Kernel Semi-Implicit Variational Inference

Ziheng Cheng[* 1]   **Longlin Yu**[* 1]   **Tianyu Xie**[1]   **Shiyue Zhang**[1]   **Cheng Zhang**[1 2]

Abstract

Semi-implicit variational inference (SIVI) extends traditional variational families with semi-implicit distributions defined in a hierarchical manner. Due to the intractable densities of semi-implicit distributions, classical SIVI often resorts to surrogates of evidence lower bound (ELBO) that would introduce biases for training. A recent advancement in SIVI, named SIVI-SM, utilizes an alternative score matching objective made tractable via a minimax formulation, albeit requiring an additional lower-level optimization. In this paper, we propose kernel SIVI (KSIVI), a variant of SIVI-SM that eliminates the need for lower-level optimization through kernel tricks. Specifically, we show that when optimizing over a reproducing kernel Hilbert space (RKHS), the lower-level problem has an explicit solution. This way, the upper-level objective becomes the kernel Stein discrepancy (KSD), which is readily computable for stochastic gradient descent due to the hierarchical structure of semi-implicit variational distributions. An upper bound for the variance of the Monte Carlo gradient estimators of the KSD objective is derived, which allows us to establish novel convergence guarantees of KSIVI. We demonstrate the effectiveness and efficiency of KSIVI on both synthetic distributions and a variety of real data Bayesian inference tasks.

1. Introduction

Variational inference (VI) is an optimization based approach widely used to approximate posterior densities for Bayesian models (Jordan et al., 1999; Wainwright & Jordan, 2008; Blei et al., 2016). The idea behind VI is to first posit a family of variational distributions over the model parameters (or latent variables), and then to find a member from this family that is close to the target posterior, where the closeness is often measured by the Kullback-Leibler (KL) divergence. Since the posterior generally does not possess analytical densities, in practice people often maximize the evidence lower bound (ELBO) instead, which is an equivalent but tractable reformulation (Jordan et al., 1999).

In classical VI, a common assumption is that variational distributions are factorized over the parameters (or latent variables) (Bishop & Tipping, 2000; Hinton & van Camp, 1993; Peterson & Hartman, 1989). When combined with conditional conjugate models, this mean-field assumption allows for a straightforward coordinate-ascent optimization scheme with closed-form update rules (Blei et al., 2016). However, conditional conjugacy may not hold in practice, and a factorized variational distribution may fail to capture the complex posterior due to the ignorance of the dependencies between different factorization components. Recent years have witnessed significant advancements in the field of VI, with numerous efforts aimed at alleviating these constraints. To address the conjugacy condition, black-box VI methods have been proposed for a broad class of models that allow generic training algorithms via Monte Carlo gradient estimators (Paisley et al., 2012; Nott et al., 2012; Ranganath et al., 2014; Rezende et al., 2014; Kingma & Welling, 2014; Titsias & Lázaro-Gredilla, 2014). In parallel, more flexible variational families have emerged, either explicitly incorporating certain structures among the parameters or drawing inspiration from invertible transformations of probability distributions (Jaakkola & Jordan, 1998; Saul & Jordan, 1996; Giordano et al., 2015; Tran et al., 2015; Rezende & Mohamed, 2015; Dinh et al., 2017; Kingma et al., 2016; Papamakarios et al., 2019). Despite their successes, it is noteworthy that all these approaches assume tractable densities of variational distributions.

To further enhance the capacity of variational families, one strategy is to incorporate the implicit models defined through neural network mappings (Huszár, 2017; Tran et al., 2017; Mescheder et al., 2017; Shi et al., 2018a;b; Song et al., 2019). These implicit models can provide more flexible variational approximations that allow easy sampling procedures by design. However, their probability densities are often intractable, rendering direct computation of the log density ratio, crucial for evaluating the ELBO, impossible. Implicit

---

[*]Equal contribution (Alphabetical) [1]School of Mathematical Sciences, Peking University, China [2]Center for Statistical Science, Peking University, China. Correspondence to: Cheng Zhang <chengzhang@math.pku.edu.cn>.

VI methods, therefore, generally rely on density ratio estimation for training, which not only introduces additional optimization complexity but also is known to be difficult in high dimensional settings (Sugiyama et al., 2012).

To circumvent the need for density ratio estimation, semi-implicit variational inference (SIVI) has been introduced (Yin & Zhou, 2018; Moens et al., 2021). In SIVI, variational distributions are crafted using a semi-implicit hierarchical framework which allows efficient computation of surrogate ELBOs for training. As surrogate ELBOs are inherently biased, Titsias & Ruiz (2019) proposed an unbiased estimator of the gradient of the exact ELBO, albeit requiring an inner-loop MCMC sampling from a reversed conditional distribution that can easily become expensive in high-dimensional regimes. Alternatively, Yu & Zhang (2023) introduced SIVI-SM, a variant of SIVI that uses an objective based on Fisher divergence rather than KL divergence. They showed that this new objective facilitates a simple minimax formulation, where the intractability of densities can be naturally handled similarly to denoising score matching (DSM) (Vincent, 2011), due to the hierarchical construction of semi-implicit distributions. However, this comes at the price of an additional lower-level problem that requires optimization over a family of functions parameterized via neural networks.

As a popular approach in machine learning, kernel methods have shown great advantage when it comes to optimization over function spaces (Gretton et al., 2012; Liu & Wang, 2016). In this paper, we propose kernel semi-implicit variational inference (KSIVI), a variant of SIVI-SM that eliminates the need for lower-level optimization through kernel tricks. More specifically, we show that there exists an explicit solution to the lower-level problem when optimizing over a reproducing kernel Hilbert space (RKHS). Upon substitution, the objective for the upper-level problem becomes the kernel Stein discrepancy (KSD), a widely used kernel-induced dissimilarity measure between probability distributions (Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017b). Moreover, similarly as done in DSM, this solution has a tractable form that solely involves conditional densities due to the hierarchical structure of semi-implicit variational families, making KSD readily computable for stochastic gradient descent (SGD). We derive an upper bound for the variance of the Monte Carlo gradient estimators under mild assumptions. This bound allows us to establish convergence guarantees of KSIVI to a stationary point using standard analysis in stochastic optimization. We show empirically that KSIVI performs on par or better than SIVI-SM on both synthetic distributions and a variety of real data Bayesian inference tasks while enjoying more efficient computation, more stable training, and fewer hyperparameter tuning headaches.

2. Background

Notations Let $\|\cdot\|$ be the standard Euclidean norm of a vector and the operator norm of a matrix or high-dimensional tensor. Let $\|\cdot\|_\infty$ denote the ℓ_∞ norm of a vector. For any $x, y \in \mathbb{R}^d$, the expressions $x \odot y, \frac{x}{y}, x^2$ stand for element-wise product, division, and square, respectively. The RKHS induced by kernel $k(\cdot, \cdot)$ is denoted as $\mathcal{H}_0$ and the corresponding Cartesian product is defined as $\mathcal{H} := \mathcal{H}_0^{\otimes d}$. Let $\nabla_1 k(\text{resp.} \nabla_2 k)$ be the gradient of the kernel w.r.t. the first (resp. second) variable. We denote the inner product in $\mathbb{R}^d$ and Hilbert space $\mathcal{F}$ by $\langle \cdot, \cdot \rangle$ and $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, respectively. For a probabilistic density function q, we use s_q to denote its score function $\nabla \log q$. Finally, we use standard $\lesssim$ and $\gtrsim$ to omit constant factors.

2.1. Semi-Implicit Variational Inference

Given an observed data set $\mathcal{D}$, the classical VI methods find the best approximation in a variational family $\{q_\phi(x)\}_{\phi \in \Phi}$ to the posterior distribution $p(x|\mathcal{D})$ by maximizing the following evidence lower bound (ELBO)

$$\mathcal{L}(q_\phi(x)) = \mathbb{E}_{q_\phi(x)}\left[\log p(\mathcal{D}, x) - \log q_\phi(x)\right], \quad (1)$$

where ϕ are the variational parameters. To expand the capacity of variational families, SIVI (Yin & Zhou, 2018) constructs a semi-implicit variational family with hierarchical structure as

$$q_\phi(x) = \int q_\phi(x|z)q(z)\mathrm{d}z, \quad (2)$$

where $q(z)$ is the mixing distribution which is easy to sample from and can be implicit, and $q_\phi(x|z)$ is the conditional layer that is required to be explicit. Compared to classical VI which requires tractable variational distributions with explicit densities, the semi-implicit construction greatly enhances the approximation ability, making it able to capture the complex dependence among different variables.

Since $q_\phi(x)$ in the ELBO (1) is no longer tractable, Yin & Zhou (2018) considers a sequence of surrogates $\underline{\mathcal{L}}^{(K)}(q_\phi(x))$ of the ELBO (1) defined as

$$\underline{\mathcal{L}}^{(K)}(q_\phi(x)) = \mathbb{E}_{q(z^{(0:K)})q_\phi(x|z^{(0)})} \log \frac{p(\mathcal{D}, x)}{\frac{1}{K+1}\sum_{k=0}^{K} q_\phi(x|z^{(k)})}, \quad (3)$$

where $q(z^{(0:K)}) = \prod_{i=0}^{K} q(z^{(k)})$. The introduced surrogate $\underline{\mathcal{L}}^{(K)}(q_\phi(x))$ is a lower bound of the $\mathcal{L}(q_\phi(x))$ and is asymptotically exact in the sense that $\lim_{K \to \infty} \underline{\mathcal{L}}^{(K)}(q_\phi(x)) = \mathcal{L}(q_\phi(x))$.

Besides the ELBO-based objectives, the score-based measure has also been introduced for VI (Liu et al., 2016; Zhang et al., 2018; Hu et al., 2018; Korba et al., 2021) where the score function $s_p(x)$ of the target posterior distribution

$p(x) = p(x|\mathcal{D})$ is generally assumed to be tractable. In particular, SIVI-SM (Yu & Zhang, 2023) learns the semi-implicit approximation by minimizing the Fisher divergence via a minimax formulation

$$\min_{\phi} \max_{f} \ \mathbb{E}_{q_{\phi}(x)} \left[2f(x)^T [s_p(x) - s_{q_{\phi}}(x)] - \|f(x)\|^2 \right].$$

(4)

Using a similar trick in DSM (Vincent, 2011), (4) has a tractable form as follows

$$\min_{\phi} \max_{f} \ \mathbb{E}_{q_{\phi}(x,z)} \left[2f(x)^T [s_p(x) - s_{q_{\phi}(\cdot|z)}(x)] - \|f(x)\|^2 \right],$$

(5)

where $q_{\phi}(x, z) = q_{\phi}(x|z)q(z)$. The introduced auxiliary function $f(x)$ is implemented as a neural network $f_{\psi}(x)$ in practice, necessitating additional attention to training deep models and tuning hyperparameters.

2.2. Kernel Stein Discrepancy

Consider a continuous and positive semi-definite kernel $k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ and its corresponding RKHS $\mathcal{H}_0$ of real-valued functions in $\mathbb{R}^d$. The reproducing property of $\mathcal{H}_0$ indicates that for any function $f \in \mathcal{H}_0$, $\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}_0} = f(x)$. Given a measure q such that $\int k(x, x) \mathrm{d}q(x) < \infty$, it follows that $\mathcal{H}_0 \subset L^2(q)$, and the integral operator $S_{q,k} : L^2(q) \rightarrow \mathcal{H}_0$ is defined as

$$(S_{q,k}f)(\cdot) := \int k(\cdot, y) f(y) \mathrm{d}q(y).$$

(6)

The Cartesian product $\mathcal{H} := \mathcal{H}_0^{\otimes d}$, which consists $f = (f_1, \cdots, f_d)$ with $f_i \in \mathcal{H}_0$, is an extension to vector-valued functions. The inner product in $\mathcal{H}$ is defined by $\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{d} \langle f_i, g_i \rangle_{\mathcal{H}_0}$. For $f \in \mathcal{H}$ and $g_0 \in \mathcal{H}_0$, we extend g_0 and define $\langle f, g_0 \rangle_{\mathcal{H}} = (\langle f_1, g_0 \rangle_{\mathcal{H}_0}, \cdots, \langle f_d, g_0 \rangle_{\mathcal{H}_0})$. For vector-valued inputs f, we reload the definition of the operator $S_{q,k}$ and apply it element-wisely. The Stein discrepancy (Gorham & Mackey, 2015; Liu & Wang, 2016; Liu et al., 2016) which measures the difference between q and p is defined as

$$\mathrm{S}(q\|p) := \max_{f \in \mathcal{F}} \ \mathbb{E}_q[\nabla \log p(x)^T f(x) + \nabla \cdot f(x)], \quad (7)$$

where $\mathcal{F}$ is a pre-defined function space. To tackle the minimization problem of $\mathrm{S}(q\|p)$, the kernel Stein discrepancy (KSD) (Liu & Wang, 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017b) maximizes f within the unit ball of the RKHS, i.e., $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$. Since the optima of $f(x)$ has a closed form, KSD has the following equivalent form

$$\mathrm{KSD}(q\|p) = \left\| S_{q,k} \nabla \log \frac{p}{q} \right\|_{\mathcal{H}}$$

$$= \left\langle \nabla \log \frac{p}{q}, S_{q,k} \nabla \log \frac{p}{q} \right\rangle_{L^2(q)}^{1/2},$$

(8)

which can be also seen as a kernelized Fisher divergence (Korba et al., 2021). Gorham & Mackey (2017b) proves that under mild conditions of target distribution $p(x)$ and kernel k, weak convergence of the KSD implies the weak convergence of distribution.

3. Proposed Method

In this section, we present kernel semi-implicit variational inference (KSIVI), which explicitly solves the lower-level optimization in SIVI-SM through kernel tricks. We will first give the expression of this explicit solution and then discuss the practical implementation of the upper-level optimization.

3.1. Training Objective

Instead of considering $f \in L^2(q_{\phi})$ which in general does not allow an explicit solution of f in (4), we seek the optimal f^* in an RKHS $\mathcal{H}$ and reformulate this minimax problem as

$$\min_{\phi} \max_{f \in \mathcal{H}} \ \mathbb{E}_{q_{\phi}(x)} \left[2f(x)^T [s_p(x) - s_{q_{\phi}}(x)] - \|f\|_{\mathcal{H}}^2 \right].$$

(9)

The following theorem shows that the solution f^* to the lower-level optimization in (9) has an explicit form, which allows us to reduce (9) to a standard optimization problem.

Theorem 3.1. *For any variational distribution q_{ϕ}, the solution f^* to the lower-level optimization in (9) takes the form*

$$f^*(x) = \mathbb{E}_{q_{\phi}(y)} k(x, y) \left[s_p(y) - s_{q_{\phi}}(y) \right].$$

(10)

Thus the upper-level optimization problem for ϕ is

$$\min_{\phi} \ KSD(q_{\phi}\|p)^2 = \left\| S_{q_{\phi},k} \nabla \log \frac{p}{q_{\phi}} \right\|_{\mathcal{H}}^2.$$

(11)

The detailed proof is deferred to Appendix A.1. Moreover, by taking advantage of the semi-implicit structure, we have

$$\mathbb{E}_{q_{\phi}(y)} k(x, y) s_{q_{\phi}}(y)$$

$$= \mathbb{E}_{q_{\phi}(y)} \frac{k(x, y)}{q_{\phi}(y)} \int \nabla_y q_{\phi}(y|z) q(z) \mathrm{d}z$$

$$= \int k(x, y) \nabla_y q_{\phi}(y|z) q(z) \mathrm{d}z \mathrm{d}y$$

$$= \int k(x, y) s_{q_{\phi}(\cdot|z)}(y) q_{\phi}(y|z) q(z) \mathrm{d}z \mathrm{d}y$$

$$= \mathbb{E}_{q_{\phi}(y,z)} k(x, y) s_{q_{\phi}(\cdot|z)}(y).$$

(12)

This leads to the following proposition.

Proposition 3.2. *The solution f^* in Theorem 3.1 can be rewritten as*

$$f^*(x) = \mathbb{E}_{q_{\phi}(y,z)} k(x, y) \left[s_p(y) - s_{q_{\phi}(\cdot|z)}(y) \right].$$

(13)

And the $KSD(q_\phi\|p)^2$ in (11) has an equivalent expression

$$KSD(q_\phi\|p)^2 = \mathbb{E}_{q_\phi(x,z),q_\phi(x',z')}\big[k(x,x')\cdot$$
$$\langle s_p(x) - s_{q_\phi(\cdot|z)}(x), s_p(x') - s_{q_\phi(\cdot|z')}(x')\rangle\big]. \quad (14)$$

Since the semi-implicit variational distribution $q_\phi(x,z)$ enables efficient sampling, both f^* and $KSD(q_\phi\|p)^2$ can be estimated using the Monte Carlo method with samples from $q_\phi(x,z)$. This way, we have transformed the minimax problem (9) into a standard optimization problem with a tractable objective function.

3.2. Practical Implementation

Suppose the conditional $q_\phi(x|z)$ admits the reparameterization trick (Kingma & Welling, 2014; Titsias & Lázaro-Gredilla, 2014; Rezende et al., 2014), i.e., $x \sim q_\phi(\cdot|z) \Leftrightarrow x = T_\phi(z,\xi), \xi \sim q_\xi$, where T_ϕ is a parameterized transformation and q_ξ is a base distribution that does not depend on ϕ. We now show how to find an optimal variational approximation $q_\phi(x)$ that minimizes the KSD between $q_\phi(x)$ and $p(x)$ defined in (14) using stochastic optimization.

To estimate the gradient $g(\phi) = \nabla_\phi KSD(q_\phi\|p)^2$, we consider two unbiased stochastic gradient estimators in our implementations. The first one is a vanilla gradient estimator using two batches of Monte Carlo samples $(x_{ri}, z_{ri}) \overset{\text{i.i.d.}}{\sim} q_\phi(x,z)$ where $r = 1, 2$ and $1 \le i \le N$, and it is defined as

$$\hat{g}_{\text{vanilla}}(\phi) = \frac{1}{N^2} \sum_{1 \le i,j \le N} \nabla_\phi \left[k(x_{1i}, x_{2j}) \langle f_{1i}, f_{2j} \rangle \right], \quad (15)$$

where $f_{ri} = s_p(x_{ri}) - s_{q_\phi(\cdot|z_{ri})}(x_{ri})$. We can also leverage U-statistics to design an alternative estimator (Gretton et al., 2012; Liu et al., 2016). Given one batch of samples $(x_i, z_i) \overset{\text{i.i.d.}}{\sim} q_\phi(x,z)$ for $1 \le i \le N$, the U-statistic gradient estimator takes the form

$$\hat{g}_{\text{u-stat}}(\phi) = \frac{2}{N(N-1)} \sum_{1 \le i<j \le N} \nabla_\phi \left[k(x_i, x_j) \langle f_i, f_j \rangle \right], \quad (16)$$

where $f_i = s_p(x_i) - s_{q_\phi(\cdot|z_i)}(x_i)$. Note that all the gradient terms in $\hat{g}_{\text{vanilla}}(\phi)$ and $\hat{g}_{\text{u-stat}}(\phi)$ can be efficiently evaluated using the reparameterization trick.

In our implementation, we assume a diagonal Gaussian conditional layer $q_\phi(x|z) = \mathcal{N}(\mu(z;\phi), \text{diag}\{\sigma^2(z;\phi)\})$ where the mean and standard deviation $\mu(z;\phi), \sigma(z;\phi) \in \mathbb{R}^d$ are parametrized by neural networks. Hence the parameterization formula of $q_\phi(x|z)$ is $x = \mu(z;\phi) + \sigma(z;\phi) \odot \xi$ with $\xi \sim q_\xi = \mathcal{N}(0, I)$, and the corresponding conditional score takes the form $s_{q_\phi(\cdot|z)}(x) = -\xi/\sigma(z;\phi)$. Note that in principle, one can sample multiple ξ for a single z_i to get a more robust estimation as done in Yin & Zhou (2018). In our experiments, we find that sampling a single ξ works

---

Algorithm 1 KSIVI with diagonal Gaussian conditional layer and vanilla gradient estimator

---

Input: target score $s_p(x)$, number of iterations T, number of samples N for stochastic gradient.
Output: the optimal variational parameters ϕ^*.
for $t = 0, \cdots, T-1$ **do**
    Sample $\{z_{r1}, \cdots, z_{rN}\}$ from mixing distribution $q(z)$ for $r = 1, 2$.
    Sample $\{\xi_{r1}, \cdots, \xi_{rN}\}$ from $\mathcal{N}(0, I)$ for $r = 1, 2$.
    Compute $x_{ri} = \mu(z_{ri};\phi) + \sigma(z_{ri};\phi) \odot \xi_{ri}$ and $f_{ri} = s_p(x_{ri}) + \frac{\xi_{ri}}{\sigma(z_{ri};\phi)}$.
    Compute stochastic gradient $\hat{g}_{\text{vanilla}}(\phi)$ through (15).
    Set $\phi \leftarrow \text{optimizer}(\phi, \hat{g}_{\text{vanilla}}(\phi))$.
end for
$\phi^* \leftarrow \phi$

---

well. The full training procedure of KSIVI with the vanilla gradient estimator (15) is described in Algorithm 1. The algorithm with U-statistic gradient estimator (16) is deferred to Appendix B. Compared to other variants of SIVI, our proposed KSIVI optimizes the KSD objective efficiently and stably with samples from the variational distribution, without using surrogates of exact ELBO, expensive inner-loop MCMC, or additional lower-level optimization.

4. Theoretical Results

In this section, we provide a theoretical guarantee for the convergence of KSIVI as a black-box variational inference (BBVI) problem (Ranganath et al., 2014; Titsias & Lázaro-Gredilla, 2014). This is non-trivial as the conventional assumptions in stochastic optimization are challenging to meet within the context of variational inference settings. Recent works (Kim et al., 2023b; Domke et al., 2023; Kim et al., 2023a) attempt to analyze the location-scale family, which has poor approximation capacity. For SIVI, Moens et al. (2021) assumes that $q(z)$ is a discrete measure with n components and gives a sub-optimal rate depending on $poly(n)$. We are not aware of any analysis of the convergence of the full version SIVI with nonlinear reparameterization.

Similar to prior works, we first investigate the smoothness of the loss function and prove an upper bound of the variance of the stochastic gradient. Then we apply the standard analysis in stochastic optimization, which ensures $\{\phi_t\}_{t \ge 1}$ converges to a stationary point.

Define the loss function as

$$\mathcal{L}(\phi) := \text{KSD}(q_\phi\|p)^2 = \left\| S_{q_\phi,k} \nabla \log \frac{p}{q_\phi} \right\|_{\mathcal{H}}^2. \quad (17)$$

We consider a diagonal Gaussian conditional layer. The conditional score function is $s_{q_\phi(\cdot|z)}(x) = -\frac{\xi}{\sigma(z;\phi)}$ for $x =$

$\mu(z; \phi) + \sigma(z; \phi) \odot \xi$ with $\xi \sim \mathcal{N}(0, I)$. We assume an unbiased gradient estimator in (15) or (16). An SGD update $\phi_{t+1} = \phi_t - \eta \hat{g}_t$ is then applied with learning rate η for T iterations, where $\hat{g}_t$ is the stochastic gradient in the t-th iteration.

Assumption 4.1. There exists a constant $B > 0$ such that $\forall x, y \in \mathbb{R}^d, \max\{k, \|\nabla_1 k\|, \|\nabla_{11} k\|, \|\nabla_{12} k\|\} \leq B^2$.

Assumption 4.2. $\log p$ is three times continuously differentiable and $\|\nabla^2 \log p(x)\| \leq L, \|\nabla^3 \log p(x)\| \leq M$.

Assumption 4.3. There exists a constant $G \geq 1$ such that $\max\{\|\nabla_\phi \mu(z; \phi)\|, \|\nabla_\phi^2 \mu(z; \phi)\|\} \leq G(1 + \|z\|)$. The same inequalities hold for $\sigma(z; \phi)$. Besides, for any z, ϕ, $\sigma(z; \phi)$ has a uniform lower bound $1/\sqrt{L}$.

As an application to location-scale family, where $\mu(z; \phi) \equiv \mu \in \mathbb{R}^d, \sigma(z; \phi) \equiv \sigma \in \mathbb{R}^d$, we have $G = 1$. For general nonlinear neural networks, similar first order smoothness is also assumed in the analysis of GANs (Arora et al., 2017). Note that we do not directly assume a uniformly bounded constant of smoothness of $\mu(z; \phi)$ and $\sigma(z; \phi)$, but it can grow with $\|z\|$. We find in practice this assumption is valid and the constant G is in a reasonable range. See Appendix A.3 for numerical evidence. The lower bound of $\sigma(z; \phi)$ is to avoid degeneracy and is required in Domke et al. (2023); Kim et al. (2023a), in which projected SGD is applied to ensure this. Similarly, our results can be easily extended to the setting of stochastic composite optimization and projected SGD optimizer, which we omit due to limited space.

Assumption 4.4. The mixing distribution $q(z)$ and the variational distribution $q_\phi(x)$ have bounded 4th-moment, i.e., $\mathbb{E}_{q(z)} \|z\|^4 \lesssim d_z^2, \mathbb{E}_{q_\phi(x)} \|x\|^4 \leq s^4$.

Overall, Assumption 4.1 guarantees that the kernel operator is well-defined and is easy to verify for commonly used Gaussian radial basis function (RBF) kernel and the inverse multi-quadratic (IMQ) kernel (Gorham & Mackey, 2017b). Assumption 4.2 and 4.3 ensures the smoothness of target distribution and variational distribution. Assumption 4.4 is a technical and reasonable assumption to bound the variance of stochastic gradient.

Based on Assumption 4.2 and 4.4, we can give a uniform upper bound on $\mathbb{E}_{q_\phi(x)} \|s_p(x)\|^4$, which is crucial to the subsequent analysis.

Proposition 4.5. *Under Assumption 4.2 and 4.4, we have* $\mathbb{E}_{q_\phi(x)} \|s_p(x)\|^4 \lesssim L^4(s^4 + \|x^*\|^4) := C^2$, *where x^* is any zero point of s_p.*

Theorem 4.6. *The objective $\mathcal{L}(\phi)$ is L_ϕ-smooth, where*

$$L_\phi \lesssim B^2 G^2 d_z \log d \left[(1 \vee L)^3 + Ld + M^2 + C\right]. \quad (18)$$

Theorem 4.7. *Both gradient estimators $\hat{g}_{vanilla}$ and $\hat{g}_{u\text{-}stat}$ have bounded variance $\Sigma = \frac{\Sigma_0}{N}$, where*

$$\Sigma_0 \lesssim B^4 G^2 d_z \log d [L^3 d + L^2 d^2 + C^2]. \quad (19)$$

Theorem 4.8. *Under Assumption 4.1-4.4, iterates from SGD update $\phi_{t+1} = \phi_t - \eta \hat{g}_t$ with proper learning rate η include an ε-stationary point $\hat{\phi}$ such that $\mathbb{E}[\|\nabla_\phi \mathcal{L}(\hat{\phi})\|] \leq \varepsilon$, if*

$$T \gtrsim \frac{L_\phi \mathcal{L}_0}{\varepsilon^2} \left(1 + \frac{\Sigma_0}{N \varepsilon^2}\right), \quad (20)$$

where $\mathcal{L}_0 := \mathcal{L}(\phi_0) - \inf_\phi \mathcal{L}$.

The complete proofs are deferred to Appendix A.2. Theorem 4.6 and 4.7 imply that the optimization problem $\min_\phi \mathcal{L}(\phi)$ satisfies the standard assumptions in non-convex optimization literature. Therefore, classic results of SGD can be applied to our scenario (Ghadimi & Lan, 2013).

Theorem 4.8 implies that the sequence $\{\phi_t\}$ converge to a stationary point of the objective $\mathcal{L}(\phi)$. Since the training objective is squared KSD instead of traditional ELBO in BBVI, $\mathcal{L}(\phi)$ is nonconvex even for location-scale family. The convergence in loss function $\mathcal{L}$ is generally inaccessible. We hope future works can shed more insights into this issue.

5. Related Works

To address the limitation of standard VI, implicit VI constructs a flexible variational family through non-invertible mappings parameterized by neural networks. However, the main issue therein is density ratio estimation, which is difficult in high dimensions (Sugiyama et al., 2012). Besides SIVI, there are a number of recent advancements in this field. Molchanov et al. (2019) have further extended SIVI in the context of generative models. Sobolev & Vetrov (2019) introduce a new surrogate of ELBO through importance sampling. UIVI (Titsias & Ruiz, 2019) runs inner-loop MCMC to get an unbiased estimation of the gradient of ELBO. KIVI (Shi et al., 2018a) constrains the density ratio estimation within an RKHS. LIVI (Uppal et al., 2023) approximates the intractable entropy with a Gaussian distribution by linearizing the generator.

Besides ELBO-based training of VI, many works consider to minimize Fisher divergence or its variants (Yu & Zhang, 2023; Ranganath et al., 2016; Grathwohl et al., 2020; Dong et al., 2022; Cheng et al., 2023). Leveraging the minimax formulation of Fisher divergence, these methods try to alternatively optimize the variational distribution and an adversarial function in certain function classes, typically neural networks. Although it does not rely on surrogates of ELBO, the lower-level optimization can not be done accurately in general, thus inducing unavoidable bias for the training of variational distribution. More importantly, it involves expensive extra computation during training.

To fix this computation issue, another closely related work (Korba et al., 2021) proposes a particle-based VI method, KSDD, which follows the KSD flow to minimize KSD, a kernelized version of Fisher divergence. It utilizes the fact
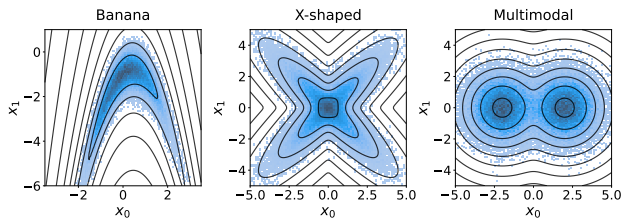
Figure 1. **Performances of KSIVI on toy examples**. The histplots in blue represent the estimated densities using 100,000 samples generated from KSIVI's variational approximation. The black lines depict the contour of the target distributions.
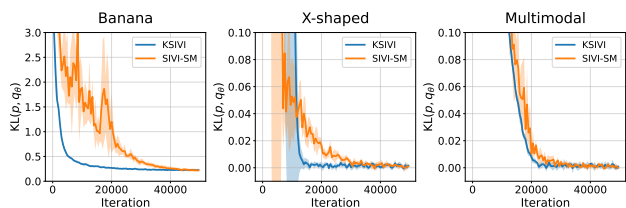

Figure 2. **Convergence of KL divergence during training obtained by different methods on toy examples.** The KL divergences are estimated using the Python ITE module (Szabó, 2014) with 100,000 samples. The results are averaged over 5 independent computations with the standard deviation as the shaded region.

that the Wasserstein gradient of KSD can be directly estimated by the empirical particle distribution. Korba et al. (2021) also discusses the theoretical properties of KSDD. However, our formulation and the "denoising" derivation are not naive extensions, since KSIVI only requires the function value of kernel, while KSDD relies on twice derivatives of the kernel. We believe this is a crucial point that distinguishes KSIVI, since less smooth kernels can be leveraged, which may have stronger power as a distance, e.g. Riesz kernel (Altekrüger et al., 2023).

As for the convergence guarantee of BBVI, previous works mainly focus on ELBO objectives and location-scale families. In particular, Domke (2019) proves the bound of the gradient variance and later Domke (2020) shows the smoothness guarantee of the loss function, both of which are under various structural assumptions. The first self-contained analysis of the convergence of location-scale family has been recently proposed by Kim et al. (2023b); Domke et al. (2023), where some important variants like STL gradient estimator and proximal gradient descent are also discussed. Still, there is no theoretical convergence guarantee of (semi) implicit VI to the best of our knowledge.

6. Experiments

In this section, we compare KSIVI to the ELBO-based method SIVI and the score-based method SIVI-SM on toy

examples and real-data problems. For the construction of the semi-implicit variational family in all these methods, we choose a standard Gaussian mixing distribution and diagonal Gaussian conditional layer (see Section 3.2) whose standard deviation $\sigma(z, \phi) = \phi_\sigma \in \mathbb{R}^d$ does not depend on z. Following the approach by Liu & Wang (2016), we dynamically set the kernel width, to the median value of variational samples' spacing. In all experiments, we use the Gaussian RBF kernel in KSIVI, following Liu & Wang (2016); Liu et al. (2016). Throughout this section, we use the vanilla gradient estimator for KSIVI, and results of the U-statistic gradient estimator can be found in Appendix C. All the experiments are implemented in PyTorch (Paszke et al., 2019). More implementation details can be found in Appendix C and https://github.com/longinYu/KSIVI.

6.1. Toy Examples

We first conduct toy experiments on approximating three two-dimensional distributions: BANANA, MULTIMODAL, and X-SHAPED, whose probability density functions are in Table 4 in Appendix C.1. We consider a temperature annealing strategy (Rezende & Mohamed, 2015) on MULTIMODAL to facilitate exploration. The results are collected after 50,000 iterations with a learning rate of 0.001 for all the methods.

Figure 1 shows approximation performances of KSIVI on the toy examples. We see that KSIVI provides favorable variational approximations for the target distributions. Figure 2 displays the KL divergence as a function of the number of iterations obtained by KSIVI and SIVI-SM. Compared to SIVI-SM, KSIVI is more stable in training, partly because it does not require additional lower-level optimization. KSIVI also tends to converge faster than SIVI-SM, although this advantage becomes less evident as the target distribution gets more complicated. Figure 8 in Appendix C.1 illustrates the convergence of maximum mean discrepancy (MMD) during training obtained by KSIVI and SIVI-SM, which mostly aligns with the behavior of KL divergence. We also conduct experiments with the IMQ kernel (Gorham & Mackey, 2017b), and do not notice a significant difference w.r.t. the Gaussian RBF kernel, which is consistent with the findings in Korba et al. (2021).

6.2. Bayesian Logistic Regression

Our second experiment is on the Bayesian logistic regression problem with the same experimental setting in Yin & Zhou (2018). Given the explanatory variable $x_i \in \mathbb{R}^d$ and the observed binary response variable $y_i \in \{0, 1\}$, the log-likelihood function takes the form

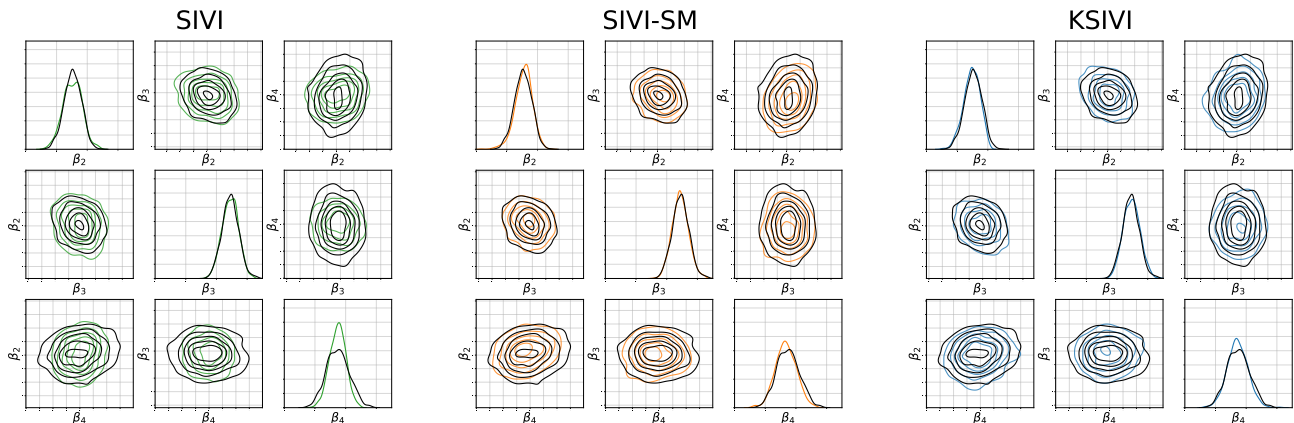$$\log p(y_i | x_i', \beta) = y_i \beta^T \bar{x}_i - \log(1 + \exp(\beta^T \bar{x}_i)),$$

Figure 3. **Marginal and pairwise variational approximations of** $\beta_2, \beta_3, \beta_4$ **on the Bayesian logistic regression task.** The contours of the pairwise posterior approximation produced by SIVI-SM (in orange), SIVI (in green), and KSIVI (in blue) are graphed in comparison to the ground truth (in black). The sample size is 1000.
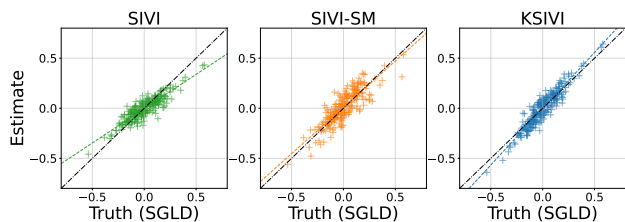

Figure 4. **Comparison between the estimated pairwise correlation coefficients and the ground truth on the Bayesian logistic regression task.** Each scatter represents the estimated correlation coefficient (y-axis) and the ground truth correlation coefficient (x-axis) of some pair (β_i, β_j). The lines in the same color as the scatters represent the regression lines. The sample size is 1000.

where $\bar{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix} \in \mathbb{R}^{d+1}$ is the covariate and $\beta \in \mathbb{R}^{d+1}$ is the variable we want to infer. The prior distribution of β is set to $p(\beta) = \mathcal{N}(0, \alpha^{-1}I)$ where the inverse variance $\alpha = 0.01$. We consider the WAVEFORM[1] dataset of $\{x_i, y_i\}_{i=1}^N$ where the dimension of the explanatory variable x_i is $d = 21$. Then different SIVI variants with the same architecture of semi-implicit variational family are applied to infer the posterior distribution $p(\beta | \{x_i, y_i\}_{i=1}^N)$. The learning rate for variational parameters ϕ is chosen as 0.001 and the batch size of particles is chosen as 100 during the training. For all the SIVI variants, the results are collected after 40,000 parameter updates. The ground truth consisting of 1000 samples is established by simulating parallel stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011) with 400,000 iterations, 1000 independent particles, and a small step size of 0.0001. Additionally, we assessed the performance of parallel Metropolis-adjusted Langevin

algorithm (MALA) (Rossky et al., 1978). The calculated KL divergence between samples from MALA and SGLD is 0.0289, suggesting their proximity.

Figure 3 demonstrates the marginal and pairwise posterior approximations for $\beta_2, \beta_3, \beta_4$ obtained by the aforementioned SIVI variants in comparison to the ground truth. We see that KSIVI agrees well with the ground truth and performs comparably to SIVI-SM. Notice that SIVI-SM also requires tuning additional hyper-parameters (e.g., the learning rate of $f_\psi(x)$ and the number of lower-level gradient steps), which may be challenging as commonly observed in minimax optimization (Goodfellow et al., 2014; Mescheder et al., 2017). In contrast, SIVI slightly underestimates the variance of β_4 in both marginal and pairwise joint distributions, as shown by the left plot in Figure 3. The results for more components of β can be found in Figure 9 in Appendix C.2. Additionally, we also investigate the pairwise correlation coefficients of β defined as

$$\boldsymbol{\rho} = \left\{ \rho_{i,j} := \frac{\text{cov}(\beta_i, \beta_j)}{\sqrt{\text{cov}(\beta_i, \beta_i)\text{cov}(\beta_j, \beta_j)}} \right\}_{1 \le i < j \le 22}$$

and compare the estimated pairwise correlation coefficients produced by different methods to the ground truth in Figure 4. We see that KSIVI provides better correlation coefficient approximations than SIVI and SIVI-SM, as evidenced by the scatters more concentrated around the diagonal.

6.3. Conditioned Diffusion Process

Our next example is a higher-dimensional Bayesian inference problem arising from the following Langevin stochastic differential equation (SDE) with state $x_t \in \mathbb{R}$

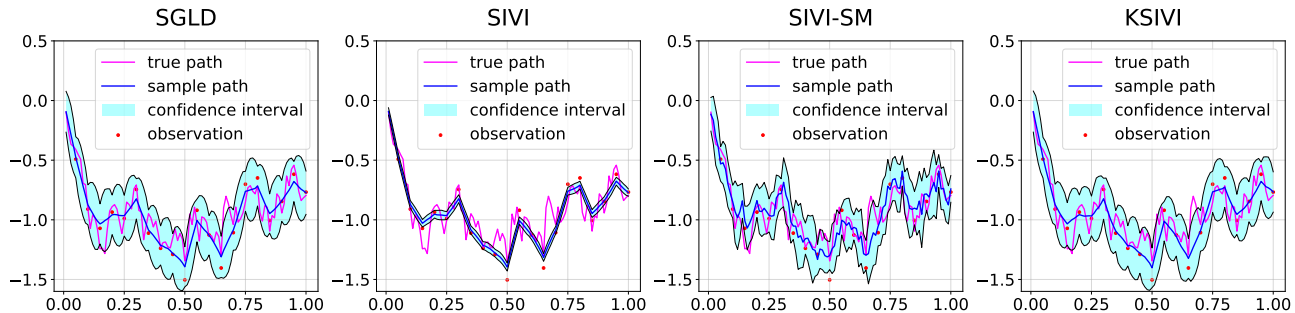$$dx_t = 10x_t(1 - x_t^2)dt + dw_t, \ 0 \le t \le 1, \quad (21)$$

---

[1]https://archive.ics.uci.edu/ml/machine-learning-databases/waveform

Figure 5. **Variational approximations of different methods for the discretized conditioned diffusion process.** The magenta trajectory represents the ground truth via parallel SGLD. The blue line corresponds to the estimated posterior mean of different methods, and the shaded region denotes the 95% marginal posterior confidence interval at each time step. The sample size is 1000.

where $x_0 = 0$ and w_t is a one-dimensional standard Brownian motion. Equation (21) describes the motion of a particle with negligible mass trapped in an energy potential, with thermal fluctuations represented by the Brownian forcing (Detommaso et al., 2018; Cui et al., 2016; Yu et al., 2023). Using the Euler-Maruyama scheme with step size $\Delta t = 0.01$, we discretize the SDE into $x = (x_{\Delta t}, x_{2\Delta t}, \cdots, x_{100\Delta t})$, which defines the prior distribution $p_{\text{prior}}(x)$ of the 100-dimensional variable x. The perturbed 20-dimensional observation is $y = (y_{5\Delta t}, y_{10\Delta t}, \ldots, y_{100\Delta t})$ where $y_{5k\Delta t} \sim \mathcal{N}(x_{5k\Delta t}, \sigma^2)$ with $1 \leq k \leq 20$ and $\sigma = 0.1$, which gives the likelihood function $p(y|x)$. Given the perturbed observations y, our goal is to infer the posterior of the discretized path of conditioned diffusion process $p(x|y) \propto p_{\text{prior}}(x)p(y|x)$. We simulate a long-run parallel SGLD of 100,000 iterations with 1000 independent particles and a small step size of 0.0001 to form the ground truth of 1000 samples. For all SIVI variants, we update the variational parameters ϕ for 100,000 iterations to ensure convergence (Appendix C.3).

Table 2 shows the training time per 10,000 iterations of SIVI variants on a 3.2 GHz CPU. For a fair time comparison, we use the score-based training of SIVI discussed in Yu et al. (2023), which computes the $\nabla \log p(x)$ instead of $\log p(x)$ to derive the gradient estimator. We see that KSIVI achieves better computational efficiency than SIVI-SM and comparable training time to SIVI. Figure 5 shows the approximation for the discretized conditional diffusion process of all methods. We see that the posterior mean estimates given by SIVI-SM are considerably bumpier compared to the ground truth, while SIVI fails to capture the uncertainty with a severely underestimated variance. In contrast, the results from KSIVI align well with the SGLD ground truth.

6.4. Bayesian Neural Network

In the last experiment, we compare KSIVI against SGLD, SIVI, and SIVI-SM on sampling from the posterior of a

Bayesian neural network (BNN) across various benchmark UCI datasets[2]. Following Liu & Wang (2016); Wang & Li (2022), we use a two-layer neural network consisting of 50 hidden units and ReLU activation functions for the BNN model. The datasets are randomly partitioned into 90% for training and 10% for testing. Additionally, gradient clipping is applied during the training process, and the exponential moving average (EMA) trick (Huang et al., 2017; Izmailov et al., 2019) is employed in the inference stage for all SIVI variants.

The averaged test rooted mean squared error (RMSE) and negative log-likelihood (NLL) over 10 random runs are reported in Table 1. We see that KSIVI can provide results on par with SIVI and SIVI-SM on all datasets. However, it does not exhibit a significant advantage within the context of Bayesian neural networks. This phenomenon is partially attributable to the evaluation metrics. Since test RMSE and test NLL are assessed from a predictive standpoint rather than directly in terms of the posterior distribution, they do not fully capture the approximation accuracy of the posterior distribution. Additionally, the performance of KSIVI may be contingent on the choice of kernels, particularly in higher-dimensional cases, as suggested by research on unbounded kernels (Hertrich et al., 2024). We leave a more thorough investigation for future work.

7. Ablation Study

Kernel Choice in Heavy Tails Distribution Considering the pivotal role of kernel function selection in the efficacy of kernel-based methods(Gorham & Mackey, 2017a), we conducted an ablation study on a two-dimmensional distribution constructed as the product of two Student's t-distributions, following the setting described by Li et al. (2023). For KSIVI, we apply a practical regularization fix to thin the kernel Stein discrepancy across all the kernel

---

[2]https://archive.ics.uci.edu/ml/datasets.php

Table 1. **Test RMSE and test NLL of Bayesian neural networks on several UCI datasets**. The results are averaged from 10 independent runs with the standard deviation in the subscripts. For each data set, the best result is marked in **black bold font** and the second best result is marked in **brown bold font**.

Dataset	Test RMSE ($\downarrow$)				Test NLL ($\downarrow$)			
	SIVI	SIVI-SM	SGLD	KSIVI	SIVI	SIVI-SM	SGLD	KSIVI
BOSTON	$2.621_{\pm 0.02}$	$2.785_{\pm 0.03}$	$2.857_{\pm 0.11}$	$\mathbf{2.555}_{\pm 0.02}$	$\mathbf{2.481}_{\pm 0.00}$	$2.542_{\pm 0.01}$	$3.094_{\pm 0.01}$	$2.506_{\pm 0.01}$
CONCRETE	$6.932_{\pm 0.02}$	$5.973_{\pm 0.04}$	$6.861_{\pm 0.19}$	$\mathbf{5.750}_{\pm 0.03}$	$3.337_{\pm 0.00}$	$\mathbf{3.229}_{\pm 0.01}$	$4.036_{\pm 0.01}$	$3.309_{\pm 0.01}$
POWER	$\mathbf{3.861}_{\pm 0.01}$	$4.009_{\pm 0.00}$	$3.916_{\pm 0.01}$	$3.868_{\pm 0.01}$	$2.791_{\pm 0.00}$	$2.822_{\pm 0.00}$	$2.944_{\pm 0.00}$	$2.797_{\pm 0.00}$
WINE	$0.597_{\pm 0.00}$	$0.605_{\pm 0.00}$	$0.597_{\pm 0.00}$	$\mathbf{0.595}_{\pm 0.00}$	$0.904_{\pm 0.00}$	$0.916_{\pm 0.00}$	$0.904_{\pm 0.00}$	$\mathbf{0.901}_{\pm 0.00}$
YACHT	$1.505_{\pm 0.07}$	$\mathbf{0.884}_{\pm 0.01}$	$2.152_{\pm 0.09}$	$1.237_{\pm 0.05}$	$1.721_{\pm 0.03}$	$\mathbf{1.432}_{\pm 0.01}$	$2.873_{\pm 0.03}$	$1.752_{\pm 0.03}$
PROTEIN	$\mathbf{4.669}_{\pm 0.00}$	$5.087_{\pm 0.00}$	$4.777_{\pm 0.00}$	$5.027_{\pm 0.01}$	$\mathbf{2.967}_{\pm 0.00}$	$3.047_{\pm 0.00}$	$2.984_{\pm 0.00}$	$3.034_{\pm 0.00}$

Table 2. **Training time (per 10,000 iterations, in seconds) for the conditioned diffusion process inference task.** For all the methods, the batch size for Monte Carlo estimation is set to $N = 128$.

Method \Dimensionality	50	100	200
SIVI	58.00	88.12	113.46
SIVI-SM	70.42	128.13	149.61
KSIVI	56.67	90.48	107.84

Table 3. **Estimated Wasserstein distances for the product of two Student's t-distributions.** The results were averaged from 10 independent runs.

Methods \Edge width	5	8	10
MIED	$\mathbf{0.0366}_{\pm 0.01}$	$\mathbf{0.0778}_{\pm 0.03}$	$0.1396_{\pm 0.04}$
KSDD	$0.1974_{\pm 0.04}$	$0.3187_{\pm 0.10}$	$0.3910_{\pm 0.11}$
KSIVI (Gaussian kernel)	$0.1048_{\pm 0.03}$	$0.1976_{\pm 0.08}$	$0.2546_{\pm 0.09}$
KSIVI (Reisz kernel)	$0.0451_{\pm 0.01}$	$0.0816_{\pm 0.04}$	$\mathbf{0.1136}_{\pm 0.05}$

functions(Bénard et al., 2023). We use the implementation from Li et al. (2023) to reproduce the results for two particle-based variational inference methods MIED (Li et al., 2023) and KSDD (Korba et al., 2021). In table 3, we report the estimated Wasserstein distances from the target distributions to the variational posteriors (using the metric implementation provided in (Li et al., 2023) with 1000 samples). We found that compared to the Gaussian kernel, the Reisz kernel leads to a more pronounced improvement for this task.

8. Conclusion

This paper proposed a novel framework of semi-implicit variational inference, KSIVI, which takes the KSD as the training objective instead of the ELBO or Fisher divergence. As a variant of SIVI-SM, KSIVI eliminates the need for lower-level optimization using the kernel trick. By taking advantage of the hierarchical structure of the semi-implicit variational family, the KSD objective is readily computable. We provided efficient Monte Carlo gradient estimators for stochastic optimization. We also derived an upper bound for the variance of the Monte Carlo gradient estimators,

which allows us to establish the convergence guarantee to a stationary point using techniques in stochastic optimization. Extensive numerical experiments validate the efficiency and effectiveness of KSIVI.

9. Limitation

Our current work has several limitations. Firstly, since KSIVI uses KSD to measure dissimilarity between distributions, it may encounter drawbacks like stationary points that differ from the target distributions, especially in high-dimensional, non-convex tasks. Secondly, our use of the Gaussian kernel in KSIVI may be less effective in high dimensions due to its fast-decaying tails. Exploring unbounded kernel functions could be a promising direction for future research. Additionally, our theoretical analysis depends on strong assumptions, such as the Lipschitz smoothness of the neural network, which may be too strict for deep neural networks in high dimensions. Investigating simpler kernel machines or random feature regression models as alternatives could also be valuable.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Altekrüger, F., Hertrich, J., and Steidl, G. Neural wasserstein gradient flows for maximum mean discrepancies with riesz kernels. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:256358344.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pp. 224–232. PMLR, 2017.

Bishop, C. M. and Tipping, M. E. Variational relevance vector machines. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pp. 46–53, 2000.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859 – 877, 2016.

Bénard, C., Staber, B., and Veiga, S. D. Kernel stein discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization, 2023.

Cheng, Z., Zhang, S., Yu, L., and Zhang, C. Particle-based variational inference with generalized wasserstein gradient flow. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR.

Cui, T., Law, K. J., and Marzouk, Y. M. Dimension-independent likelihood-informed mcmc. *Journal of Computational Physics*, 304:109–137, 2016.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. A stein variational newton method. *Advances in Neural Information Processing Systems*, 31, 2018.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.

Domke, J. Provable gradient variance guarantees for black-box variational inference. *Advances in Neural Information Processing Systems*, 32, 2019.

Domke, J. Provable smoothness guarantees for black-box variational inference. In *International Conference on Machine Learning*, pp. 2587–2596. PMLR, 2020.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Domke, J., Garrigos, G., and Gower, R. Provable convergence guarantees for black-box variational inference. *arXiv preprint arXiv:2306.03638*, 2023.

Dong, H., Wang, X., Yong, L., and Zhang, T. Particle-based variational inference with preconditioned functional gradient flow. In *The Eleventh International Conference on Learning Representations*, 2022.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Giordano, R. J., Broderick, T., and Jordan, M. I. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing System 27 (NIPS)*, 2014.

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

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1292–1301. PMLR, 06–11 Aug 2017a.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017b.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pp. 3732–3747. PMLR, 2020.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Hertrich, J., Wald, C., Altekrüger, F., and Hagemann, P. Generative sliced MMD flows with Riesz kernels. In *The Twelfth International Conference on Learning Representations*, 2024.

Hinton, G. E. and van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Annual Conference Computational Learning Theory*, 1993. URL https://api.semanticscholar.org/CorpusID:9346534.

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

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free, 2017.

Huszár, F. Variational inference using implicit distributions. *arXiv preprint arXiv: 1702.08235*, 2017.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization, 2019.

Jaakkola, T. S. and Jordan, M. I. Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, pp. 173–173, 1998.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Kim, K., Ma, Y., and Gardner, J. R. Linear convergence of black-box variational inference: Should we stick the landing? *arXiv preprint arXiv:2307.14642*, 2023a.

Kim, K., Oh, J., Wu, K., Ma, Y., and Gardner, J. R. On the convergence of black-box variational inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. Kernel stein discrepancy descent. In *International Conference on Machine Learning*, pp. 5719–5730. PMLR, 2021.

Li, L., qiang liu, Korba, A., Yurochkin, M., and Solomon, J. Sampling with mollified interaction energy descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zWy7dqOcel.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016.

Mescheder, L. M., Nowozin, S., and Geiger, A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 2017.

Moens, V., Ren, H., Maraval, A., Tutunov, R., Wang, J., and Ammar, H. Efficient semi-implicit variational inference. *arXiv preprint arXiv:2101.06070*, 2021.

Molchanov, D., Kharitonov, V., Sobolev, A., and Vetrov, D. Doubly semi-implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2593–2602. PMLR, 2019.

Nott, D. J., Tan, S. L., Villani, M., and Kohn, R. Regression density estimation with variational methods and stochastic approximation. *Journal of Computational and Graphical Statistics*, 21(3):797–820, 2012.

Paisley, J. W., Blei, D. M., and Jordan, M. I. Variational bayesian inference with stochastic search. In *Proceedings of the 29th International Conference on Machine Learning ICML*, 2012.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Peterson, C. and Hartman, E. Explorations of the mean field theory learning algorithm. *Neural Networks*, 2(6):475–494, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90045-2. URL https://www.sciencedirect.com/science/article/pii/0893608089900452.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822. PMLR, 2014.

Ranganath, R., Tran, D., Altosaar, J., and Blei, D. Operator variational inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, pp. 1530–1538, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Rossky, P. J., Doll, J. D., and Friedman, H. L. Brownian dynamics as smart monte carlo simulation. *Journal of Chemical Physics*, 69:4628–4633, 1978. URL https://api.semanticscholar.org/CorpusID:14967940.

Saul, L. K. and Jordan, M. I. Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems*, 1996.

Shi, J., Sun, S., and Zhu, J. Kernel implicit variational inference. In *International Conference on Learning Representations*, 2018a.

Shi, J., Sun, S., and Zhu, J. A spectral approach to gradient estimation for implicit distributions. In *International Conference on Machine Learning*, 2018b.

Sobolev, A. and Vetrov, D. P. Importance weighted hierarchical variational inference. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2019.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Szabó, Z. Information theoretical estimators toolbox. *Journal of Machine Learning Research*, 15:283–287, 2014.

Titsias, M. and Lázaro-Gredilla, M. Doubly stochastic variational bayes for non-conjugate inference. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1971–1979, Bejing, China, 22–24 Jun 2014. PMLR.

Titsias, M. K. and Ruiz, F. Unbiased implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 167–176. PMLR, 2019.

Tran, D., Blei, D. M., and Airoldi, E. M. Copula variational inference. In *Advances in Neural Information Processing Systems*, 2015.

Tran, D., Ranganath, R., and Blei, D. M. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, 2017.

Uppal, A., Stensbo-Smidt, K., Boomsma, W. K., and Frellsen, J. Implicit variational inference for high-dimensional posteriors. *arXiv preprint arXiv:2310.06643*, 2023.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Maching Learning*, 1(1-2):1–305, 2008.

Wang, Y. and Li, W. Accelerated information gradient flow, 2022.

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Yin, M. and Zhou, M. Semi-implicit variational inference. In *International Conference on Machine Learning*, pp. 5660–5669. PMLR, 2018.

Yu, L. and Zhang, C. Semi-implicit variational inference via score matching. *arXiv preprint arXiv:2308.10014*, 2023.

Yu, L., Xie, T., Zhu, Y., Yang, T., Zhang, X., and Zhang, C. Hierarchical semi-implicit variational inference with application to diffusion model acceleration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ghIBaprxsV.

Zhang, C., Shahbaba, B., and Zhao, H. Variational hamiltonian monte carlo via score matching. *Bayesian Analysis*, 13(2):485–506, 2018.

A. Proofs

A.1. Proof of Theorem 3.1

Theorem A.1. *Consider the min-max problem*

$$\min_{\phi} \max_{f} \quad \mathbb{E}_{q_\phi} \left[2f(x)^T [s_p(x) - s_{q_\phi}(x)] - \|f\|_{\mathcal{H}}^2 \right]. \tag{22}$$

Given variational distribution q_ϕ, the optimal f^* is given by

$$f^*(x) = \mathbb{E}_{y \sim q_\phi(y)} k(x, y) \left[s_p(y) - s_{q_\phi}(y) \right]. \tag{23}$$

Thus the upper-level problem for ϕ is

$$\min_{\phi} \quad KSD(q_\phi \| p)^2 = \left\| S_{q_\phi, k} \nabla \log \frac{p}{q_\phi} \right\|_{\mathcal{H}}^2. \tag{24}$$

Proof. For any $f \in \mathcal{H}$, we have $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ by reproducing property. Since

$$\begin{aligned}
\mathbb{E}_{q_\phi} f(x)^T [s_p(x) - s_{q_\phi}(x)] &= \mathbb{E}_{q_\phi} \langle f, k(\cdot, x) \rangle_{\mathcal{H}}^T [s_p(x) - s_{q_\phi}(x)] \\
&= \left\langle f, \mathbb{E}_{q_\phi} k(\cdot, x) [s_p(x) - s_{q_\phi}(x)] \right\rangle_{\mathcal{H}} \\
&= \left\langle f, S_{q_\phi, k} \nabla \log \frac{p}{q_\phi} \right\rangle_{\mathcal{H}},
\end{aligned} \tag{25}$$

the lower-level problem is

$$\max_{f} \quad 2 \left\langle f, S_{q_\phi, k} \nabla \log \frac{p}{q_\phi} \right\rangle_{\mathcal{H}} - \|f\|_{\mathcal{H}}^2. \tag{26}$$

Therefore, the optimal $f^* = S_{q_\phi, k} \nabla \log \frac{p}{q_\phi}$ and the upper-level problem is

$$\min_{\phi} \quad \left\| S_{q_\phi, k} \nabla \log \frac{p}{q_\phi} \right\|_{\mathcal{H}}^2. \tag{27}$$

$\square$

A.2. Proofs in Section 4

Proposition A.2. *Under Assumption 4.2 and 4.4, we have $\mathbb{E}_{q_\phi(x)} \|s_p(x)\|^4 \lesssim L^4(s^4 + \|x^*\|^4) := C^2$, where x^* is any zero point of s_p.*

Proof. By AM-GM inequality, and since x^* is a zero point of s_p,

$$\mathbb{E}_{q_\phi(x)} \|s_p(x)\|^4 = \mathbb{E}_{q_\phi(x)} \|s_p(x) - s_p(x^*)\|^4 \lesssim L^4 (\mathbb{E}_{q_\phi(x)} \|x\|^4 + \|x^*\|^4) \lesssim L^4(s^4 + \|x^*\|^4). \tag{28}$$

$\square$

Theorem A.3. *The objective $\mathcal{L}(\phi)$ is L_ϕ-smooth, where $L_\phi \lesssim B^2 G^2 d_z \log d \left[(1 \vee L)^3 + Ld + M^2 + \mathbb{E}\|s_p(x)\|^2 \right]$.*

Proof. Let $u = (z, \xi)$ and $x = T(u; \phi) = \mu(z; \phi) + \sigma(z; \phi) \odot \xi$. Then the objective has the following representation:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{z, z'} \mathbb{E}_{q_\phi(x|z), q_\phi(x'|z')} \left[k(x, x')(s_p(x) - s_{q_\phi(\cdot|z)}(x))^T (s_p(x') - s_{q_\phi(\cdot|z')}(x')) \right] \\
&= \mathbb{E}_{u, u'} \left[k(x, x')(s_p(x) + \xi/\sigma)^T (s_p(x') + \xi'/\sigma') \right],
\end{aligned} \tag{29}$$

which implies

$$
\begin{aligned}
\nabla_\phi^2 \mathcal{L} = \mathbb{E} \Bigg\{ & \underbrace{\nabla_\phi^2[k(x,x')](s_p(x) + \xi/\sigma)^T(s_p(x') + \xi'/\sigma')}_{\textcircled{1}} \\
& + \underbrace{\nabla_\phi k(x,x') \otimes \nabla_\phi \left[(s_p(x) + \xi/\sigma)^T(s_p(x') + \xi'/\sigma')\right] + \nabla_\phi \left[(s_p(x) + \xi/\sigma)^T(s_p(x') + \xi'/\sigma')\right] \otimes \nabla_\phi k(x,x')}_{\textcircled{2}} \\
& + \underbrace{k(x,x')\nabla_\phi^2 \left[(s_p(x) + \xi/\sigma)^T(s_p(x') + \xi'/\sigma')\right]}_{\textcircled{3}} \Bigg\} .
\end{aligned}
\tag{30}
$$

For $\textcircled{1}$, we have

$$
\begin{aligned}
\nabla_\phi^2[k(x,x')] = \nabla_\phi^2 x \nabla_1 k + \nabla_\phi^T x \left(\nabla_{11} k \nabla_\phi x + \nabla_{21} k \nabla_\phi x'\right) \\
+ \nabla_\phi^2 x' \nabla_2 k + \nabla_\phi^T x' \left(\nabla_{22} k \nabla_\phi x' + \nabla_{12} k \nabla_\phi x\right).
\end{aligned}
\tag{31}
$$

Here $\nabla_\phi^2 x \nabla_1 k$ is a matrix with entry $[\nabla_\phi^2 x \nabla_1 k]_{ij} = \sum_{l=1}^d \nabla_\phi^2 x_l \nabla_{x_l} k$.

Note that by Assumption 4.3, $\|\nabla_\phi x\| \le \|\nabla_\phi \mu\| + \|\nabla_\phi \sigma \odot \xi\| \le G(1 + \|z\|)(1 + \|\xi\|_\infty)$ and $\|\nabla_\phi^2 x\| \le \|\nabla_\phi^2 \mu\| + \|\nabla_\phi^2 \sigma \odot \xi\| \le G(1 + \|z\|)(1 + \|\xi\|_\infty)$. Then by Assumption 4.1, $\|\nabla_1 k\| \le B^2, \|\nabla_{11} k\| \le B^2$,

$$
\begin{aligned}
\mathbb{E}[\|\textcircled{1}\|] &\le \mathbb{E}[\|\nabla_\phi^2 k\| \cdot \|s_p(x) + \xi/\sigma\| \|s_p(x') + \xi'/\sigma'\|] \\
&\le \sqrt{\mathbb{E}[\|\nabla_\phi^2 k\|^2]} \sqrt{\mathbb{E}[\|s_p(x) + \xi/\sigma\|^2 \|s_p(x') + \xi'/\sigma'\|^2]} \\
&\lesssim \left[GB^2 \sqrt{\mathbb{E}(1 + \|\xi\|_\infty)^2(1 + \|z\|)^2} + G^2 B^2 \sqrt{\mathbb{E}(1 + \|\xi\|_\infty)^4(1 + \|z\|)^4}\right] \mathbb{E}[\|s_p(x) + \xi/\sigma\|^2] \\
&\lesssim G^2 B^2 d_z \log d \, \mathbb{E}[\|s_p(x) + \xi/\sigma\|^2] \\
&\lesssim G^2 B^2 d_z \log d \, [\mathbb{E}\|s_p(x)\|^2 + Ld].
\end{aligned}
\tag{32}
$$

In the inequality second to last, we utilize the well-known fact $\mathbb{E}\|\xi\|_\infty^4 \lesssim \log^2 d$ and Assumption 4.4.

For $\textcircled{2}$, we have

$$
\begin{aligned}
\|\nabla_\phi k(x,x')\| &\le \|\nabla_\phi^T x \nabla_1 k\| + \|\nabla_\phi^T x' \nabla_2 k\| \\
&\lesssim B^2 G[(1 + \|z\|)(1 + \|\xi\|_\infty) + (1 + \|z'\|)(1 + \|\xi'\|_\infty)],
\end{aligned}
\tag{33}
$$

and

$$
\begin{aligned}
\nabla_\phi \left[(s_p(x) + \xi/\sigma)^T(s_p(x') + \xi'/\sigma')\right] &= \left[\nabla^2 \log p(x)\nabla_\phi x - diag\{\xi/\sigma^2\}\nabla_\phi \sigma\right]^T (s_p(x') + \xi'/\sigma') \\
&+ \left[\nabla^2 \log p(x')\nabla_\phi x' - diag\{\xi'/\sigma'^2\}\nabla_\phi \sigma'\right]^T (s_p(x) + \xi/\sigma).
\end{aligned}
\tag{34}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}[\|\textcircled{2}\|] &\lesssim B^2 G \sqrt{\mathbb{E}[(1 + \|z\|)^2(1 + \|\xi\|_\infty)^2]} \sqrt{\mathbb{E}\left[\|\left[\nabla^2 \log p(x)\nabla_\phi x - diag\{\xi/\sigma^2\}\nabla_\phi \sigma\right]^T (s_p(x') + \xi'/\sigma')\|^2\right]} \\
&\lesssim B^2 G \sqrt{d_z \log d} \cdot LG \sqrt{\mathbb{E}[(1 + \|z\|)^2(1 + \|\xi\|_\infty)^2 \|s_p(x') + \xi'/\sigma'\|^2]} \\
&\lesssim LG^2 B^2 d_z \log d \sqrt{\mathbb{E}\|s_p(x)\|^2 + Ld}.
\end{aligned}
\tag{35}
$$

For $\textcircled{3}$, we have

$$
\nabla_\phi^2 \left[(s_p(x) + \xi/\sigma)^T(s_p(x') + \xi'/\sigma')\right] = \nabla_\phi^2 \left[s_p(x)^T s_p(x') + \frac{\xi}{\sigma} \cdot s_p(x') + \frac{\xi'}{\sigma'} \cdot s_p(x) + \frac{\xi}{\sigma} \cdot \frac{\xi'}{\sigma'}\right]
\tag{36}
$$

Firstly,

$$
\begin{aligned}
\nabla_\phi^2[s_p(x)^T s_p(x')] = \nabla_\phi^2 x[\nabla^2 \log p(x) s_p(x')] + \nabla_\phi^T x[\nabla^3 \log p(x) s_p(x')]\nabla_\phi x \\
+ \nabla_\phi^2 x'[\nabla^2 \log p(x') s_p(x)] + \nabla_\phi^T x'[\nabla^3 \log p(x') s_p(x)]\nabla_\phi x' \\
+ \nabla_\phi^T x \nabla^2 \log p(x)\nabla^2 \log p(x')\nabla_\phi x' + \nabla_\phi^T x' \nabla^2 \log p(x')\nabla^2 \log p(x)\nabla_\phi x.
\end{aligned}
\tag{37}
$$

$$\mathbb{E}\left[k(x,x')\|\nabla_\phi^2[s_p(x)^T s_p(x')]\|\right] \lesssim B^2\mathbb{E}\left[\left[LG(1+\|z\|)(1+\|\xi\|_\infty) + MG^2(1+\|z\|)^2(1+\|\xi\|_\infty^2)\right]\|s_p(x')\|\right]$$
$$+ B^2 L^2 G^2 \mathbb{E}[(1+\|z\|)(1+\|\xi\|_\infty)(1+\|z'\|)(1+\|\xi'\|_\infty)] \tag{38}$$
$$\lesssim GB^2(L\sqrt{d_z\log d} + GMd_z\log d)\sqrt{\mathbb{E}\|s_p(x)\|^2} + B^2 L^2 G^2 d_z\log d.$$

Then, since

$$\nabla_\phi^2[\frac{\xi}{\sigma}\cdot s_p(x')] = \nabla_\phi^2[s_p(x')]\frac{\xi}{\sigma} + \nabla_\phi^2[\frac{\xi}{\sigma}]s_p(x') + \nabla_\phi[s_p(x')]\otimes\nabla_\phi[\frac{\xi}{\sigma}] + \nabla_\phi[\frac{\xi}{\sigma}]\otimes\nabla_\phi[s_p(x')], \tag{39}$$

it holds that

$$\|\nabla_\phi^2[\frac{\xi}{\sigma}\cdot s_p(x')]\| \lesssim \|\nabla_\phi^2[s_p(x')]\|\cdot\|\frac{\xi}{\sigma}\| + \|\nabla_\phi^2[\frac{\xi}{\sigma}]\|\cdot\|s_p(x')\| + \|\nabla_\phi[\frac{\xi}{\sigma}]\|\cdot\|\nabla_\phi[s_p(x')]\|$$
$$\lesssim \left[M\|\nabla_\phi x'\|^2 + L\|\nabla_\phi^2 x'\|\right]\cdot L^{1/2}\|\xi\| + (1\vee L)^{3/2}G^2(1+\|z\|)^2\|\xi\|_\infty\cdot\|s_p(x')\| \tag{40}$$
$$+ L\|\nabla_\phi x'\|\cdot LG(1+\|z\|)\|\xi\|_\infty,$$

and thus

$$\mathbb{E}\left[k(x,x')\|\nabla_\phi^2[\frac{\xi}{\sigma}\cdot s_p(x')]\|\right] \lesssim B^2 G(L\sqrt{d_z\log d} + GMd_z\log d)L^{1/2}d^{1/2}$$
$$+ B^2 G^2(1\vee L)^{3/2}d_z\log^{1/2}d\sqrt{\mathbb{E}\|s_p(x)\|^2} + B^2 G^2 L^2 d_z\log d. \tag{41}$$

Lastly, we have

$$\nabla_\phi^2[\frac{\xi}{\sigma}\cdot\frac{\xi'}{\sigma'}] = \nabla_\phi^2[\frac{\xi}{\sigma}]\frac{\xi'}{\sigma'} + \nabla_\phi^2[\frac{\xi'}{\sigma'}]\frac{\xi}{\sigma} + \nabla_\phi[\frac{\xi'}{\sigma'}]\otimes\nabla_\phi[\frac{\xi}{\sigma}] + \nabla_\phi[\frac{\xi}{\sigma}]\otimes\nabla_\phi[\frac{\xi'}{\sigma'}]. \tag{42}$$

And

$$\mathbb{E}\left[k(x,x')\|\nabla_\phi^2[\frac{\xi}{\sigma}\cdot\frac{\xi'}{\sigma'}]\|\right] \lesssim B^2\mathbb{E}\left[(1\vee L)^{3/2}G^2(1+\|z\|)^2\|\xi\|_\infty\cdot L^{1/2}\|\xi'\| + \|\xi\|_\infty LG(1+\|z\|)\cdot\|\xi'\|_\infty LG(1+\|z'\|)\right]$$
$$\lesssim B^2\left[(1\vee L)^{3/2}L^{1/2}G^2 d_z d^{1/2}\log^{1/2}d + L^2 G^2 d_z\log d\right]$$
$$\lesssim B^2(1\vee L)^{3/2}L^{1/2}G^2 d_z d^{1/2}\log^{1/2}d. \tag{43}$$

Therefore, we get

$$\mathbb{E}[\|③\|] \lesssim B^2 G^2 d_z\log d\left((1\vee L)^{3/2} + M\right)\sqrt{\mathbb{E}\|s_p(x)\|^2 + Ld} \tag{44}$$

Combining ①, ②, ③, we can conclude that

$$\|\nabla_\phi^2\mathcal{L}\| \lesssim B^2 G^2 d_z\log d\left[(1\vee L)^3 + Ld + M^2 + \mathbb{E}\|s_p(x)\|^2\right]. \tag{45}$$

$\square$

Theorem A.4. *Both gradient estimators $\hat{g}_{vanilla}$ and $\hat{g}_{u\text{-}stat}$ have bounded variance:*

$$\mathrm{Var}(\hat{g}) \lesssim \frac{B^4 G^2 d_z\log d[L^3 d + L^2 d^2 + \mathbb{E}\|s_p(x)\|^4]}{N}. \tag{46}$$

Proof. It is sufficient to look at the upper bound of variance with two samples. Note that

$$\nabla_\phi\left[k(x,x')(s_p(x)+\xi/\sigma)^T(s_p(x')+\xi'/\sigma')\right] = \underbrace{\nabla_\phi[k(x,x')](s_p(x)+\xi/\sigma)^T(s_p(x')+\xi'/\sigma')}_{①}$$
$$+ \underbrace{k(x,x')\nabla_\phi\left[(s_p(x)+\xi/\sigma)^T(s_p(x')+\xi'/\sigma')\right]}_{②}. \tag{47}$$

For ①, we again utilize the fact that

$$\|\nabla_\phi k(x,x')\| \lesssim B^2 G[(1+\|z\|)(1+\|\xi\|_\infty) + (1+\|z'\|)(1+\|\xi'\|_\infty)], \tag{48}$$

and thus

$$
\begin{aligned}
\mathbb{E}[\|\textcircled{1}\|^2] &\lesssim \mathbb{E}[\|\nabla_\phi k\|^2 \|s_p(x) + \xi/\sigma\|^2 \|s_p(x') + \xi'/\sigma'\|^2] \\
&\lesssim B^4 G^2 d_z \log d [\mathbb{E}\|s_p(x) + \xi/\sigma\|^4]^{1/2} \mathbb{E}\|s_p(x) + \xi/\sigma\|^2 \\
&\lesssim B^4 G^2 d_z \log d \left[\sqrt{\mathbb{E}\|s_p(x)\|^4} + Ld\right] \mathbb{E}\|s_p(x) + \xi/\sigma\|^2.
\end{aligned}
\tag{49}
$$

For ②, we have

$$
\begin{aligned}
\nabla_\phi \left[(s_p(x) + \xi/\sigma)^T (s_p(x') + \xi'/\sigma')\right] &= \left[\nabla^2 \log p(x)\nabla_\phi x - diag\{\xi/\sigma^2\}\nabla_\phi \sigma\right]^T (s_p(x') + \xi'/\sigma') \\
&+ \left[\nabla^2 \log p(x')\nabla_\phi x' - diag\{\xi'/\sigma'^2\}\nabla_\phi \sigma'\right]^T (s_p(x) + \xi/\sigma).
\end{aligned}
\tag{50}
$$

Hence

$$
\begin{aligned}
\mathbb{E}[\|\textcircled{2}\|^2] &\lesssim B^4 \mathbb{E}\left[\left\|\left[\nabla^2 \log p(x)\nabla_\phi x - diag\{\xi/\sigma^2\}\nabla_\phi \sigma\right]^T (s_p(x') + \xi'/\sigma')\right\|^2\right] \\
&\lesssim B^4 L^2 G^2 \mathbb{E}[(1 + \|z\|)^2(1 + \|\xi\|_\infty)^2 \|s_p(x') + \xi'/\sigma'\|^2] \\
&\lesssim B^4 L^2 G^2 d_z \log d \left[\mathbb{E}\|s_p(x) + \xi/\sigma\|^2\right].
\end{aligned}
\tag{51}
$$

Therefore, combining ①, ②, we conclude that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\nabla_\phi \left[k(x,x')(s_p(x) + \xi/\sigma)^T (s_p(x') + \xi'/\sigma')\right]\right\|^2\right] \\
\lesssim B^4 G^2 d_z \log d \left[Ld + L^2 + \sqrt{\mathbb{E}\|s_p(x)\|^4}\right] \left[\mathbb{E}\|s_p(x) + \xi/\sigma\|^2\right],
\end{aligned}
\tag{52}
$$

which implies

$$
\begin{aligned}
\mathrm{Var}(\hat{g}) &\lesssim \frac{1}{N}\mathrm{Var}\left(\left[k(x,x')(s_p(x) + \xi/\sigma)^T (s_p(x') + \xi'/\sigma')\right]\right) \\
&\lesssim \frac{B^4 G^2 d_z \log d[L^3 d + L^2 d^2 + \mathbb{E}\|s_p(x)\|^4]}{N}.
\end{aligned}
\tag{53}
$$

$\square$

Theorem A.5. *Under Assumption 4.1-4.4, the iteration sequence generated by SGD $\phi_{t+1} = \phi_t - \eta \hat{g}_t$ with proper learning rate η includes an ε-stationary point $\hat{\phi}$ such that $\mathbb{E}[\|\nabla_\phi \mathcal{L}(\hat{\phi})\|] \leq \varepsilon$, if*

$$
T \gtrsim \frac{L_\phi \mathcal{L}_0}{\varepsilon^2}\left(1 + \frac{\Sigma_0}{N\varepsilon^2}\right).
\tag{54}
$$

Here $\mathcal{L}_0 := \mathcal{L}(\phi_0) - \inf_\phi \mathcal{L}$.

Proof. Since $\mathcal{L}(\cdot)$ is L_ϕ-smooth and $\hat{g}_t$ is unbiased with $\mathrm{Var}(\hat{g}_t) \leq \frac{\Sigma_0}{N}$, we have

$$
\begin{aligned}
\mathbb{E}_t \mathcal{L}(\phi_{t+1}) &\leq E_t \left[\mathcal{L}(\phi_t) + \langle \nabla_\phi \mathcal{L}(\phi_t), \phi_{t+1} - \phi_t \rangle + \frac{L_\phi}{2}\|\phi_{t+1} - \phi_t\|^2\right] \\
&\leq \mathcal{L}(\phi_t) - (\eta - \frac{\eta^2 L_\phi}{2})\|\nabla_\phi \mathcal{L}(\phi_t)\|^2 + \frac{\eta^2 L_\phi \Sigma_0}{2N}.
\end{aligned}
\tag{55}
$$

Let $\eta \leq 1/L_\phi$. Take full expectation and sum from 0 to $T-1$.

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\|\nabla_\phi \mathcal{L}(\phi_t)\|^2] &\lesssim \frac{\mathcal{L}(\phi_0) - \mathbb{E}\mathcal{L}(\phi_T)}{\eta T} + \frac{\eta L_\phi \Sigma_0}{N} \\
&\lesssim \sqrt{\frac{\Sigma_0 L_\phi \mathcal{L}_0}{NT}} + \frac{L_\phi \mathcal{L}_0}{T}.
\end{aligned}
\tag{56}
$$

In the last inequality, we plug optimal $\eta \asymp \min\left\{1/L_\phi, \sqrt{\frac{N\mathcal{L}_0}{L_\phi \sigma_0 T}}\right\}$. We conclude the proof by substituting the RHS with ε^2 and solving T. $\square$

A.3. Justification for Assumption 4.3

In all our experiments, the variance $\sigma(\cdot; \phi)$ is a constant vector $\sigma \in \mathbb{R}^d$, rather than a neural network. Therefore the smoothness assumption for $\sigma(\cdot; \phi)$ is valid for $G = 1$. As for $\mu(\cdot; \phi)$, we compute $\|\nabla_\phi \mu(z; \phi)\|$ and $\|\nabla_\phi^2 \mu(z; \phi)\|$ for randomly sampled z and ϕ in the training process. Note that for simplicity, we compute the Frobeneious norm and estimate the Lipschitz constant of Jacobian $\nabla_\phi \mu(z; \phi)$, which are larger than the operator norms. We plot the result in BLR experiment in Figure 6 and 7, showing that G remains within a certain range during the early, middle, and late stages of training. Therefore, Assumption 4.3 is reasonable.
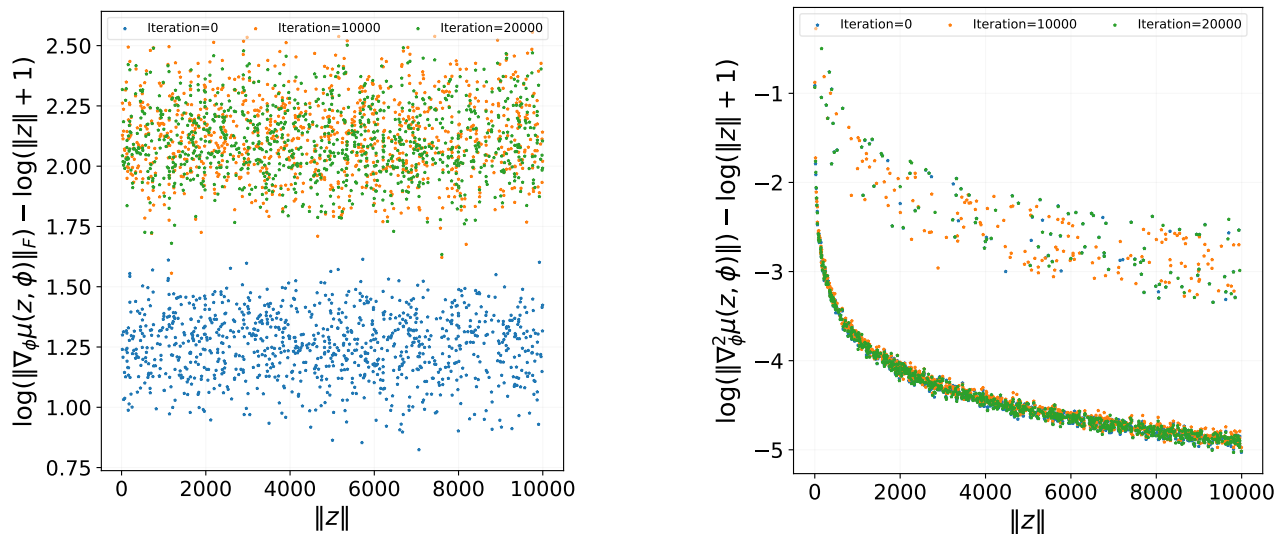

Figure 6. Smoothness of neural network $\mu(\cdot; \phi)$ in KSIVI ($\hat{g}_{\text{vanilla}}$) trajectory.
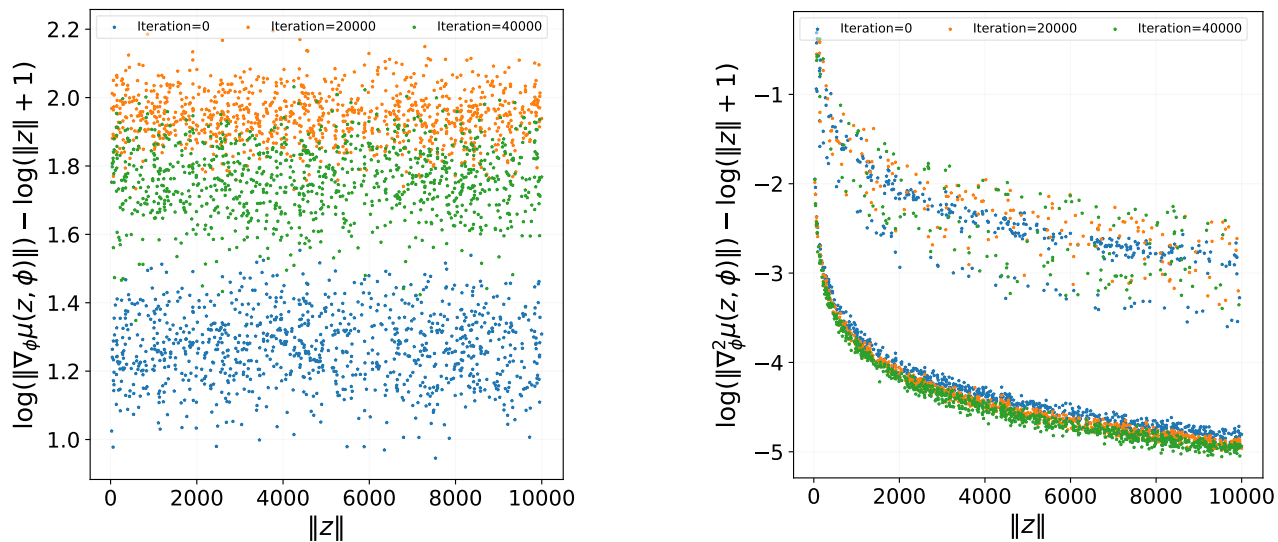

Figure 7. Smoothness of neural network $\mu(\cdot; \phi)$ in KSIVI ($\hat{g}_{\text{u-stat}}$) trajectory.
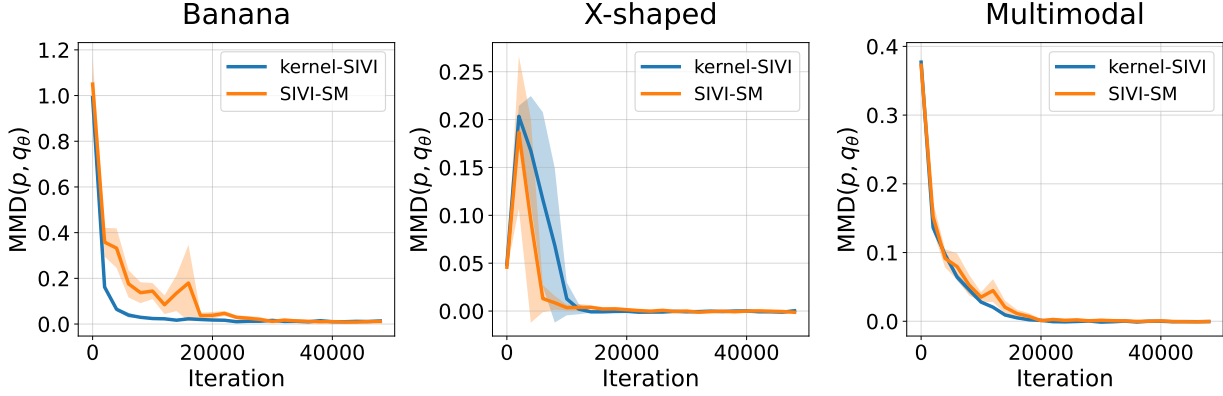
17

Figure 8. Convergence of MMD divergence of different methods on 2-D toy examples. The MMD objectives are estimated using 1000 samples. The results are averaged over 5 independent computations with the standard deviation as the shaded region.

B. KSIVI Algorithm

Algorithm 2 KSIVI with diagonal Gaussian conditional layer and U-statistic gradient estimator

---

Input: target score $s_p(x)$, number of iterations T, number of samples N for stochastic gradient.
Output: the optimal variational parameters ϕ^*.
for $t = 0, \cdots, T-1$ **do**
    Sample $\{z_1, \cdots, z_N\}$ from mixing distribution $q(z)$;
    Sample $\{\xi_1, \cdots, \xi_N\}$ from $\mathcal{N}(0, I)$;
    Compute $x_i = \mu(z_i; \phi) + \sigma(z_i; \phi) \odot \xi_i$ and $f_i = s_p(x_i) + \xi_i / \sigma(z_i; \phi)$;
    Compute the gradient estimate $\hat{g}_{\text{u-stat}}(\phi)$ through (16);
    Set $\phi \leftarrow \text{optimizer}(\phi, \hat{g}_{\text{u-stat}}(\phi))$;
end for
$\phi^* \leftarrow \phi$.

C. Experiment Setting and Additional Results

C.1. Toy Experiments

Setting details In the three 2-D toy examples, we set the conditional layer to be $\mathcal{N}(\mu(z; \phi), \text{diag}\{\phi_\sigma^2\})$. The $\mu(z; \phi)$ in SIVI-SM and KSIVI all have the same structures of multi-layer perceptrons (MLPs) with layer widths $[3, 50, 50, 2]$ and ReLU activation functions. The initial value of ϕ_σ is set to 1 except for banana distribution on which we use $\frac{1}{2}$. We set the learning rate of variational parameters ϕ to 0.001 and the learning rate of ψ in SIVI-SM to 0.002. In the lower-level optimization of SIVI-SM, ψ is updated after each update of ϕ.

Additional results Figure 8 depicts the descent of maximum mean discrepancy (MMD) during the training process of SIVI-SM and KSIVI.

Table 4. Three 2-D target distributions implemented in the 2-D toy experiments.

Name	Density	
BANANA	$x = \left(v_1, v_1^2 + v_2 + 1\right)^T, v \sim \mathcal{N}(0, \Sigma)$	$\Sigma = \left[\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 1 \end{smallmatrix}\right]$
MULTIMODAL	$x \sim \frac{1}{2}\mathcal{N}(x|\mu_1, I) + \frac{1}{2}\mathcal{N}(x|\mu_2, I)$	$\mu_1 = [-2, 0]^T, \mu_2 = [2, 0]^T$
X-SHAPED	$x \sim \frac{1}{2}\mathcal{N}(x|0, \Sigma_1) + \frac{1}{2}\mathcal{N}(x|0, \Sigma_2)$	$\Sigma_1 = \left[\begin{smallmatrix} 2 & 1.8 \\ 1.8 & 2 \end{smallmatrix}\right], \Sigma_2 = \left[\begin{smallmatrix} 2 & -1.8 \\ -1.8 & 2 \end{smallmatrix}\right]$

Table 5. **Estimated sliced Wasserstein distances (Sliced-WD) of KSIVI and SIVI-SM on Bayesian logistic regression.** The Sliced-WD objectives are estimated using 1000 samples.

Methods \Step sizes	0.00001	0.0001	0.001	0.005	0.008
SIVI-SM (iteration = 10k)	2.5415	24.2952	0.1203	0.6107	0.9906
SIVI-SM (iteration = 20k)	5.6184	28.6335	**0.0938**	0.3283	0.9145
KSIVI (iteration = 10k)	1.5196	0.6863	0.1509	**0.1064**	0.6059
KSIVI (iteration = 20k)	**1.2600**	**0.1120**	0.0965	0.2628	**0.3186**

C.2. Bayesian Logistic Regression

Setting details For the experiment on Bayesian logistic regression, $\mu(z; \phi)$ for all the SIVI variants are parameterized as MLPs with layer widths $[10, 100, 100, 22]$ and ReLU activation functions. For SIVI-SM, the f_ψ is three-layer MLPs with 256 hidden units and ReLU activation functions. The initial value of ϕ_σ^2 is e^{-5} for KSIVI and 1 for SIVI and SIVI-SM. These initial values are determined through grid search on the set $\{e^0, e^{-2}, e^{-5}\}$. For SIVI, the learning rate of ϕ_μ is set to 0.01, and the learning rate of ϕ_σ is set to 0.001, following the choices made in Yin & Zhou (2018). The learning rate for the variational parameters ϕ, comprising ϕ_μ and ϕ_σ, is set to 0.001 for both SIVI and SIVI-SM. The learning rate of ψ for SIVI-SM is set to 0.002, and we update ψ after each update of ϕ. For SIVI, the auxiliary sample size is set to $K = 100$. For all methods, we train the variational distributions for 20,000 iterations with a batch size of 100.

Additional results In line with Figure 5 in Domke & Sheldon (2018), we provide the results of KSIVI using varying step sizes of 0.00001, 0.0001, 0.001, 0.005, and 0.008. The results in Table 5 indicate that KSIVI demonstrates more consistent convergence behavior across different step sizes. Figure 9 shows the marginal and pairwise joint density of the well-trained posterior $q_\phi(\beta_1, \beta_2, \cdots, \beta_6)$ for SIVI, SIVI-SM, KSIVI with vanilla gradient estimator and KSIVI with U-statistic gradient estimator. Figure 10 presents the results of the estimated pairwise correlation coefficients by KSIVI using the U-statistic gradient estimator.

C.3. Conditioned Diffusion Process

Setting details For the experiment on conditioned diffusion process, $\mu(z; \phi)$ for all the SIVI variants are parameterized as MLPs with layer widths $[100, 128, 128, 100]$ and ReLU activation functions. For SIVI-SM, the f_ψ is a MLPs with layer widths $[100, 512, 512, 100]$ and ReLU activation functions. The initial value of ϕ_σ^2 is e^{-2} for all the SIVI variants. We set the learning rate of variational parameters ϕ to 0.0002 for KSIVI and 0.0001 for SIVI and SIVI-SM. The learning rate of ψ of SIVI-SM is set to 0.0001 and the number of lower-level gradient steps for ψ is 1. For all methods, we conduct training on variational distributions using 100,000 iterations with a batch size of 128, and the Adam optimizer is employed. For SIVI, we use the score-based training as done in Yu et al. (2023) for a fair time comparison.

Additional results Table 6 shows the estimated sliced Wasserstein distances from the target distributions to different variational approximations with increasing dimensions d = 50, 100, and 200 on conditional diffusion. Table 7 shows the training time per 10,000 iterations of KSIVI with the vanilla estimator and KSIVI with the u-stat estimator. We see that there exists a tradeoff between computation cost and estimation variance between the vanilla and u-stat overall. As shown in the equations (15) and (16), the vanilla estimator uses two batches of N samples and requires N^2 computation complexity for backpropagation, while the u-stat estimator uses one batch of N samples and requires $\frac{N^2}{2}$ computation. For convergence, the variance of the vanilla estimator (15) is slightly smaller than the variance of the u-stat estimator (16).

Figure 11 depicts the well-trained variational posteriors of SIVI, SIVI-SM, and KSIVI with the U-statistic gradient estimator on the conditioned diffusion process problem. Figure 12 shows the training loss of SIVI, SIVI-SM, and KSIVI. We observe that the training losses of all the methods converge after 100,000 iterations.

C.4. Bayesian Neural Network

Setting details For the experiments on the Bayesian neural network, we follow the setting in Liu & Wang (2016). For the hyper-parameters of the inverse covariance of the BNN, we select them as the estimated mean of 20 particles generated by SVGD (Liu & Wang, 2016). $\mu(z; \phi)$ for all the SIVI variants are parameterized as MLPs with layer widths $[3, 10, 10, d]$ and ReLU activation functions, where d is the number of parameters in Bayesian neural network. For SIVI-SM, the f_ψ takes

Table 6. **Estimated sliced Wasserstein distances (Sliced-WD) on conditional diffusion.** The Sliced-WD objectives are estimated using 1000 samples.

Methods \Dimensionality	50	100	200
SIVI	0.1266	0.0981	0.0756
SIVI-SM	0.0917	0.0640	0.0628
UIVI	0.0314	0.0426	0.0582
KSIVI (u-stat)	**0.0134**	0.0182	**0.0551**
KSIVI (vanilla)	**0.0140**	**0.0115**	0.0493

Table 7. **Training time (per 10,000 iterations) for the conditioned diffusion process inference task.** For all the methods, the batch size for Monte Carlo estimation is set to $N = 128$.

Methods \Dimensionality	50	100	200
KSIVI (u-stat)	39.09	58.13	64.17
KSIVI (vanilla)	56.67	90.48	107.84

three-layer MLPs with 16 hidden units and ReLU activation functions. For KSIVI, we apply a practical regularization fix to thin the kernel Stein discrepancy, as described in (Bénard et al., 2023). For all the SIVI variants, we choose the learning rate of ϕ from $\{0.00001, 0.0001, 0.0002, 0.0005, 0.001\}$ and run 20,000 iterations for training. For SGLD, we choose the step size from the set $\{0.00002, 0.00004, 0.00005, 0.0001\}$, and the number of iterations is set to 10,000 with 100 particles.

Additional results Table 8 shows the additional results of KSIVI with the U-statistic gradient estimator.

Table 8. **Test RMSE and test NLL of Bayesian neural networks on several UCI datasets.** The results are averaged from 10 independent runs with the standard deviation in the subscripts. For each data set, the best result is marked in **black bold font** and the second best result is marked in **brown bold font**.

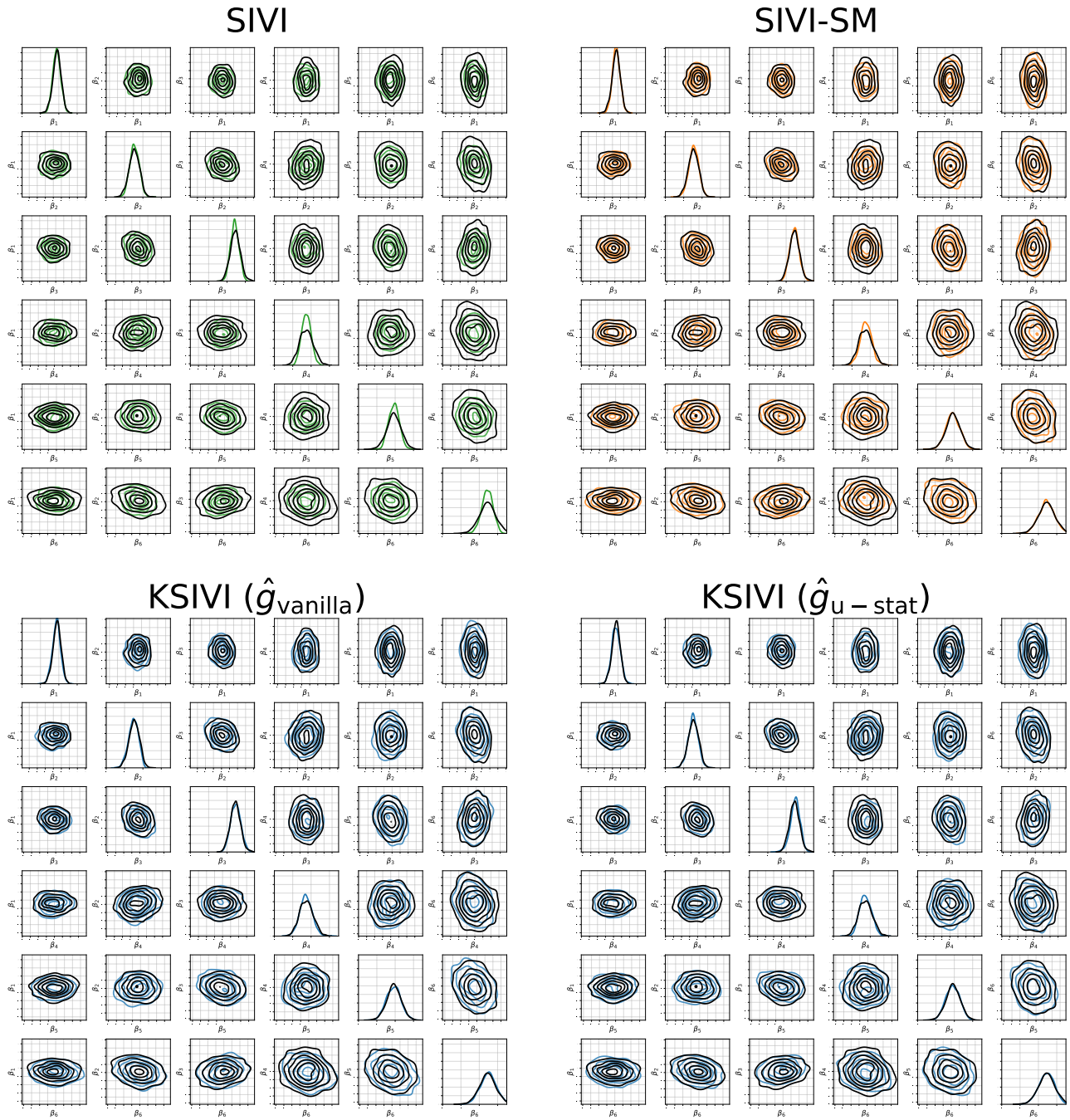| Dataset	Test RMSE ($\downarrow$)				Test NLL ($\downarrow$)			
	SIVI	SIVI-SM	KSIVI ($\hat{g}_{\text{u-stat}}$)	KSIVI ($\hat{g}_{\text{vanilla}}$)	SIVI	SIVI-SM	KSIVI ($\hat{g}_{\text{u-stat}}$)	KSIVI ($\hat{g}_{\text{vanilla}}$)
BOSTON	$2.621_{\pm 0.02}$	$2.785_{\pm 0.03}$	$2.857_{\pm 0.11}$	$\mathbf{2.555}_{\pm 0.02}$	$\mathbf{2.481}_{\pm 0.00}$	$2.542_{\pm 0.01}$	$3.094_{\pm 0.01}$	$2.506_{\pm 0.01}$
CONCRETE	$6.932_{\pm 0.02}$	$5.973_{\pm 0.04}$	$6.861_{\pm 0.19}$	$\mathbf{5.750}_{\pm 0.03}$	$3.337_{\pm 0.00}$	$\mathbf{3.229}_{\pm 0.01}$	$4.036_{\pm 0.01}$	$3.309_{\pm 0.01}$
POWER	$\mathbf{3.861}_{\pm 0.01}$	$4.009_{\pm 0.00}$	$3.916_{\pm 0.01}$	$3.868_{\pm 0.01}$	$2.791_{\pm 0.00}$	$2.822_{\pm 0.00}$	$2.944_{\pm 0.00}$	$2.797_{\pm 0.00}$
WINE	$0.597_{\pm 0.00}$	$0.605_{\pm 0.00}$	$0.597_{\pm 0.00}$	$\mathbf{0.595}_{\pm 0.00}$	$0.904_{\pm 0.00}$	$0.916_{\pm 0.00}$	$0.904_{\pm 0.00}$	$\mathbf{0.901}_{\pm 0.00}$
YACHT	$1.505_{\pm 0.07}$	$\mathbf{0.884}_{\pm 0.01}$	$2.152_{\pm 0.09}$	$1.237_{\pm 0.05}$	$1.721_{\pm 0.03}$	$\mathbf{1.432}_{\pm 0.01}$	$2.873_{\pm 0.03}$	$1.752_{\pm 0.03}$
PROTEIN	$\mathbf{4.669}_{\pm 0.00}$	$5.087_{\pm 0.00}$	$4.777_{\pm 0.00}$	$5.027_{\pm 0.01}$	$\mathbf{2.967}_{\pm 0.00}$	$3.047_{\pm 0.00}$	$2.984_{\pm 0.00}$	$3.034_{\pm 0.00}$

Figure 9. **Marginal and pairwise joint posteriors on $\beta_1, \cdots, \beta_6$.** The contours of the pairwise empirical densities produced by three baseline algorithms, i.e. SIVI-SM (in orange), SIVI (in green), and KSIVI (in blue) including both KSIVI ($\hat{g}_{\text{vanilla}}$) and KSIVI ($\hat{g}_{\text{u-stat}}$) are graphed in comparison to the ground truth(in black).
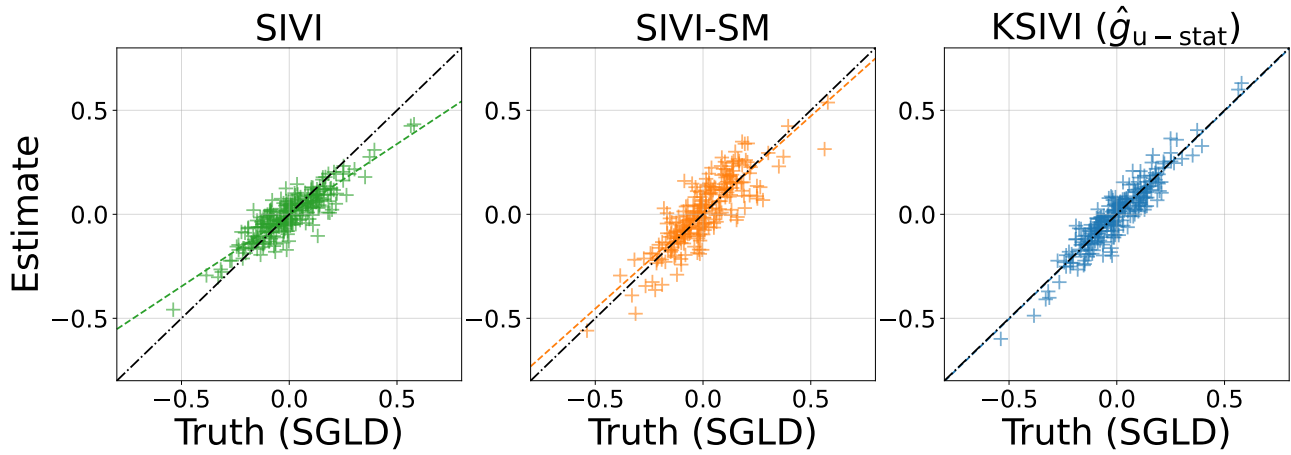
Figure 10. **Comparison result between the estimated pairwise correlation coefficients on SIVI, SIVI-SM, and KSIVI with U-statistic gradient estimator.** The correlation coefficients are estimated with 1000 samples.
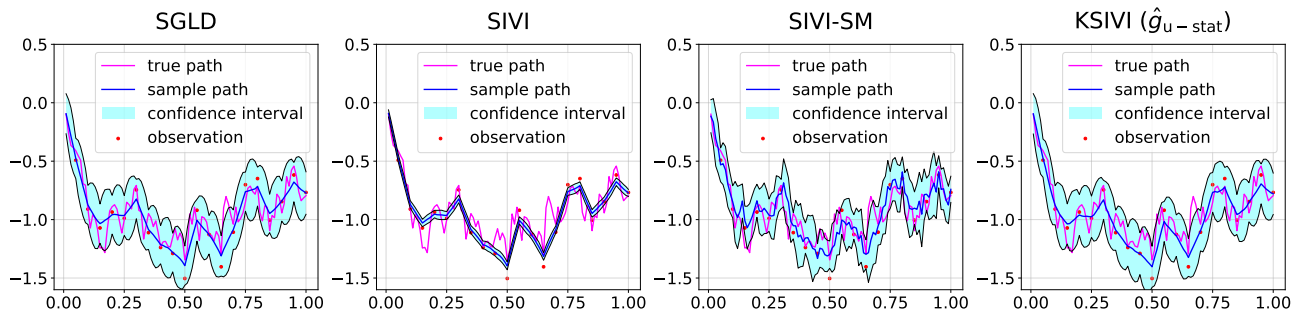

Figure 11. **Variational approximations of different methods for the discretized conditioned diffusion process.** The magenta trajectory represents the ground truth via parallel SGLD. The blue line corresponds to the estimated posterior mean of different methods, and the shaded region denotes the 95% marginal posterior confidence interval at each time step. The sample size is 1000.

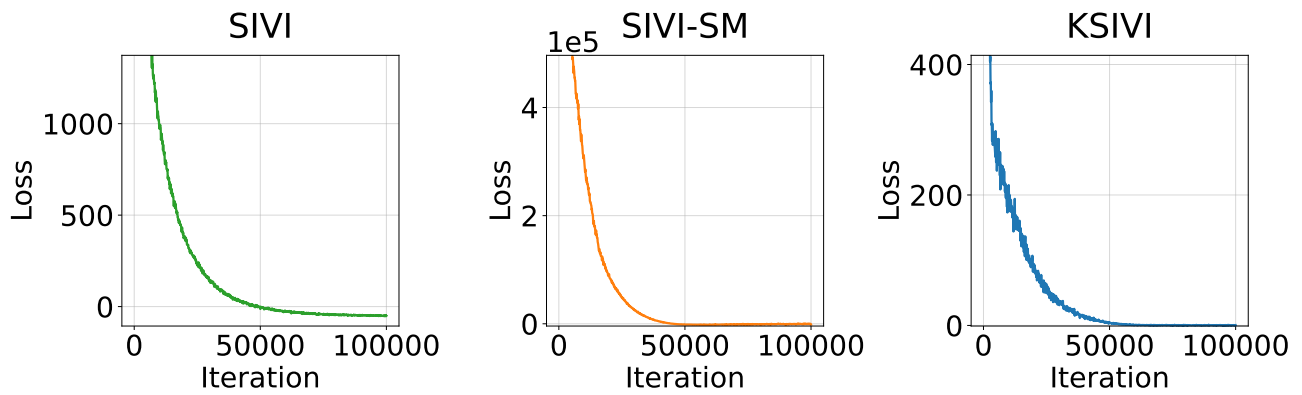*Figure 12.* **Training loss of SIVI in conditioned diffusion process experiment.** The sample size is 1000 for all the methods.