANCY

Before moving to the details, we give a brief overview of the intuition on how to tackle the curse-of-fimensionality issue of SD (The right figure of Figure 1). For detailed explanation, refer to appendix B.1. This issue of Stein discrepancy (Eq.2) comes from two sources: the score function $s_p(\boldsymbol{x})$ and

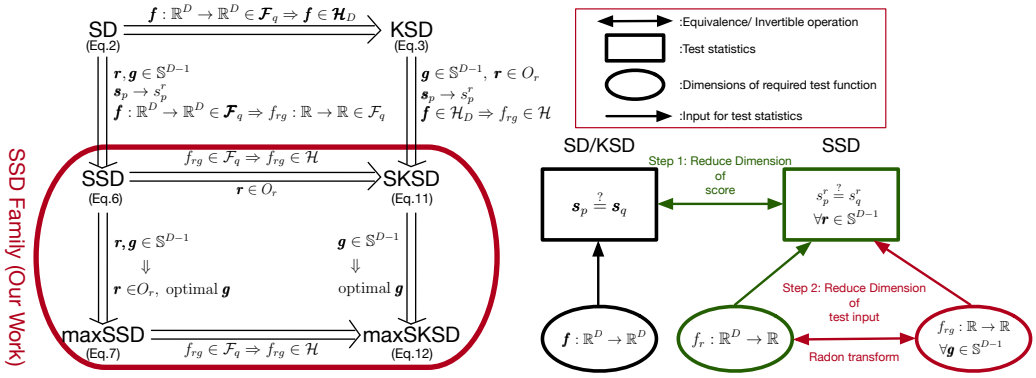

Figure 1: **(Left)** The connections between SD, KSD and the proposed SSD family. **(Right)** The intuition of SSD. The rectangular boxes indicate what statistics the discrepancy wants to test. The circle represents the dimension of the test function required for such test. The double arrow means equivalence relations or invertable operations.

the test function $f(\boldsymbol{x})$ defined on $\mathcal{X} \subset \mathbb{R}^D$. First, we notice that comparing $\boldsymbol{s}_p$ and $\boldsymbol{s}_q$ is equivalent to comparing projected score $s_p^r = \boldsymbol{s}_p^T \boldsymbol{r}$ and s_q^r for all $\boldsymbol{r} \in \mathbb{S}^{D-1}$ on an hyper-sphere (Green square in Figure 1 (Right)). This operation reduces the test function's output from $\mathbb{R}^D$ to $\mathbb{R}$ (Green circle in Figure 1 (Right)). However, its input dimension is not affected. Reducing the input dimension of test functions is non-trivial, as directly removing input dimensions results in the test power decrease of test functions (see examples in appendix B.1). Instead, a better toolkit for this dimension reduction task is Radon transform, which is an invertible transform such that the original function can be reconstructed by its one-dimensional projections (projected by $\boldsymbol{g} \in \mathbb{S}^{D-1}$). Thus, the above two steps not only reduce the dimensions of the test function's output and input, but also maintain the validity of the resulting discrepancy as each step is either equivalent or invertible.

In detail, assume two distributions p and q supported on $\mathbb{R}^D$ with differentiable densities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, and define the test functions $f(\cdot; \boldsymbol{r}, \boldsymbol{g}) : \mathbb{R}^D \to \mathbb{R}$ such that $f(\boldsymbol{x}; \boldsymbol{r}, \boldsymbol{g}) = f_{rg}(\boldsymbol{x}^T \boldsymbol{g})$. The proposed sliced Stein discrepancy (SSD), defined using two uniform distributions $p_r(\boldsymbol{r})$ and $p_g(\boldsymbol{g})$ over the hypersphere $\mathbb{S}^{D-1}$, is given by the following, with $f_{rg} \in \mathcal{F}_q$ meaning $f(\cdot; \boldsymbol{r}, \boldsymbol{g}) \in \mathcal{F}_q$:

$$S(q, p) = \mathbb{E}_{p_r, p_g} \left[\sup_{f_{rg} \in \mathcal{F}_q} \mathbb{E}_q [s_p^r(\boldsymbol{x}) f_{rg}(\boldsymbol{x}^T \boldsymbol{g}) + \boldsymbol{r}^T \boldsymbol{g} \nabla_{\boldsymbol{x}^T \boldsymbol{g}} f_{rg}(\boldsymbol{x}^T \boldsymbol{g})] \right]. \tag{5}$$

We verify the proposed SSD is a valid discrepancy measure, namely, $S(q, p) = 0$ iff. $q = p$ a.e.

Theorem 1. *(SSD Validity) If assumptions 1-4 in appendix A are satisfied, then for two probability distributions p and q, $S(q, p) \geq 0$, and $S(q, p) = 0$ if and only if $p = q$ a.e.*

Despite this attractive theoretical result, SSD is difficult to compute in practice. Specifically, the expectations over $\boldsymbol{r}$ and $\boldsymbol{g}$ can be approximated by Monte Carlo but this typically requires a very large number of samples in high dimensions (Deshpande et al., 2019). We propose to relax such limitations by using only a finite number of slicing directions $\boldsymbol{r}$ from an orthogonal basis O_r of $\mathbb{R}^D$, e.g. the standard basis of one-hot vectors, and the corresponding optimal test direction $\boldsymbol{g}_r$ for each $\boldsymbol{r}$. We call this variant maxSSD, which is defined as follows and validated in Corollary 1.1:

$$S_{max}(q, p) = \sum_{\boldsymbol{r} \in O_r} \sup_{f_{rg_r} \in \mathcal{F}_q, \boldsymbol{g}_r \in \mathbb{S}^{D-1}} \mathbb{E}_q [s_p^r(\boldsymbol{x}) f_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r) + \boldsymbol{r}^T \boldsymbol{g}_r \nabla_{\boldsymbol{x}^T \boldsymbol{g}_r} f_{rg_r}(\boldsymbol{x}^T \boldsymbol{g}_r)]. \tag{6}$$

Corollary 1.1. *(maxSSD) Assume the conditions in Theorem 1, then $S_{max}(q, p) = 0$ iff. $p = q$ a.e.*

3.2 Closed form Solution with the Kernel Trick

The optimal test function given $\boldsymbol{r}$ and $\boldsymbol{g}$ is intractable without further assumptions on the test function families. This introduces another scalability issue as optimizing these test functions explicitly

can be time consuming. Fortunately, we can apply the kernel trick to obtain its analytic form. Assume for each test function $f_{rg} \in \mathcal{H}_{rg}$, where $\mathcal{H}_{rg}$ is a scalar-valued RKHS equipped with kernel $k(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{r}, \boldsymbol{g}) = k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \boldsymbol{x}'^T\boldsymbol{g})$ that satisfies assumption 5 in appendix A and $f_{rg}(\boldsymbol{x}^T\boldsymbol{g}) = \langle f_{rg}, k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \cdot)\rangle_{\mathcal{H}_{rg}}$. We define the following quantities:

$$\xi_{p,r,g}(\boldsymbol{x}, \cdot) = s_p^r(\boldsymbol{x})k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \cdot) + \boldsymbol{r}^T\boldsymbol{g}\nabla_{\boldsymbol{x}^T\boldsymbol{g}}k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \cdot), \tag{7}$$

$$\begin{aligned} h_{p,r,g}(\boldsymbol{x}, \boldsymbol{y}) = {} & s_p^r(\boldsymbol{x})k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \boldsymbol{y}^T\boldsymbol{g})s_p^r(\boldsymbol{y}) + \boldsymbol{r}^T\boldsymbol{g}s_p^r(\boldsymbol{y})\nabla_{\boldsymbol{x}^T\boldsymbol{g}}k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \boldsymbol{y}^T\boldsymbol{g}) + \\ & \boldsymbol{r}^T\boldsymbol{g}s_p^r(\boldsymbol{x})\nabla_{\boldsymbol{y}^T\boldsymbol{g}}k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \boldsymbol{y}^T\boldsymbol{g}) + (\boldsymbol{r}^T\boldsymbol{g})^2\nabla^2_{\boldsymbol{x}^T\boldsymbol{g}, \boldsymbol{y}^T\boldsymbol{g}}k_{rg}(\boldsymbol{x}^T\boldsymbol{g}, \boldsymbol{y}^T\boldsymbol{g}). \end{aligned} \tag{8}$$

The following theorem describes the optimal test function inside SSD (Eq.(5)) and maxSSD (Eq.(6)).

Theorem 2. *(Closed form solution) If $\mathbb{E}_q[h_{p,r,g}(\boldsymbol{x}, \boldsymbol{x})] < \infty$, then*

$$\begin{aligned} D^2_{rg}(q, p) &= ||\sup_{f_{rg} \in \mathcal{H}_{rg}, ||f_{rg}|| \leq 1} \mathbb{E}_q[s_p^r(\boldsymbol{x})f_{rg}(\boldsymbol{x}^T\boldsymbol{g}) + \boldsymbol{r}^T\boldsymbol{g}\nabla_{\boldsymbol{x}^T\boldsymbol{g}}f_{rg}(\boldsymbol{x}^T\boldsymbol{g})]||^2 \\ &= ||\mathbb{E}_q[\xi_{p,r,g}(\boldsymbol{x})]||^2_{\mathcal{H}_{rg}} = \mathbb{E}_{q(\boldsymbol{x})q(\boldsymbol{x}')}[h_{p,r,g}(\boldsymbol{x}, \boldsymbol{x}')]. \end{aligned} \tag{9}$$

Next, we propose the kernelized version of SSD with orthogonal basis O_r, called SKSD.

Theorem 3. *(SKSD as a discrepancy) For two probability distributions p and q, given assumptions 1,2 and 5 in appendix A and $\mathbb{E}_q[h_{p,r,g}(\boldsymbol{x}, \boldsymbol{x})] < \infty$ for all $\boldsymbol{r}$ and $\boldsymbol{g}$, we define SKSD as*

$$SK_o(q, p) = \sum_{\boldsymbol{r} \in O_r} \int_{\mathbb{S}^{D-1}} p_g(\boldsymbol{g}) D^2_{rg}(q, p) d\boldsymbol{g}, \tag{10}$$

which is equal to 0 if and only if $p = q$ a.e.

Following the same idea of maxSSD (Eq.6), it suffices to use optimal slice direction $\boldsymbol{g}_r$ for each $\boldsymbol{r} \in O_r$, resulting in a slicing matrix $\boldsymbol{G} \in \mathbb{R}^{D \times D}$. We name this discrepancy as maxSKSD, or *maxSKSD-g* when we need to distinguish it from another variant described later.

Corollary 3.1. *(maxSKSD) Assume the conditions in Theorem 3 are satisfied. Then*

$$SK_{max}(q, p) = \sum_{\boldsymbol{r} \in O_r} \sup_{\boldsymbol{g}_r} D^2_{rg_r}(q, p) \tag{11}$$

is equal to 0 if and only if $p = q$ a.e.

Figure 1 (Left) clarifies the connections between the mentioned discrepancies. We emphasise that using a single projection $\boldsymbol{g}$ in maxSKSD may be insufficient when no single projected feature $\boldsymbol{x}^T\boldsymbol{g}$ is informative enough to describe the difference between p and q. Instead, in maxSKSD, for each score projection $\boldsymbol{r} \in O_r$, we have a corresponding $\boldsymbol{g}_r$. One can also use the optimal $\boldsymbol{r}$ to replace the summation over O_r, which provides additional benefits in certain GOF tests. We call this discrepancy *maxSKSD-rg*, and its validity can be proved accordingly. Interestingly, in appendix F, we show that there exists limitations of *maxSKSD-g* which can be efficiently addressed by *maxSKSD-rg*.

Kernel choice RBF kernel with median heuristics is a common choice, which satisfies the conditions in Theorem 3. However, better kernels, e.g. deep kernels which evaluate a given kernel on the transformed input $\phi(\boldsymbol{x})$, might be preferred. It is non-trivial to directly use such kernel on SKSD or maxSKSD. We propose a an adapted form of Eq.(10) to incorporate such kernel and maintain its validity. We include the details in appendix D and leave the experiments for future work.

3.3 Application of maxSKSD

Goodness-of-fit Test Assume the optimal test directions $\boldsymbol{g}_r \in \boldsymbol{G}$ are available, maxSKSD (Eq.(11)) can then be estimated using U-statistics (Hoeffding, 1992; Serfling, 2009). Given i.i.d. samples $\{\boldsymbol{x}_i\}_{i=1}^N \sim q$, we have an unbiased minimum variance estimator:

$$\widehat{SK}_{max}(q, p) = \frac{1}{N(N-1)} \sum_{\boldsymbol{r} \in O_r} \sum_{1 \leq i \neq j \leq N} h_{p,r,g_r}(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{12}$$

The asymptotic behavior of the estimator is analyzed in appendix E.1. We use bootstrap (Liu et al., 2016; Huskova & Janssen, 1993; Arcones & Gine, 1992) to determine the threshold for rejecting the null hypothesis, see Algorithm 1 in appendix E.1.

Model Learning The proposed maxSKSD can be applied to model learning in two ways. First, it can be directly used as a training objective, in such case q is the data distribution and p is the model to be learned, and the learning algorithm performs $\min_p SK_{max}(q, p)$. The second model learning scheme is to leverage the particle inference for latent variables and train the model parameters using an EM-like (Dempster et al., 1977) algorithm. Similar to the relation between SVGD and KSD, we can derive a corresponding particle inference algorithm based on maxSKSD, called *sliced-SVGD* (S-SVGD). In short, we define a specific form of the perturbation as $\boldsymbol{\phi}(\boldsymbol{x}) = [\phi_{g_i}(\boldsymbol{x}^T \boldsymbol{g}_i), \dots, \phi_{g_D}(\boldsymbol{x}^T \boldsymbol{g}_D)]^T$ and modify the proofs of Lemma 1 accordingly. The resulting S-SVGD algorithm uses kernels defined on one dimensional projected samples, which sidesteps the vanishing repulsive force problem of SVGD in high dimensions (Zhuo et al., 2017; Wang et al., 2018). We illustrate this in Figure 2 by estimating the variance of a standard Gaussian with the particles obtained by SVGD or S-SVGD (see appendix H.1). We see that as the dimension increases, SVGD severely under-estimates the variance of p, while the S-SVGD remains robust. Furthermore, its validity is justified since in such case the KL gradient equals to maxSKSD which is a valid discrepancy. Readers are referred to appendix E.2 for the derivations.

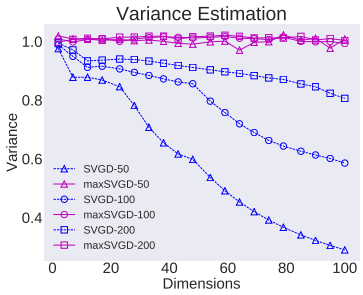

Figure 2: Estimating the average variance of $p(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ across dimensions using SVGD particles. SVGD-50 means the variance are estimated using 50 samples.

4 Experiments

4.1 Goodness of fit test

We evaluate maxSKSD (Eq.(11)) for GOF tests in high dimensional problems. First, we demonstrate its robustness to the increasing dimensionality using the Gaussian GOF benchmarks (Jitkrittum et al., 2017; Huggins & Mackey, 2018; Chwialkowski et al., 2016). Next, we show the advantage of our method for GOF tests on 50-dim *Restricted Boltzmann Machine* (RBM) (Liu et al., 2016; Huggins & Mackey, 2018; Jitkrittum et al., 2017). We included in comparison extensive baseline test statistics for GOF test: Gaussian or Cauchy random Fourier features (RFF) (Rahimi & Recht, 2008), KSD with RBF kernel (Liu et al., 2016; Chwialkowski et al., 2016), finite set Stein discrepancy (FSSD) with random or optimized test locations (Jitkrittum et al., 2017), random feature Stein discrepancy (RFSD) with L2 SechExp and L1 IMQ kernels (Huggins & Mackey, 2018), and maximum mean discrepancy (MMD) (Gretton et al., 2012) with RBF kernel. Notice that we use gradient descent to obtain the test directions $\boldsymbol{g}_r$ (and potentially the slicing directions $\boldsymbol{r}$) for Eq.(11).

4.1.1 GOF Tests with High dimensional Gaussian Benchmarks

We conduct 4 different benchmark tests with $p = \mathcal{N}(0, \boldsymbol{I})$: (1) **Null test**: $q = p$; (2) **Laplace**: $q(\boldsymbol{x}) = \prod_{d=1}^{D} \text{Lap}(x_d|0, 1/\sqrt{2})$ with mean/variance matched to p; (3) **Multivariate-t**: q is multivariate-t with 5 degrees of freedom; (4) **Diffusion**: $q(\boldsymbol{x}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ where the variance of 1st-dim is 0.3 and the rest is the same as in $\boldsymbol{I}$. For the testing setup, we set the significance level $\alpha = 0.05$. For FFSD and RFSD, we use the open-sourced code from the original publications. We only consider maxSKSD-g here as it already performs nearly optimally. We refer to appendix G.1 for details.

Figure 3 shows the GOF test performances and the corresponding discrepancy values. In summary, the proposed maxSKSD outperforms the baselines in all tests, where the result is robust to the increasing dimensions and the discrepancy values match the expected behaviours.

Null The left-most column in Figure 3 shows that all methods behave as expected as the rejection rate is closed to the significance level, except for RFSD with L2 SechExp kernel. All the discrepancy values oscillate around 0, with the KSD being less stable.

Laplace and Multivariate-t The two middle columns of Figure 3 show that maxSKSD-g achieves a nearly perfect rejection rate consistently as the dimension increases, while the test power for all baselines decreases significantly. For the discrepancy values, similar to the KL divergence between q and p, maxSKSD-g linearly increases with dimensions due to the independence assumptions..

Diffusion This is a more challenging setting since p and q only differ in one of their marginal distributions, which can be easily buried in high dimensions. As shown in the rightmost column of Figure 3, all methods failed in high dimensions except maxSKSD-g, which still consistently

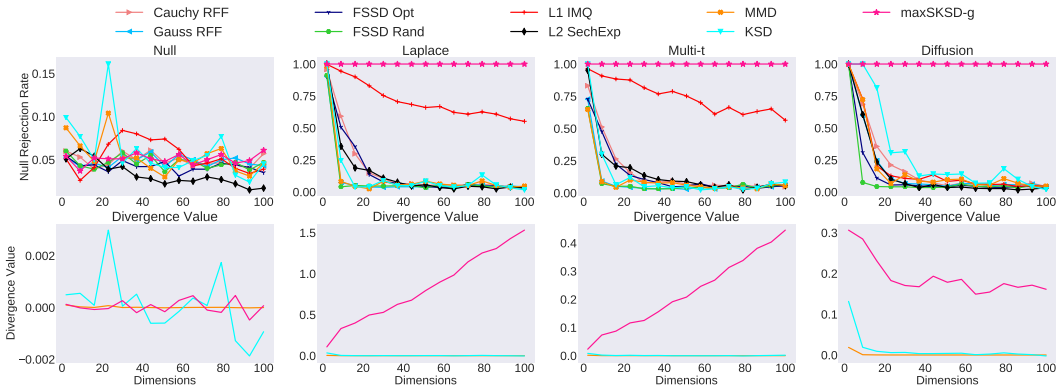

Figure 3: Each column reports GOF test results for a different alternative hypothesis, with the upper panel showing the rejection rate of the Null hypothesis and the lower panel showing the discrepancy value averaged over all trials. Both quantities are plotted w.r.t. the number of dimensions.

achieves optimal performance. For the discrepancy values, we expect a positive constant due to the one marginal difference between p and q. Only maxSKSD-g behaves as expected as the problem dimension increases. The decreasing value at the beginning is probably due to the difficulty in finding the optimal direction $\boldsymbol{g}$ in high dimensions when the training set is small.

4.1.2 RBM GOF TEST

We demonstrate the power maxSKSD for GOF tests on RBMs, but we now also include results for *maxSKSD-rg*. We follow the test setups in Liu et al. (2016); Jitkrittum et al. (2017); Huggins & Mackey (2018) where different amounts of noise are injected into the weights to form the alternative hypothesis q. The samples are drawn using block Gibbs samplers. Refer to appendix G.2 for details. Figure 4 shows that maxSKSD based methods dominate the baselines, especially with maxSKSD-rg significantly outperforming the others. At perturbation level 0.01, maxSKSD-rg achieves 0.96 rejection rate, while others are all below 0.5. This result shows the advantages of optimizing the slicing directions $\boldsymbol{r}$.

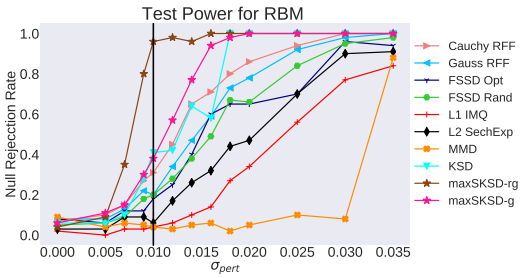

Figure 4: RBM GOF Test with different levels of perturbation noise. The black vertical line indicates the perturbation level at 0.01.

4.2 MODEL LEARNING

We evaluate the efficiency of maxSKSD-based algorithms in training machine learning models. First, we use *independent component analysis* (ICA) which is often used as a benchmark for evaluating training methods for energy-based model (Gutmann & Hyvärinen, 2010; Hyvärinen, 2005; Ceylan & Gutmann, 2018). Our approach trains the ICA model by directly minimizing maxSKSD. Next, we evaluate the proposed S-SVGD particle inference algorithm, when combined with *amortization* (Feng et al., 2017; Pu et al., 2017), in the training of a *variational autoencoder* (VAE) (Kingma & Welling, 2013; Rezende et al., 2014) on binarized MNIST. Appendix H.5 also shows superior results for S-SVGD when training a Bayesian neural network (BNN) on UCI datasets (Dua & Graff, 2017).

4.2.1 ICA

ICA consists of a simple generative process $\boldsymbol{z} \sim \text{Lap}(0, 1)$ and $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$, where the model parameters are a non-singular matrix $\boldsymbol{W} \in \mathbb{R}^{D \times D}$. The log density for $\boldsymbol{x}$ is $\log p(\boldsymbol{x}) = \log p_z(\boldsymbol{W}^{-1}\boldsymbol{x}) + C$, where the normalization constant C can be ignored when training with Stein discrepancies. We train the models on data sampled from a randomly initialized ICA model and evaluate the corresponding test log likelihoods. We compare maxSKSD with KSD and the state-of-the-art LSD (Grathwohl et al., 2020). For more details on the setup, we refer the reader to appendix H.2.

Table 1 shows that both maxSKSD and LSD are robust to increasing dimensions, with maxSKSD being better when D is very large. Also at $D = 200$, maxSKSD converges significantly faster than

Table 1: Test NLL for different dimensional ICA with different objective functions. The above results are averaged over 5 independent runs of each methods.

Method	Dimension						
	$D = 10$	$D = 20$	$D = 40$	$D = 60$	$D = 80$	$D = 100$	$D = 200$
KSD	**-10.23**	-15.98	-34.50	-56.87	-86.09	-116.51	-329.49
LSD	-10.42	-14.54	**-17.16**	**-15.05**	-12.39	-5.49	46.63
maxSKSD	-10.45	**-14.50**	-17.28	-15.70	**-11.91**	**-4.21**	**47.72**

Table 2: Average log likelihood on first $5,000$ test images for different D of latent dimensions.

Method	Latent Dim			
	D=16	D=32	D=48	D=64
Vanilla VAE	-91.50	-90.39	-90.58	-91.50
SVGD VAE	**-88.58**	-90.43	-93.47	-94.88
S-SVGD VAE	-89.17	**-87.55**	**-87.74**	**-87.78**

Table 3: Label entropy and accuracy for imputed images.

Method	Entropy	Accuracy
Vanilla VAE	0.297	0.718
SVGD VAE	0.538	0.691
S-SVGD VAE	**0.542**	**0.728**

LSD (see Figure 9 in appendix H.3). This faster convergence is due to the closed-form solution for the optimal test functions, whereas LSD requires adversarial training. While KSD is also kernel-based, it suffers from the curse-of-dimensionality and fails to train the model properly for $D > 20$. Instead the proposed maxSKSD can successfully avoid the problems of KSD with high dimensional data.

4.2.2 Amortized SVGD

Finally, we consider training VAEs with implicit encoders on dynamically binarized MNIST. The decoder is trained as in vanilla VAEs, but the encoder is trained by amortization (Feng et al., 2017; Pu et al., 2017), which minimizes the mean square error between the initial samples from the encoder, and the modified samples driven by the SVGD/S-SVGD dynamics (Algorithm 3 in appendix H.4).

We report performance in terms of test log-likelihood (LL). Furthermore we consider an imputation task, by removing the pixels in the lower half of the image and imputing the missing values using (approximate) posterior sampling from the VAE models. The performance is measured in terms of imputation diversity and correctness, using label entropy and accuracy. For fair comparisons, we do not tune the coefficient of the repulsive force. We refer to appendix H.4 for details.

Table 2 reports the average test LL. We observe that S-SVGD is much more robust to the increasing latent dimensions compared to SVGD. To be specific, with $D = 16$, SVGD performs the best where S-SVGD performs slightly worse than SVGD. However, when the dimension starts to increase, LL of SVGD drops significantly. For $D = 64$, a common choice for latent space, it performs even significantly worse than vanilla VAE. On the other hand, S-SVGD is much more robust. Notice that the purpose of this experiment is to show compare their robustness instead of achieving the state-of-the-art performance. Still the performance can be easily boosted, e.g. running longer S-SVGD steps before encoder update, we leave it for the future work.

For the imputation task, we compute the label entropy and accuracy for the imputed images (Table 3). We observe S-SVGD has higher label entropy compared to vanilla VAE and better accuracy compared to SVGD. This means both S-SVGD and SVGD capture the muli-modality nature of the posterior compared to uni-modal Gaussian distribution. However, high label entropy itself may not be a good indicator for the quality of the learned posterior. One can think of a counter-example that the imputed images are diverse but does not look like any digits. This may also gives a high label entropy but the quality of the posterior is poor. Thus, we use the accuracy to indicate the "correctness" of the imputed images, with higher label accuracy meaning the imputed images are closed to the original image. Together, a good model should give a higher label entropy along with the high label accuracy. We observe S-SVGD has more diverse imputed images with high imputation accuracy.

4.3 Summary of the experiments in appendix

We present further empirical results on GOF tests and model learning in the appendix to demonstrate the advantages of the proposed maxSKSD. As a summary glance of the results:

- In appendix F, we analyse the potential limitations of maxSKSD-g and show that they can be mitigated by maxSKSD-rg, i.e. optimising the slicing direction $\boldsymbol{r}$;

- In appendix G.3, we successfully apply maxSKSD to selecting the step size for *stochastic gradient Hamiltonian Monte Carlo* (SGHMC) (Chen et al., 2014);
- In appendix H.5, we show that the proposed S-SVGD approach out-performs the original SVGD on Bayesian neural network regression tasks.

5 Related Work

Stein Discrepancy SD (Gorham & Mackey, 2015) and KSD (Liu et al., 2016; Chwialkowski et al., 2016) are originally proposed for GOF tests. Since then research progress has been made to improve these two discrepancies. For SD, LSD (Grathwohl et al., 2020; Hu et al., 2018) is proposed to increase the capacity of test functions using neural networks with L_2 regularization. On the other hand, FSSD (Jitkrittum et al., 2017) and RFSD (Huggins & Mackey, 2018) aim to reduce the computation cost of KSD from $O(n^2)$ to $O(n)$ where n is the number of samples. Still the curse-of-dimensionality issue remains to be addressed in KSD, and the only attempt so far (to the best of our knowledge) is the *kernelized complete conditional Stein discrepancy* (KCC-SD (Singhal et al., 2019)), which share our idea of avoiding kernel evaluations on high dimensional inputs but through comparing conditional distributions. KCC-SD requires the sampling from $q(x_d|\boldsymbol{x}_{-d})$, which often needs significant approximations in practice due to its intractability. This makes KCC-SD less suited for GOF test due to estimation quality in high dimensions. On the other hand, our approach does not require this approximation, and the corresponding estimator is well-behaved asymptotically.

Wasserstein Distance and Score matching Sliced Wasserstein distance (SWD) (Kolouri et al., 2016) and sliced score matching (SSM) (Song et al., 2019) also uses the "slicing" idea. However, their motivation is to address the computational issues rather than statistical difficulties in high dimensions. SWD leveraged the closed-form solution of 1D Wasserstein distance by projecting distributions onto 1D slices. SSM uses Hutchson's trick (Hutchinson, 1990) to approximate the trace of Hessian.

Particle Inference Zhuo et al. (2017); Wang et al. (2018) proposed *message passing SVGD* to tackle the well-known mode collapse problem of SVGD using local kernels in the graphical model. However, our work differs significantly in both theory and applications. Theoretically, the discrepancy behind their work is only valid if p and q have the same Markov blanket structure (refer to Section 3 in Wang et al. (2018) for detailed discussion). Thus, unlike our method, no GOF test and practical inference algorithm can be derived for generic cases. Empirically, the Markov blanket structure information is often unavailable, whereas our method only requires projections that can be easily obtained using optimizations. *Projected SVGD* (pSVGD) is a very recent attempt (Chen & Ghattas, 2020) which updates the particles in an adaptively constructed low dimensional space, resulting in a biased inference algorithm. The major difference compared to S-SVGD is that our work still updates the particles in the original space with kernel being evaluated in 1D projections. Furthermore, S-SVGD can theoretically recover the correct target distribution. There is no real-world experiments provided in (Chen & Ghattas, 2020), and a stable implementation of pSVGD is non-trivial, so we did not consider pSVGD when selecting the baselines.

6 Conclusion

We proposed sliced Stein discrepancy (SSD), as well as its scalable and kernelized version maxSKSD, to address the curse-of-dimensionality issues in Stein discrepancy. The key idea is to project the score function on one-dimensional slices and define (kernel-based) test functions on one-dimensional projections. We also theoretically prove their validity as a discrepancy measure. We conduct extensive experiments including GOF tests and model learning to show maxSKSD's improved performance and robustness in high dimensions. There are three exciting avenues of future research. First, although validated by our theoretical study in appendix D, practical approaches to incorporate deep kernels into SSD remains an open question. Second, the performance of maxSKSD crucially depends on the optimal projection direction, so better optimization methods to efficiently construct this direction is needed. Lastly, we believe "slicing" is a promising direction for kernel design to increase the robustness to high dimensional problems in general. For example, MMD can be easily extended to high dimensional two-sample tests using this kernel design trick.

REFERENCES

Miguel A Arcones and Evarist Gine. On the bootstrap of u and v statistics. *The Annals of Statistics*, pp. 655–674, 1992.

Adi Ben-Israel. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.

Ronald N Bracewell. Strip integration in radio astronomy. *Australian Journal of Physics*, 9(2): 198–217, 1956.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Ciwan Ceylan and Michael U Gutmann. Conditional noise-contrastive estimation of unnormalised models. *arXiv preprint arXiv:1806.03664*, 2018.

Peng Chen and Omar Ghattas. Projected Stein variational gradient descent. *arXiv preprint arXiv:2002.03469*, 2020.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691, 2014.

Andreas Christmann and Ingo Steinwart. Support vector machines.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings, 2016.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22, 1977.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-sliced Wasserstein distance and its use for gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized Stein variational gradient descent. *arXiv preprint arXiv:1707.06626*, 2017.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pp. 226–234, 2015.

Will Grathwohl, Kuan-Chieh Wang, Jorn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Cutting out the middle-man: Training and evaluating energy-based models without sampling. *arXiv preprint arXiv:2002.05616*, 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in Statistics*, pp. 308–334. Springer, 1992.

Tianyang Hu, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems*, pp. 1899–1909, 2018.

Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate u-statistics. *The Annals of Statistics*, pp. 1811–1823, 1993.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pp. 262–271, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, pp. 261–272, 2019.

Qiang Liu and Yihao Feng. Two methods for wild variational inference. *arXiv preprint arXiv:1612.00081*, 2016.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in neural information processing systems*, pp. 2378–2386, 2016.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284, 2016.

Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. VAE learning via Stein variational gradient descent. In *Advances in Neural Information Processing Systems*, pp. 4236–4245, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pp. 496–504, 2016.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Raghav Singhal, Xintian Han, Saad Lahlou, and Rajesh Ranganath. Kernelized complete conditional Stein discrepancy. *arXiv preprint arXiv:1904.04478*, 2019.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. *arXiv preprint arXiv:1905.07088*, 2019.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Charles Stein, Persi Diaconis, Susan Holmes, Gesine Reinert, et al. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pp. 1–25. Institute of Mathematical Statistics, 2004.

Charles Stein et al. 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*. The Regents of the University of California, 1972.

Dilin Wang, Zhe Zeng, and Qiang Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning*, pp. 5219–5227. PMLR, 2018.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*, 2016.

Jingwei Zhuo, Chang Liu, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Message passing Stein variational gradient descent. *arXiv preprint arXiv:1711.04425*, 2017.