# Stein Variational Gradient Descent as Moment Matching

**Qiang Liu, Dilin Wang**
Department of Computer Science
The University of Texas at Austin
Austin, TX 78712
{lqiang, dilin}@cs.utexas.edu

## Abstract

Stein variational gradient descent (SVGD) is a non-parametric inference algorithm that evolves a set of particles to fit a given distribution of interest. We analyze the non-asymptotic properties of SVGD, showing that there exists a set of functions, which we call the *Stein matching set*, whose expectations are *exactly* estimated by any set of particles that satisfies the fixed point equation of SVGD. This set is the image of Stein operator applied on the feature maps of the positive definite kernel used in SVGD. Our results provide a theoretical framework for analyzing properties of SVGD with different kernels, shedding insight into optimal kernel choice. In particular, we show that SVGD with linear kernels yields exact estimation of means and variances on Gaussian distributions, while random Fourier features enable probabilistic bounds for distributional approximation. Our results offer a refreshing view of the classical inference problem as fitting Stein's identity or solving the Stein equation, which may motivate more efficient algorithms.

## 1 Introduction

One of the core problems of modern statistics and machine learning is to approximate difficult-to-compute probability distributions. Two fundamental ideas have been extensively studied and used in the literature: variational inference (VI) and Markov chain Monte Carlo (MCMC) sampling (e.g., Koller & Friedman, 2009; Wainwright et al., 2008). MCMC has the advantage of being non-parametric and asymptotically exact, but often suffers from difficulty in convergence, while VI frames the inference into a parametric optimization of the KL divergence and works much faster in practice, but loses the asymptotic consistency. An ongoing theme of research is to combine the advantages of these two methodologies.

Stein variational gradient descent (SVGD) (Liu & Wang, 2016) is a synthesis of MCMC and VI that inherits the non-parametric nature of MCMC while maintaining the optimization perspective of VI. In brief, SVGD for distribution $p(x)$ updates a set of particles $\{\boldsymbol{x}_i\}_{i=1}^n$ parallelly with a velocity field $\boldsymbol{\phi}(\cdot)$ that balances the gradient force and repulsive force,

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \epsilon \boldsymbol{\phi}(\boldsymbol{x}_i), \qquad \boldsymbol{\phi}(\cdot) = \frac{1}{n}\sum_{j=1}^{n} \nabla_{\boldsymbol{x}_j} \log p(\boldsymbol{x}_j) k(\boldsymbol{x}_j, \cdot) + \nabla_{\boldsymbol{x}_j} k(\boldsymbol{x}_j, \cdot),$$

where $\epsilon$ is a step size and $k(\boldsymbol{x}, \boldsymbol{x}')$ is a positive definite kernel defined by the user. This update is derived as approximating a kernelized Wasserstein gradient flow of KL divergence (Liu et al., 2017) with connection to Stein's method (Stein, 1972) and optimal transport (Ollivier et al., 2014); see also Anderes & Coram (2002). SVGD has been applied to solve challenging inference problems in various domains; examples include Bayesian inference (Liu & Wang, 2016; Feng et al., 2017),

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

uncertainty quantification (Zhu & Zabaras, 2018), reinforcement learning (Liu et al., 2017; Haarnoja et al., 2017), learning deep probabilistic models (Wang & Liu, 2016; Pu et al., 2017) and Bayesian meta learning (Feng et al., 2017; Kim et al., 2018).

However, the theoretical properties of SVGD are still largely unexplored. The only exceptions are Liu et al. (2017); Lu et al. (2018), which studied the partial differential equation that governs the evolution of the limit densities of the particles, with which the convergence to the distribution of interest can be established. However, the results in Liu et al. (2017); Lu et al. (2018) are asymptotic in nature and hold only when the number of particles is very large. A theoretical understanding of SVGD in the finite sample size region is still missing and of great practical importance, because the particle sizes used in practice are often relatively small, given that SVGD with a single particle exactly reduces to finding the mode (a.k.a. maximum a posteriori (MAP)).

**Our Results** We analyze the finite sample properties of SVGD. In contrast to the dynamical perspective of Liu et al. (2017), we directly study what properties a set of particles would have if it satisfies the fixed point equation of SVGD, regardless of how we obtain them algorithmically, or whether the fixed point is unique. Our analysis indicates that the fixed point equation of SVGD is essentially a moment matching condition which ensures that the fixed point particles $\{\boldsymbol{x}_i^*\}_{i=1}^n$ *exactly* estimate the expectations of all the functions in a special function set $\mathcal{F}^*$,

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i^*) = \mathbb{E}_p f, \quad \forall f \in \mathcal{F}^*.$$

This set $\mathcal{F}^*$, which we call the *Stein matching set*, consists of functions obtained by applying Stein operator on the linear span of feature maps of the kernel used by SVGD.

This framework allows us to understand properties of different kernels (and the related feature maps) by studying their Stein matching sets $\mathcal{F}^*$, which should ideally either match the test functions that we are actually interested in estimating, or is as large as possible to approximate the overall distribution. This process is difficult in general, but we make two observations in this work:

*i*) We show that, by using linear kernels (features), SVGD can *exactly* estimate the mean and variance of Gaussian distributions when the number of particles is larger than the dimension. Since Gaussian-like distributions appear widely in practice, and the estimates of mean and variance are often of special importance, linear kernels can provide a significant advantage over the typical Gaussian RBF kernels, especially in estimating the variance.

*ii*) Linear features are not sufficient to approximate the whole distributions. We show that, by using random features of strictly positive definite kernels, the fixed points of SVGD approximate the whole distribution with an $O(1/\sqrt{n})$ rate in kernelized Stein discrepancy.

Overall, our framework reveals a novel perspective that reduces the inference problem to either a *regression problem* of fitting Stein identities, or *inverting the Stein operator* which is framed as solving a differential equation called Stein equation. These ideas are significantly different from the traditional MCMC and VI that are currently popular in machine learning literature, and draw novel connections to Quasi Monte Carlo and quadrature methods, among other techniques in applied mathematics. New efficient approximate inference methods may be motivated with our new perspectives.

## 2 Background

We introduce the basic background of the Stein variational method, a framework of approximate inference that integrates ideas from Stein's method, kernel methods, and variational inference. The readers are referred to Liu et al. (2016); Liu & Wang (2016); Liu et al. (2017) and references therein for more details. For notation, all vectors are assumed to be column vectors. The differential operator $\nabla_{\boldsymbol{x}}$ is viewed as a column vector of the same size as $\boldsymbol{x} \in \mathbb{R}^d$. For example, $\nabla_{\boldsymbol{x}}\phi$ is a $\mathbb{R}^d$-valued function when $\phi$ is a scalar-valued function, and $\nabla_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^d \partial_{x_i} \boldsymbol{\phi}(\boldsymbol{x})$ is a scalar-valued function when $\boldsymbol{\phi}$ is $\mathbb{R}^d$-valued.

**Stein's Identity** Stein's identity forms the foundation of our framework. Given a positive differentiable density $p(\boldsymbol{x})$ on $\mathcal{X} \subseteq \mathbb{R}^d$, one form of Stein's identity is

$$\mathbb{E}_p[\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x})] = 0, \quad \forall \boldsymbol{\phi},$$

which holds for any differentiable, $\mathbb{R}^d$-valued function $\boldsymbol{\phi}$ that satisfies a proper zero-boundary condition. Stein's identity can be proved by a simple exercise of integration by parts. We may write Stein's identity in a more compact way by defining a Stein operator $\mathcal{P}_{\boldsymbol{x}}$:

$$\mathbb{E}_p[\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x})] = 0, \qquad \text{where} \qquad \mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x}),$$

where $\mathcal{P}_{\boldsymbol{x}}$ is formally viewed as a $d$-dimensinoal column vector like $\nabla_{\boldsymbol{x}}$, and hence $\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}$ is the inner product of $\mathcal{P}_{\boldsymbol{x}}$ and $\boldsymbol{\phi}$, yielding a scalar-valued function.

The power of Stein's identity is that, for a given distribution $p$, it defines an *infinite* number of functions of form $\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}$ that has zero expectation under $p$, all of which only depend on $p$ through the Stein operator $\mathcal{P}_{\boldsymbol{x}}$, or the score function $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) = \frac{\nabla p(\boldsymbol{x})}{p(\boldsymbol{x})}$, which is independent of the normalization constant in $p$ that is often difficult to calculate.

**Stein Discrepancy on RKHS** Stein's identity can be leveraged to characterize the discrepancy between different distributions. The idea is that, for two different distributions $p \neq q$, there shall exist a function $\boldsymbol{\phi}$ such that $\mathbb{E}_q[\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}] \neq 0$. Consider functions $\boldsymbol{\phi}$ in a $\mathbb{R}^d$-valued reproducing kernel Hilbert space (RKHS) of form $\mathcal{H} = \mathcal{H}_0 \times \cdots \mathcal{H}_0$ where $\mathcal{H}_0$ is a $\mathbb{R}$-valued RKHS with positive definite kernel $k(\boldsymbol{x}, \boldsymbol{x}')$. We may define a *kernelized Stein discrepancy* (KSD) (Liu et al., 2016; Chwialkowski et al., 2016; Oates et al., 2017):

$$\mathbb{D}_k(q \,||\, p) = \max_{\boldsymbol{\phi} \in \mathcal{H}} \big\{ \mathbb{E}_q[\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x})] \;:\; ||\boldsymbol{\phi}||_{\mathcal{H}} \leq 1 \big\}, \tag{1}$$

The optimal $\boldsymbol{\phi}$ in (1) can be solved in closed form:

$$\boldsymbol{\phi}^*_{q,p}(\cdot) \propto \mathbb{E}_{\boldsymbol{x} \sim q}[\mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)], \tag{2}$$

which yields a simple kernel-based representation of KSD:

$$\mathbb{D}_k^2(q \,||\, p) = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{x}' \sim q}[\kappa_p(\boldsymbol{x}, \boldsymbol{x}')], \qquad \text{with} \qquad \kappa_p(\boldsymbol{x}, \boldsymbol{x}') = \mathcal{P}_{\boldsymbol{x}}^\top (\mathcal{P}_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}')), \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are i.i.d. draws from $q$, and $\kappa_p(\boldsymbol{x}, \boldsymbol{x}')$ is a new "Steinalized" positive definite kernel obtained by applying the Stein operator twice, first w.r.t. variable $\boldsymbol{x}$ and then $\boldsymbol{x}'$. It turns out that the RKHS related to kernel $\kappa_p(\boldsymbol{x}, \boldsymbol{x}')$ is exactly the space of functions obtained by applying Stein operator on functions in $\mathcal{H}$, that is,

$$\mathcal{H}_p = \{\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi} \colon \forall \boldsymbol{\phi} \in \mathcal{H}\}.$$

By Stein's identity, all the functions in $\mathcal{H}_p$ have zero expectation under $p$. We can also define $\mathcal{H}_p^+$ to be the space of functions in $\mathcal{H}_p$ adding arbitrary constants, that is, $\mathcal{H}_p^+ := \{f(\boldsymbol{x}) + c \colon f \in \mathcal{H}_p, \;\; c \in \mathbb{R}\}$, which can also be viewed as a RKHS, with kernel $\kappa_p(\boldsymbol{x}, \boldsymbol{x}') + 1$. Stein discrepancy can be viewed as a maximum mean discrepancy (MMD) on the Steinalized RKHS $\mathcal{H}_p^+$ (or equivalently $\mathcal{H}_p$):

$$\mathbb{D}_k(q \,||\, p) = \max_{f \in \mathcal{H}_p^+} \big\{ \mathbb{E}_q f - \mathbb{E}_p f \colon \quad ||f||_{\mathcal{H}_p^+} \leq 1 \big\}. \tag{4}$$

Different from typical MMD, here the RKHS space depends on distribution $p$. In order to make Stein discrepancy discriminative, in that $\mathbb{D}_k(q \,||\, p) = 0$ implies $q = p$, we need to take kernels $k(\boldsymbol{x}, \boldsymbol{x}')$ so that $\mathcal{H}_p^+$ is sufficiently large. It has been shown that this can be achieved if $k(\boldsymbol{x}, \boldsymbol{x}')$ is strictly positive definite or universal, in a proper technical sense (Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017; Oates et al., 2017).

It is useful to consider the kernels in a random feature representation (Rahimi & Recht, 2007),

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{w} \sim p_{\boldsymbol{w}}}[\phi(\boldsymbol{x}, \boldsymbol{w}) \phi(\boldsymbol{x}', \boldsymbol{w})], \tag{5}$$

where $\phi(\boldsymbol{x}, \boldsymbol{w})$ is a set of features indexed by a random parameter $\boldsymbol{w}$ drawn from a distribution $p_{\boldsymbol{w}}$. For example, the Gaussian RBF kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{1}{2h^2}||\boldsymbol{x} - \boldsymbol{x}'||_2^2)$ admits

$$\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{w}) = \sqrt{2} \cos(\frac{1}{h} \boldsymbol{w}_1^\top \boldsymbol{x} + w_0), \tag{6}$$

where $w_0 \sim \mathrm{Unif}([0, 2\pi])$ and $\boldsymbol{w}_1 \sim \mathcal{N}(0, I)$. With the random feature representation, KSD can be rewritten into

$$\mathbb{D}_k^2(q \,||\, p) = \mathbb{E}_{\boldsymbol{w} \sim p_{\boldsymbol{w}}} \Big[ ||\mathbb{E}_{\boldsymbol{x} \sim q}[\mathcal{P}_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{w})]||^2 \Big], \tag{7}$$

which can be viewed as the mean square error of Stein's identity $\mathbb{E}_{\boldsymbol{x} \sim q}[\mathcal{P}_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{w})] = 0$ over the random features. $\mathbb{D}_k^2(q \,||\, p) = 0$ shall imply $q = p$ if the feature set $\mathcal{G} = \{\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{w}) \colon \forall \boldsymbol{w}\}$ is rich enough. Note that the RKHS $\mathcal{H}$ and feature set $\mathcal{G}$ are different; Stein discrepancy is an *expected* loss function on $\mathcal{G}$ as shown in (7), but a *worst-case* loss on $\mathcal{H}$ as shown in (1).

**Stein Variational Gradient Descent (SVGD)** SVGD is a deterministic sampling algorithm motivated by Stein discrepancy. It is based on the following basic observation: given a distribution $q$, assume $q_{[\boldsymbol{\phi}]}$ is the distribution of $\boldsymbol{x}' = \boldsymbol{x} + \epsilon\boldsymbol{\phi}(\boldsymbol{x})$ obtained by updating $\boldsymbol{x}$ with a velocity field $\boldsymbol{\phi}$, where $\epsilon$ is a small step size, then we have

$$\mathrm{KL}(q_{[\boldsymbol{\phi}]} \,||\, p) = \mathrm{KL}(q \,||\, p) - \epsilon\mathbb{E}_q[\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}] + O(\epsilon^2),$$

which shows that the decrease of KL divergence is dominated by $\epsilon\mathbb{E}_q[\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}]$. In order to choose $\boldsymbol{\phi}$ to make $q_{[\boldsymbol{\phi}]}$ move towards $p$ as fast as possible, we should choose $\boldsymbol{\phi}$ to maximize $\mathbb{E}_q[\mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}]$, whose solution is exactly $\boldsymbol{\phi}^*_{q,p}(\cdot) \propto \mathbb{E}_{\boldsymbol{x}\sim q}[\mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \cdot)]$ as shown in (2). This suggests that $\boldsymbol{\phi}^*_{q,p}$ happens to be the best velocity field that pushes the probability mass of $q$ towards $p$ as fast as possible.

Motivated by this, SVGD approximates $q$ with the empirical distribution of a set of particles $\{\boldsymbol{x}_i\}_{i=1}^n$, and iteratively updates the particles by

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \frac{\epsilon}{n}\sum_{j=1}^n [\mathcal{P}_{\boldsymbol{x}_j} k(\boldsymbol{x}_j, \boldsymbol{x}_i)]. \tag{8}$$

Liu et al. (2017) studied the asymptotic properties of the dynamic system underlying SVGD, showing that the evolution of the limit density of the particles when $n \to \infty$ can be captured by a nonlinear Fokker-Planck equation, and established its weak convergence to the target distribution $p$.

However, the analysis in Liu et al. (2017) and Lu et al. (2018) do not cover the case when the sample size $n$ is finite, which is more relevant to the practical performance. We address this problem by directly analyzing the properties of the fixed point equation of SVGD, yielding results that work for finite sample size $n$, also independent of the update rule used to arrive the fixed points.

## 3 SVGD as Moment Matching

This section presents our main results on the moment matching properties of SVGD and the related Stein matching sets. We start with Section 3.1 which introduces the basic idea and characterizes the Stein matching set of SVGD with general positive definite kernels. We then analyze in Section 3.2 the special case when the rank of the kernel is less than the particle size, in which case the Stein matching set is independent of the fixed points themselves. Section 3.3 shows that SVGD with linear features exactly estimates the first two second-order moments of Gaussian distributions. Section 3.4 establishes a probabilistic bound when random features are used.

### 3.1 Fixed Point of SVGD

Our basic idea is rather simple to illustrate. Assume $X^* = \{\boldsymbol{x}_i^*\}_{i=1}^n$ is the fixed point of SVGD and $\hat{\mu}_{X^*}$ its related empirical measure, then according to (8), the fixed point condition of SVGD ensures

$$\mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[\mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}_i^*)] = 0, \qquad \forall i = 1, \dots, n. \tag{9}$$

On the other hand, by Stein's identity, we have

$$\mathbb{E}_{\boldsymbol{x}\sim p}[\mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}_i^*)] = 0, \qquad \forall i = 1, \dots, n.$$

This suggests that $\hat{\mu}_{X^*}$ *exactly* estimates the expectation of functions of form $f(\boldsymbol{x}) = \mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}_i^*)$ under $p$, all of which are zero. By the linearity of expectation, the same holds for all the functions in the linear span of $\mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}_i^*)$.

**Lemma 3.1.** *Assume $X^* = \{\boldsymbol{x}_i^*\}_{i=1}^n$ satisfies the fixed point equation* (9) *of SVGD. We have*

$$\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i^*) = \mathbb{E}_p f, \qquad \forall f \in \mathcal{F}^*,$$

*where the Stein matching set $\mathcal{F}^*$ is the linear span of $\{\mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}_i^*)\}_{i=1}^n \cup \{1\}$, that is, $\mathcal{F}^*$ consists of*

$$f(\boldsymbol{x}) = \sum_{i=1}^n \boldsymbol{a}_i^\top \mathcal{P}_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{x}_i^*) \;+\; b, \quad \forall \boldsymbol{a}_i \in \mathbb{R}^d,\; b \in \mathbb{R}.$$

*Equivalently, $f(\boldsymbol{x}) = \mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x}) + b$ and $\boldsymbol{\phi}$ is in the linear span of $\{k(\boldsymbol{x}, \boldsymbol{x}_i^*)\}_{i=1}^n$, that is, $\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^n \boldsymbol{a}_i k(\boldsymbol{x}, \boldsymbol{x}_i^*)$.*

Extending Lemma 3.1, one can readily see that the SVGD fixed points can approximate the expectation of functions that are close to $\mathcal{F}^*$. Specifically, let $\mathcal{F}^*_\epsilon$ be the $\epsilon$ neighborhood of $\mathcal{F}^*$, that is, $\mathcal{F}^*_\epsilon = \{f\colon \inf_{f'\in\mathcal{F}} ||f - f'||_\infty \le \epsilon\}$, then it is easily shown that

$$\left|\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i^*) - \mathbb{E}_p f\right| \le 2\epsilon, \qquad \forall f \in \mathcal{F}^*_\epsilon.$$

Therefore, the SVGD approximation can be viewed as *prioritizing* the functions within, or close to, $\mathcal{F}^*$. This is different in nature from Monte Carlo, which approximates the expectation of *all bounded variance functions* with the same $O(1/\sqrt{n})$ error rate. Instead, SVGD shares more similarity with the *quadrature* and *sigma point methods*, which also find points (particles) to match the expectation on certain class of functions, but mostly only on polynomial functions and for simple distributions such as uniform or Gaussian distributions. SVGD provides a more general approach that can match moments of richer classes of functions for more general complex multivariate distributions. As we show in Section 3.3, when using polynomial kernels, SVGD reduces to matching polynomials when applied to multivariate Gaussian distributions.

In this view, the performance of SVGD is essentially decided by the Stein matching set $\mathcal{F}^*$. We shall design the algorithm, by engineering the kernels or feature maps, to make $\mathcal{F}^*$ as large as possible in order to approximate the distribution well, or include the test functions of actual interest, such as mean and variance.

### 3.2 Fixed Point of Feature-based SVGD

One undesirable property of $\mathcal{F}^*$ in Lemma 3.1 is that it depends on the values of the fixed point particles $X^*$, whose properties are difficult to characterize *a priori*. This makes it difficult to infer what kernel should be used to obtain a desirable $\mathcal{F}^*$. It turns out the dependency of $\mathcal{F}$ on $X^*$ can be essentially decoupled by using *degenerated kernels* corresponding to a finite number of feature maps. Specifically, we consider kernels of form

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\ell=1}^m \phi_\ell(\boldsymbol{x})\phi_\ell(\boldsymbol{x}'),$$

where we assume the number $m$ of features is no larger than the particle size $n$. Then, the fixed point of SVGD reduces to

$$\mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[\sum_{\ell=1}^m \mathcal{P}_{\boldsymbol{x}}\phi_\ell(\boldsymbol{x})\phi_\ell(\boldsymbol{x}_j^*)] = 0, \quad \forall j \in [n]. \tag{10}$$

Define $\Phi = [\phi_\ell(\boldsymbol{x}_j^*)]_{\ell,j}$ which is a matrix of size $(m \times n)$ . If $rank(\Phi) \ge m$, then (10) reduces to

$$\mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[\mathcal{P}_{\boldsymbol{x}}\phi_\ell(\boldsymbol{x})] = 0, \quad \forall \ell = 1, \dots, m, \tag{11}$$

where the test function $f(\boldsymbol{x}) := \mathcal{P}_{\boldsymbol{x}}\phi_\ell(\boldsymbol{x})$ no longer depends on the fixed point $X^*$.

**Theorem 3.2.** *Assume $X^*$ is a fixed point of SVGD with kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\ell=1}^m \phi_\ell(\boldsymbol{x})\phi_\ell(\boldsymbol{x}')$. Define the $(m \times n)$ matrix $\Phi = [\phi_\ell(\boldsymbol{x}_i^*)]_{\ell\in[m],i\in[n]}$. If $\mathrm{rank}(\Phi) \ge m$, then*

$$\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i^*) = \mathbb{E}_p f, \quad \forall f \in \mathcal{F}^*,$$

*where the Stein matching set $\mathcal{F}^*$ is the linear span of $\{\mathcal{P}_{\boldsymbol{x}}\phi_\ell(\boldsymbol{x})\}_{\ell=1}^m \cup \{1\}$, that is, it is set of the functions of form*

$$f(\boldsymbol{x}) = \sum_{\ell=1}^m \boldsymbol{a}_\ell^\top \mathcal{P}_{\boldsymbol{x}}\phi_\ell(\boldsymbol{x}) \ + \ b, \qquad \forall\ \boldsymbol{a}_\ell \in \mathbb{R}^d,\ b \in \mathbb{R}. \tag{12}$$

Note that the rank condition implies that we must have $m \le n$. The idea is that $n$ particles can at most match $n$ linearly independent features exactly. Here, although the rank condition still depends on the fixed point $X^* = \{\boldsymbol{x}_i^*\}_{i=1}^n$ and cannot be guaranteed *a priori*, it can be numerically verified once we obtain the values of $X^*$. In our experiments, we find that the rank condition tends to always hold practically when $n = m$. In cases when it does fail to satisfy, we can always rerun the algorithm with a larger $n$ until it is satisfied. Intuitively, it seems to require bad luck to have $\Phi$ low rank when there are more particles than features ($n \ge m$), although a theoretical guarantee is still missing.

**Query-Specific Inference as Solving Stein Equation** Assume we are interested in a *query-specific* task of estimating $\mathbb{E}_p f$ for a specific test function $f$. In this case, we should ideally select the features $\{\phi_\ell\}_\ell$ such that (12) holds to yield an exact estimation of $\mathbb{E}_p f$. By the linearity of the Stein operator, (12) is equivalent to

$$\texttt{Stein Equation:} \qquad f(\boldsymbol{x}) = \mathcal{P}_{\boldsymbol{x}}^\top \boldsymbol{\phi}(\boldsymbol{x}) + b, \tag{13}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{\ell=1}^m \boldsymbol{a}_\ell \phi_\ell(\boldsymbol{x})$. Eq (13) is known as *Stein Equation* when solving $\boldsymbol{\phi}$ and $b$ with a given $f$, which effectively calculates the inverse of Stein operator.

Stein equation plays a central role in Stein's method as a theoretical tool (Barbour & Chen, 2005). Here, we highlight its fundamental connection to the approximate inference problem: if we can exactly solve $\boldsymbol{\phi}$ and $b$ for a given $f$, then the inference problem regarding $f$ is already solved (without running SVGD), since we can easily see that $\mathbb{E}_p f = b$ by taking expectation from both sides of (13).

Mathematically, this reduces the integration problem of estimating $\mathbb{E}_p f$ into solving a differential equation. It suggests that Stein equation is at least as hard as the inference problem itself, and we should not expect a tractable way to solve it in general cases. On the other hand, it suggested that efficient ways of approximate inference may be developed by approximate solutions of Stein equation. Similar idea has been investigated in Oates et al. (2017), which developed a kernel approximation of Stein equation in the case based on a given set of points. SVGD allows us to further extend this idea by optimizing the set of points (particles) on which approximation is defined.

### 3.3 Linear Feature SVGD is Exact for Gaussian

Although Stein equation is difficult to solve in general, it is significantly simplified when the distribution $p$ of interest is Gaussian. In the following, we show that when $p$ is a multivariate Gaussian distribution, we can use linear features, relating to a linear kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}' + 1$, to ensure that SVGD exactly estimates all the first and second order moments of $p$. This insight provides an important practical guidance on the optimal kernel choices for Gaussian-like distributions.

**Theorem 3.3.** *Assume $X^*$ is a fixed point of SVGD with polynomial kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}' + 1$. Let $\mathcal{F}^*$ be the Stein matching set in Theorem 3.2. If $p$ is a multivariate normal distribution on $\mathbb{R}^d$, then $\mathcal{F} \subseteq \mathrm{Poly}(2)$, where $\mathrm{Poly}(2)$ is the set of all polynomials upto the second order, that is, $\mathrm{Poly}(2) = \{\boldsymbol{x}^\top A \boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c \colon A \in \mathbb{R}^{d\times d},\ \boldsymbol{b} \in \mathbb{R}^d,\ c \in \mathbb{R}\}$.*

*Further, denote by $\Phi$ the $(d+1) \times n$ matrix defined by*

$$\Phi = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_n \\ 1 & 1 & \cdots & 1 \end{bmatrix}.$$

*If $\mathrm{rank}(\Phi) \geq d + 1$, then $\mathcal{F} = \mathrm{Poly}(2)$. In this case, any fixed point of SVGD exactly estimates both the mean and the covariance matrix of the target distribution.*

More generally, if the features are polynomials of order $j$, its related Stein matching set should be polynomials of order $j + 1$ for Gaussian distributions. We do not investigate this further because it is less common to estimate higher order moments in multivariate settings.

Theorem 3.3 suggests that it is a good heuristic to include linear features in SVGD, because Gaussian-like distributions appear widely thanks to the central limit theorem and Bernstein–von Mises theorem, and the main goal of inference is often to estimate the mean and variance. In contrast, the more commonly used Gaussian RBF kernel does not have similar exact recovery results for the mean and variance, even for Gaussian distributions.

A nice property of our result is that once we use fewer features than the particles and solve the fixed point exactly, the features do not "interfere" with each other. This allows us to "program" our algorithm by adding different types of features that serve different purposes in different cases.

### 3.4 Random feature SVGD

The linear features are not sufficient for providing the consistent estimation of the whole distribution, even for Gaussian distributions. Non-degenerate kernels are required to obtain bounds on the whole distributions, but they complicate the analysis because their Stein matching set depends on the

solution $X^*$ as shown in Lemma C.1. Random features can be used to sidestep this difficulty (Rahimi & Recht, 2007), enabling us to analyze a random feature variant of SVGD with probabilistic bounds.

To set up, assume $k(\boldsymbol{x}, \boldsymbol{x}')$ is a universal kernel whose Stein discrepancy $\mathbb{D}_k(q \,||\, p)$ yields a discriminative measure of differences between distributions. Assume $k(\boldsymbol{x}, \boldsymbol{x}')$ yields the random feature representation in (5), and we can approximate it by drawing $m$ random features,

$$\hat{k}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{m} \sum_{\ell=1}^{m} \phi(\boldsymbol{x}, \boldsymbol{w}_\ell)\phi(\boldsymbol{x}', \boldsymbol{w}_\ell),$$

where $\boldsymbol{w}_\ell$ are i.i.d. drawn from $p_{\boldsymbol{w}}$. We assume $m \leq n$, then running SVGD with kernel $\hat{k}(\boldsymbol{x}, \boldsymbol{x}')$ (with the random features fixed during the iterations) yields a matching set that decouples with the fixed point $X^*$. In this way, our result below establish that $\mathbb{D}_k(\hat{\mu}_{X^*} \,||\, p) = \tilde{O}(1/\sqrt{n})$ with high probability. According to (4), this provides a uniform bound of $\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f$ for all functions in the unit ball of $\mathcal{H}_p^+$.

Here, random features are introduced mainly for facilitating theoretical analysis, but we also find random feature SVGD works comparably, and sometimes even better than SVGD with the original non-degenerate kernel (see Appendix). This is because with a finite number $n$ of particles, at most $n$ function basis of $k(x, x')$ can be effectively used, even if $k(x, x')$ itself has an infinite rank. From the perspective of moment matching, there is no benefit to use universal kernels when the particle size $n$ is finite.

In the sequel, we first explain the intuitive idea behind our result, highlighting a perspective that views inference as fitting a zero-valued curve with Stein's identity, and then introduce technical details.

**Distributional Inference as Fitting Stein's Identity** Recall that our goal can be viewed as finding particles $X^* = \{\boldsymbol{x}_i^*\}$ such that their empirical $\hat{\mu}_{X^*}$ approximates the target distribution $p$. We re-frame this into finding $\hat{\mu}_{X^*}$ such that Stein's identity holds (approximately):

$$\text{Find } \hat{\mu}_{X^*} \qquad s.t. \qquad \mathbb{E}_{\hat{\mu}_{X^*}}[\mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})] \approx 0, \quad \forall \boldsymbol{w}.$$

We may view this as a special curve fitting problem: considering $\boldsymbol{g}_X(\boldsymbol{w}) = \mathbb{E}_{\hat{\mu}_X}[\mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})]$, we want to find "parameter" $X$ such that $\boldsymbol{g}_X(\boldsymbol{w}) \approx 0$ for all inputs $\boldsymbol{w}$. The kernelized Stein discrepancy (KSD), as shown in (7), can be viewed as the expected rooted mean square loss of this fitting problem:

$$\mathbb{D}_k^2(\hat{\mu}_X \,||\, p) = \mathbb{E}_{\boldsymbol{w} \sim p_{\boldsymbol{w}}} \left[ ||\boldsymbol{g}_X(\boldsymbol{w})||_2^2 \right] \tag{14}$$

When replacing $k(\boldsymbol{x}, \boldsymbol{x}')$ with its random feature approximation $\hat{k}(\boldsymbol{x}, \boldsymbol{x}')$, the corresponding KSD can be viewed as an empirical loss on random sample $\{\boldsymbol{w}_\ell\}$ from $p_{\boldsymbol{w}}$:

$$\mathbb{D}_{\hat{k}}^2(\hat{\mu}_X \,||\, p) = \frac{1}{m} \sum_{\ell=1}^{m} \left[ ||\boldsymbol{g}_X(\boldsymbol{w}_\ell)||_2^2 \right].$$

By running SVGD with $\hat{k}(\boldsymbol{x}, \boldsymbol{x}')$, we acheive $\boldsymbol{g}_{X^*}(\boldsymbol{w}_\ell) = 0$ for all $\ell$ at the fixed point, implying a zero empirical loss $\mathbb{D}_{\hat{k}}(\hat{\mu}_{X^*} \,||\, p) = 0$ assuming the rank condition holds.

The key question, however, is to bound the expected loss $\mathbb{D}_k(\hat{\mu}_{X^*} \,||\, p)$, which can be achieved using generalization bounds in statistical learning theory. In fact, standard results in learning theory suggests that the difference between the empirical loss and expected loss is $O(m^{-1/2})$, yielding $\mathbb{D}_k^2(\hat{\mu}_{X^*} \,||\, p) = O(m^{-1/2})$. However, following (4), this implies $\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f = O(m^{-1/4})$ for $f \in \mathcal{H}_p^+$, which does not acheive the standard $O(m^{-1/2})$. Fortunately, note that our setting is noise-free, and we achieve zero empirical loss; thus, we can get a better rate of $\mathbb{D}_k^2(\hat{\mu}_X \,||\, p) = \tilde{O}(m^{-1})$ using the techniques in Srebro et al. (2010).

**Bound for Random Features** We now present our concentration bounds of random feature SVGD.

**Assumption 3.4.** *1) Assume $\{\phi(\boldsymbol{x}, \boldsymbol{w}_\ell)\}_{\ell=1}^m$ is a set of random features with $\boldsymbol{w}_\ell$ i.i.d. drawn from $p_{\boldsymbol{w}}$ on domain $\mathcal{W}$, and $X^* = \{\boldsymbol{x}_i^*\}_{i=1}^n$ is an* approximate *fixed point of SVGD with random feature $\phi(\boldsymbol{x}, \boldsymbol{w}_\ell)$ in the sense that*

$$|\mathbb{E}_{\boldsymbol{x} \sim \hat{\mu}_{X^*}} \mathcal{P}_{x^j} \phi(\boldsymbol{x}, \boldsymbol{w}_\ell)| \leq \frac{\epsilon_j}{\sqrt{m}}.$$

*where $\mathcal{P}_{x^j}$ is the Stein operator w.r.t. the $j$-th coordinate $x^j$ of $\boldsymbol{x}$. Assume $\epsilon^2 := \sum_{j=1}^d \epsilon_j^2 < \infty$.*

*2) Let* $\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{w}\in\mathcal{W}} |\mathcal{P}_{x^j}\phi(\boldsymbol{x},\boldsymbol{w})| = M_j$*, and* $M^2 := \sum_{j=1}^d M_j^2 < \infty$*. This may imply that* $\mathcal{X}$ *has to be compact since* $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ *is typically unbounded on non-compact* $\mathcal{X}$ *(e.g., when* $p$ *is standard Gaussian,*$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) = \boldsymbol{x}$*).*

*3) Define function set*

$$\mathcal{P}_j\Phi = \{\boldsymbol{w} \mapsto \mathcal{P}_{x^j}\phi(\boldsymbol{x},\boldsymbol{w})\colon \forall \boldsymbol{x} \in \mathcal{X}\}.$$

*We assume the Rademacher complexity of* $\mathcal{P}_j\Phi$ *satisfies* $\mathcal{R}_m(\mathcal{P}_j\Phi) \le R_j/\sqrt{m}$*, and* $R^2 := \sum_{j=1}^d R_j^2 < \infty$*.*

**Theorem 3.5.** *Under Assumption 3.4, for any* $\delta > 0$*, we have with at least probability* $1-\delta$ *(in terms of the randomness of feature parameters* $\{\boldsymbol{w}_\ell\}_{\ell=1}^m$*),*

$$\mathbb{D}_k(\hat{\mu}_{X^*} \,||\, p) \le \frac{C}{\sqrt{m}}\left[\epsilon^2 + \log^3 m + \log(1/\delta)\right]^{1/2}, \tag{15}$$

*where* $C$ *is a constant that depends on* $R$ *and* $M$*.*

**Remark** Recalling (4), Eq (15) provides a uniform bound

$$\sup_{||f||_{\mathcal{H}_p^+}\le 1} \left\{\mathbb{E}_{\mu_{X^*}} f - \mathbb{E}_p f\right\} = O(m^{-1/2}\log^{1.5} m).$$

This is a uniform bound that controls the worse error uniformly among all $f \in \mathcal{H}_p^+$. It is unclear if the logarithm factor $\log^{1.5} m$ is essential. In the following, we present a result that has an $O(1/\sqrt{m})$ rate, without the logarithm factor, but only holds for individual functions.

**Theorem 3.6.** *Let* $\mathcal{F}_\infty$ *be the set of linear span of the Steinalized features:*

$$f(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[\boldsymbol{v}(\boldsymbol{w})^\top \mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x},\boldsymbol{w})], \tag{16}$$

*where* $\boldsymbol{v}(\boldsymbol{w}) = [v_1(\boldsymbol{w}),\ldots,v_d(\boldsymbol{w})] \in \mathbb{R}^d$ *is the combination weights that satisfy* $\sup_{\boldsymbol{w}} ||\boldsymbol{v}(\boldsymbol{w})||_\infty < \infty$*. We may define a norm on* $\mathcal{F}_\infty$ *by* $||f||_{\mathcal{F}_\infty}^2 := \inf_{\boldsymbol{v}} \sum_{j=1}^d \sup_{\boldsymbol{w}} |v_j(\boldsymbol{w})|^2$*, where* $\inf_{\boldsymbol{v}}$ *is taken on all* $\boldsymbol{v}(w)$ *that satisfies* (16)*.*

*Assume Assumption 3.4 holds, then for any given function* $f \in \mathcal{F}_\infty$ *with* $||f||_{\mathcal{F}_\infty} \le 1$*, we have with at least probability* $1-\delta$*,*

$$|\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f| \le \frac{C}{\sqrt{m}}(1 + \epsilon + \sqrt{2\log(1/\delta)}),$$

*where* $C$ *is a constant that depends on* $R$ *and* $M$*.*

The $\mathcal{F}_\infty$ defined above is closely related to the RKHS $\mathcal{H}_p$. In fact, one can show that $\mathcal{F}_\infty$ is a dense subset of $\mathcal{H}_p$ (Rahimi & Recht, 2008) and is hence quite rich if $k(\boldsymbol{x},\boldsymbol{x}')$ is set to be universal.

## 4 Conclusion

We analyze SVGD through the eyes of moment matching. Our results are non-asymptotic in nature and provide an insightful framework for understanding the influence of kernels in the behavior of SVGD fixed points. Our framework suggests promising directions to develop systematic ways of optimizing the choice of kernels, especially for the query-specific inference that focuses on specific test functions. A particularly appealing idea is to "program" the inference algorithm by adding features that serve specific purposes so that the algorithm can be easily adapted to meet the needs of different users. In general, we expect that the connection between approximation inference and Stein's identity and Stein equation will provide further opportunities for deriving new generations of approximate inference algorithms.

Another advantage of our framework is that it separates the design of the fixed point equation with the numerical algorithm used to achieve the fixed point. In this way, the iterative algorithm does not have to be derived as an approximation of an infinite dimensional gradient flow, in contrast to the original SVGD. This allows us to apply various practical numerical methods and acceleration techniques to solve the fixed point equation faster, with convergence guarantees.

## References


Anderes, Ethan and Coram, Marc. A general spline representation for nonparametric and semiparametric density estimates using diffeomorphisms. In *arXiv preprint arXiv:1205.5314*, 2002.

Barbour, Andrew D and Chen, Louis Hsiao Yun. *An introduction to Stein's method*, volume 4. World Scientific, 2005.

Bartlett, Peter L and Mendelson, Shahar. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Chwialkowski, Kacper, Strathmann, Heiko, and Gretton, Arthur. A kernel test of goodness-of-fit. *International Conference on Machine Learning (ICML)*, 2016.

Feng, Yihao, Wang, Dilin, and Liu, Qiang. Learning to draw samples with amortized Stein variational gradient descent. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Gorham, Jackson and Mackey, Lester. Measuring sample quality with kernels. *International Conference on Machine Learning (ICML)*, 2017.

Haarnoja, Tuomas, Tang, Haoran, Abbeel, Pieter, and Levine, Sergey. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning (ICML)*, pp. 1352–1361, 2017.

Kim, Taesup, Yoon, Jaesik, Dia, Ousmane, Kim, Sungwoong, Bengio, Yoshua, and Ahn, Sungjin. Bayesian model-agnostic meta-learning. In *Advances In Neural Information Processing Systems (NIPS)*, 2018.

Koller, Daphne and Friedman, Nir. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Liu, Qiang and Wang, Dilin. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems (NIPS)*, pp. 2378–2386, 2016.

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

Liu, Yang, Ramachandran, Prajit, Liu, Qiang, and Peng, Jian. Stein variational policy gradient. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Lu, Jianfeng, Lu, Yulong, and Nolen, James. Scaling limit of the stein variational gradient descent part i: the mean field regime. *arXiv preprint arXiv:1805.04035*, 2018.

Oates, Chris J, Girolami, Mark, and Chopin, Nicolas. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 2017.

Ollivier, Yann, Pajot, Hervé, and Villani, Cédric. *Optimal Transport: Theory and Applications*, volume 413. Cambridge University Press, 2014.

Pu, Yuchen, Gan, Zhe, Henao, Ricardo, Li, Chunyuan, Han, Shaobo, and Carin, Lawrence. VAE learning via Stein variational gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 4239–4248, 2017.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1177–1184, 2007.

Rahimi, Ali and Recht, Benjamin. Uniform approximation of functions with random bases. In *Communication, Control, and Computing, 46th Annual Allerton Conference on*, pp. 555–561. IEEE, 2008.

Srebro, Nathan, Sridharan, Karthik, and Tewari, Ambuj. Smoothness, low noise and fast rates. In *Advances in neural information processing systems (NIPS)*, pp. 2199–2207, 2010.

Stein, Charles. 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*, pp. 583–602, 1972.

Wainwright, Martin J, Jordan, Michael I, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, Dilin and Liu, Qiang. Learning to draw samples: With application to amortized MLE for generative adversarial learning. *arXiv preprint arXiv:1611.01722*, 2016.

Zhu, Yinhao and Zabaras, Nicholas. Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, 366:415–447, 2018.

## A Proof of Theorem 3.3

*Proof.* Assume $p$ is multivariate normal $\mathcal{N}(\boldsymbol{\mu}, Q^{-1})$ where $Q$ is the inverse covariance matrix. We have $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) = -Q(\boldsymbol{x} - \boldsymbol{\mu})$. Since $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}' + 1$, the functions in $\mathcal{F}^*$ should have a form of

$$\begin{aligned} f(\boldsymbol{x}) &= \sum_{i=1}^{n} \boldsymbol{a}_i^\top [-Q(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x}_i^\top \boldsymbol{x} + 1) + \boldsymbol{x}_i] + b \\ &= \boldsymbol{x}^\top W \boldsymbol{x} + \boldsymbol{v}^\top \boldsymbol{x} + c, \end{aligned}$$

where

$$W = -\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{a}_i^\top Q,$$

$$\boldsymbol{v} = \sum_{i=1}^{n} (\boldsymbol{\mu}^\top \boldsymbol{x}_i - 1) Q \boldsymbol{a}_i$$

$$c = b + \sum_{i=1}^{n} \boldsymbol{a}_i^\top (Q\boldsymbol{\mu} + \boldsymbol{x}_i).$$

Denote by $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ the $(d \times n)$ matrix, $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ the $(d \times n)$ matrix, and $B = QA$. We have

$$W = -XB^\top, \tag{17}$$

$$\boldsymbol{v}^\top = (\boldsymbol{\mu}^\top X - \boldsymbol{e}^\top) B^\top, \tag{18}$$

$$c = b + \boldsymbol{e}^\top B^\top \boldsymbol{\mu} + \operatorname{tr}(XA^\top), \tag{19}$$

where $\boldsymbol{e}$ is the $\mathbb{R}^d$-vector of all ones. Eq. (17) and (18) are equivalent to

$$\begin{bmatrix} -X \\ \boldsymbol{\mu}^\top X - \boldsymbol{e} \end{bmatrix} B = \begin{bmatrix} W \\ \boldsymbol{v}^\top \end{bmatrix} \tag{20}$$

We just need to show that for any value of $W \in \mathbb{R}^{d \times d}$, $\boldsymbol{v} \in \mathbb{R}^d$ and $c \in \mathbb{R}$ there exists $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ and $b$ that satisfies the above equation. This is equivalent to

$$\begin{bmatrix} -I, 0 \\ \boldsymbol{\mu}^\top, -1 \end{bmatrix} \Phi B = \begin{bmatrix} W \\ \boldsymbol{v}^\top \end{bmatrix}$$

Since $\begin{bmatrix} -I, 0 \\ \boldsymbol{\mu}^\top, -1 \end{bmatrix}$ is always full rank, if $\Phi$ has a rank at least $d + 1$, then (20) exits a solution for $B$. We can then get $A = Q^{-1}B$ and solve $b$ from (19). □

## B Proof of Theorem 3.5

*Proof.* A loss function is $H$-smooth iff its derivative is $H$-Lipschitz. For twice differentiable $\phi$, this just means $|\phi''| \leq H$. The following result from Srebro et al. (2010) is key to our proof.

**Theorem B.1** (Srebro et al. (2010) Theorem 1)**.** *For an $H$-smooth non-negative loss $\phi$, such that $\forall_{x,y,h} |\phi(h(x), y)| \leq b$, for any $\delta > 0$, we have with probability at least $1 - \delta$ over a random sample of size $n$ that, for any $h \in \mathcal{H}$, we have*

$$L(h) \leq \hat{L}(h) + K\left[\sqrt{\hat{L}(h)}\left(\sqrt{H}\log^{1.5} n\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{b\log(1/\delta)}{n}}\right) + H\log^3 n\mathcal{R}_n^2(\mathcal{H}) + \frac{b\log(1/\delta)}{n}\right].$$

*where $K$ is a numerical constant that satisfies $K < 10^5$.*

We now apply this result to bound kernelized Stein discrepancy. Take $\phi(x, y) = (x - y)^2$, then $H = 2$. Define $g_{X,j}(\boldsymbol{w}) = \mathbb{E}_{\hat{\mu}_X}[\mathcal{P}_{x^j}\phi(\boldsymbol{x}, \boldsymbol{w})]$ and $\mathcal{G}_j = \{g_{X,j} \colon \forall X \in \mathcal{X}^n\}$. Recall that the Stein discrepancy can be viewed as the sum of mean square losses of fitting $g_{X,j}$ to the zero-valued line:

$$\mathbb{D}_k^2(\hat{\mu}_X \,||\, p) = \sum_{j=1}^{d} L_j(g_{X,j}), \qquad \text{where} \qquad L_j(g_{X,j}) = \mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[(g_{X,j}(\boldsymbol{w}) - 0)^2],$$

$$\mathbb{D}_{\hat{k}}^2(\hat{\mu}_X \,||\, p) = \sum_{j=1}^{d} \hat{L}_j(g_{X,j}), \qquad \text{where} \qquad \hat{L}_j(g_{X,j}) = \frac{1}{m}\sum_{\ell=1}^{m}[(g_{X,j}(\boldsymbol{w}_\ell) - 0)^2].$$

We now apply Theorem B.1 to each bound the difference between the expected loss $L_j(g_{X,j})$ and the empirical loss $\hat{L}_j(g_{X,j})$. From Assumption 3.4.2, we have $\sup_{X,\boldsymbol{w}} |g_{X,j}(\boldsymbol{w})| \le M_j$. This is because

$$|g_{X,j}(\boldsymbol{w})| = |\frac{1}{n}\sum_{i=1}^{n} \mathcal{P}_{x^j}\phi(\boldsymbol{x}, \boldsymbol{w})| \le \frac{1}{n}\sum_{i=1}^{n} |\mathcal{P}_{x^j}\phi(\boldsymbol{x}, \boldsymbol{w})| \le M_j.$$

Using Theorem B.1, we have with probability $1 - \delta$, for any $X$,

$$L_j(g_{X,j}) \le \hat{L}_j(g_{X,j}) + K\bigg[\hat{L}_j(g_{X,j})\bigg(\sqrt{2}\log^{1.5} m\mathcal{R}_m(\mathcal{G}_j) + \sqrt{\frac{M_j^2\log(1/\delta)}{m}}\bigg) + \; 2\log^3 m\mathcal{R}_m^2(\mathcal{G}_j) + \frac{M_j^2\log(1/\delta)}{m}\bigg].$$

By Assumption 3.4.1, $|g_{X^*,j}(\boldsymbol{w}_\ell)| \le \frac{\epsilon_j}{\sqrt{m}}$ for $\forall \ell = 1, \ldots, m$ at the approximate fixed point $X^*$. We have $\hat{L}_j(g_{X,j}) \le \frac{\epsilon_j}{\sqrt{m}}$. By Assumption 3.4.3, we have $\mathcal{R}_m(\mathcal{G}_j) \le R_j/\sqrt{m}$. Therefore,

$$L_j(g_{X^*,j}) \le \frac{\epsilon_j^2}{m} + K\bigg[\frac{\epsilon_j}{\sqrt{m}}\bigg(\sqrt{2}\log^{1.5} m\frac{R_j}{\sqrt{m}} + \sqrt{\frac{M_j^2\log(1/\delta)}{m}}\bigg) + 2\log^3 m\frac{R_j^2}{m} + \frac{M_j^2\log(1/\delta)}{m}\bigg].$$

Summing across $j = 1, \ldots, d$, we get

$$\begin{aligned}
&\mathbb{D}_k^2(\hat{\mu}_{X^*} \,||\, p) \\
&\le \frac{1}{m}\sum_{j=1}^{d}\bigg[\epsilon_j^2 + K\bigg(\sqrt{2}R_j\epsilon_j\log^{1.5} m + M_j\epsilon_j\sqrt{\log(1/\delta)} + 2R_j^2\log^3 m + M_j^2\log(1/\delta)\bigg)\bigg] \\
&\le \frac{1}{m}\bigg[\epsilon^2 + K\bigg(\sqrt{2}R\epsilon\log^{1.5} m + M\epsilon\sqrt{\log(1/\delta)} + 2R^2\log^3 m + M^2\log(1/\delta)\bigg)\bigg] \\
&\le \frac{C^2}{m}\left[\epsilon^2 + \log^3 m + \log(1/\delta)\right],
\end{aligned}$$

where $C^2 = \max\{1 + \frac{1}{\sqrt{2}}KR + \frac{1}{2}M,\ \frac{1}{\sqrt{2}}KR + 2KR^2,\ \frac{1}{2}KM + KM^2\}$. □

## C Proof of Theorem 3.6

*Proof.* By Stein's identity $\mathbb{E}_{\boldsymbol{x}\sim p}[\mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})] = 0$, we have $\mathbb{E}_p f = 0$ for $\forall f \in \mathcal{F}_\infty$. This is because, assuming $f(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[\boldsymbol{v}(\boldsymbol{w})^\top \mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})]$,

$$\mathbb{E}_p f = \mathbb{E}_{\boldsymbol{x}\sim p}\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[v(\boldsymbol{w})^\top \mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})] = \mathbb{E}_{p_{\boldsymbol{w}}}[v(\boldsymbol{w})^\top \mathbb{E}_{\boldsymbol{x}\sim p}[\mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})]] = 0.$$

Therefore,

$$\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f = \mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[f(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[\mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[v(\boldsymbol{w})^\top \mathcal{P}_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{w})]].$$

This gives

$$\begin{aligned}|\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f| &= \left| \sum_{j=1}^{d} \mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[\mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[v_j(\boldsymbol{w})\mathcal{P}_{x^j}\phi(\boldsymbol{x},\boldsymbol{w})]] \right| \\ &\leq \sum_{j=1}^{d} |\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[\mathbb{E}_{\boldsymbol{x}\sim\hat{\mu}_{X^*}}[v_j(\boldsymbol{w})\mathcal{P}_{x^j}\phi(\boldsymbol{x},\boldsymbol{w})]]| \\ &= \sum_{j=1}^{d} |\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[v_j(\boldsymbol{w})g_{X^*,j}(\boldsymbol{w})]| \,.\end{aligned}$$

Let $h_{X,j}(\boldsymbol{w}) = v_j(\boldsymbol{w})g_{X,j}(\boldsymbol{w})$. Then Assumption 3.4.1-2 gives $\sup_{\boldsymbol{w}} |h_{X^*,j}(\boldsymbol{w}_\ell)| \leq \frac{\epsilon_j M_j}{\sqrt{m}},\ \forall j = 1,\ldots,m$. We have

$$\begin{aligned}|\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[h_{X^*,j}(\boldsymbol{w})]| &\leq |\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[h_{X^*,j}(\boldsymbol{w})] - \frac{1}{m}\sum_{\ell=1}^{m} h_{X^*,j}(\boldsymbol{w}_\ell)| \ + \ |\frac{1}{m}\sum_{\ell=1}^{m} h_{X^*}(\boldsymbol{w}_\ell)| \\ &\leq \sup_{h_{X,j}\in v_j\mathcal{G}_j} |\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[h_{X,j}(\boldsymbol{w})] - \frac{1}{m}\sum_{\ell=1}^{m} h_{X,j}(\boldsymbol{w}_\ell)| \ + \ \frac{\epsilon_j M_j}{\sqrt{m}},\end{aligned}$$

where $v_j\mathcal{G}_j = \{\boldsymbol{w} \mapsto v_j(\boldsymbol{w})g_{X,j}(\boldsymbol{w}) \colon \ X \in \mathcal{X}^n\}$. Therefore, we just need bound

$$\Delta_\ell(\boldsymbol{w}_1,\ldots\boldsymbol{w}_m) \stackrel{def}{=} \sup_{h_{X,j}\in v_j\mathcal{G}_j} |\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[h_{X,j}(\boldsymbol{w})] - \frac{1}{m}\sum_{\ell=1}^{m} h_{X,j}(\boldsymbol{w}_\ell)|.$$

This can be done using standard techniques in uniform concentration bounds. To do this, note that any $\boldsymbol{w}_\ell$ and $\boldsymbol{w}'_\ell$,

$$\begin{aligned}&|\Delta_\ell(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_\ell,\ldots,\boldsymbol{w}_m) - \Delta_\ell(\boldsymbol{w}_1,\ldots,\boldsymbol{w}'_\ell,\ldots,\boldsymbol{w}_m)| \\ &\leq \frac{2\sup_{h_{X,j}\in v_j\mathcal{G}_j}\sup_{\boldsymbol{w}}|h_{X,j}(\boldsymbol{w})|}{m} \leq \frac{2V_jM_j}{m},\end{aligned}$$

where we assume $\sup_{\boldsymbol{w}} |v_j(\boldsymbol{w})| = V_j$. By Mcdiarmid's inequality, we have

$$\Pr(\Delta_\ell(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m) > \mathbb{E}[\Delta_\ell(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m)] + t) \leq \exp(-\frac{mt^2}{2V_j^2M_j^2}).$$

On the other hand, the expectation $\mathbb{E}[\Delta_\ell(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m)]$ can be bounded by Rademacher complexity of $v_j\mathcal{G}_j$:

$$\mathbb{E}[\Delta_\ell(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_m)] \leq 2\mathcal{R}_m(v_j\mathcal{G}_j).$$

Restating the result, we have with probability $1-\delta$, for $\forall\delta > 0$,

$$\sup_{h_{X,j}\in v_j\mathcal{G}_j} \left| \mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[h_{X,j}(\boldsymbol{w})] - \frac{1}{m}\sum_{\ell=1}^{m} h_{X,j}(\boldsymbol{w}_\ell) \right| \leq 2\mathcal{R}_m(v_j\mathcal{G}_j) + V_jM_j\sqrt{\frac{2\log(1/\delta)}{m}}.$$

Overall, this gives

$$\begin{aligned}|\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f| &= \sum_{j=1}^{d} |\mathbb{E}_{\boldsymbol{w}\sim p_{\boldsymbol{w}}}[h_{X^*,j}(\boldsymbol{w})]| \\ &\leq \sum_{j=1}^{d} \left( 2\mathcal{R}_m(v_j\mathcal{G}_j) + V_jM_j\sqrt{\frac{2\log(1/\delta)}{m}} + \frac{\epsilon_jM_j}{\sqrt{m}} \right). \\ &= \frac{1}{\sqrt{m}}\left( VM\sqrt{2\log(1/\delta)} + \epsilon M \right) + 2\sum_{j=1}^{d} \mathcal{R}_m(v_j\mathcal{G}_j),\end{aligned}$$

where we use the fact that $V^2 = \sum_{j=1}^d V_j^2$, $M^2 = \sum_{j=1}^d M_j^2$ and $\epsilon^2 = \sum_{j=1}^d \epsilon_j^2$.

We just need to bound the Rademacher complexity $\mathcal{R}_m(v_j\mathcal{G}_j)$. This requires recalling some properties of Rademacher complexity. Let $\{\mathcal{F}_j \colon j = 1, \ldots, n\}$ be a set of function sets, and $\frac{1}{n}\sum_{j=1}^n \mathcal{F}_i$ be the set of functions consisting of functions of form $\frac{1}{n}\sum_{j=1}^n f_i, \forall f_i \in \mathcal{F}_i$. Then we have (see, e.g., Bartlett & Mendelson (2002))

$$\mathcal{R}_m(\frac{1}{n}\sum_{j=1}^n \mathcal{F}_i) \leq \frac{1}{n}\sum_{i=1}^n \mathcal{R}_m(\mathcal{F}_i).$$

Applying this to $\mathcal{G}_j$, we have

$$\mathcal{R}_m(\mathcal{G}_j) \leq \mathcal{R}_m(\mathcal{P}_j\Phi) \leq \frac{R_j}{\sqrt{m}}.$$

Further, applying Lemma C.1.5) below, we have

$$\mathcal{R}_m(v_j\mathcal{G}_j) \leq 2(\mathcal{R}_m(\mathcal{G}_j) + \frac{V_j}{\sqrt{m}})(M_j + V_j) \leq \frac{2}{\sqrt{m}}(R_j + V_j)(M_j + V_j) \leq \frac{2}{\sqrt{m}}(R_j^2 + 2{V_j}^2 + M_j^2)$$

Therefore,

$$\sum_{j=1}^d \mathcal{R}_m(v_j\mathcal{G}_j) \leq \frac{2}{\sqrt{m}}(R^2 + 2V^2 + M^2).$$

Putting everything together, we get

$$|\mathbb{E}_{\hat{\mu}_{X^*}} f - \mathbb{E}_p f| \leq \frac{1}{\sqrt{m}}\left(VM\sqrt{2\log(1/\delta)} + \epsilon M + 2R^2 + 4V^2 + 2M^2\right).$$

This concludes the proof. □

### C.1 Rademacher Complexity

The following Lemma collects some basic properties of Rademacher complexity. See Bartlett & Mendelson (2002) for more information.

For a function set $\mathcal{F}$, its Rademacher complexity is defined as

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right|\right],$$

where the expectation is taken when $\sigma_i$ are i.i.d. uniform $\{\pm 1\}$-valued random variables and $x_i$ are i.i.d. random variables from some underlying distribution. A basic property of Rademacher complexity is that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m f(x_i) - \mathbb{E}f\right|\right] \leq 2\mathcal{R}_m(\mathcal{F}).$$

*Proof.*

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{\ell=1}^m f(x_i) - \mathbb{E}f\right|\right] &\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m (f(x_i) - f(x_i'))\right|\right] \\
&= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i(f(x_i) - f(x_i'))\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i f(x_i)\right|\right] \\
&= 2\mathcal{R}_m[\mathcal{F}].
\end{aligned}$$

□

**Lemma C.1.** *Let $\mathcal{F}$, $\mathcal{F}_1$ and $\mathcal{F}_2$ are real-valued function classes.*

*1) Define $\mathcal{F}_1 + \mathcal{F}_2 = \{f + g \colon f \in \mathcal{F}_1,\ g \in \mathcal{F}_2\}$. We have*

$$\mathcal{R}_m(\mathcal{F}_1 + \mathcal{F}_2) \leq \mathcal{R}_m(\mathcal{F}_1) + \mathcal{R}_m(\mathcal{F}_2).$$

*2) Let $\phi \colon \mathbb{R} \to \mathbb{R}$ be an $L_\phi$-Lipschitz function. Define $\phi \circ \mathcal{F} = \{\phi \circ f \colon \forall f \in \mathcal{F}\}$. We have*

$$\mathcal{R}_m(\phi \circ \mathcal{F}) \leq 2L_\phi \mathcal{R}_m(\mathcal{F}) + \frac{\phi(0)}{m}.$$

*3) For any uniformly bounded function $g$, we have*

$$\mathcal{R}_m(\mathcal{F} + g) \leq \mathcal{R}_m(\mathcal{F}) + \frac{||g||_\infty}{\sqrt{m}}.$$

*4) For constant $c \in \mathbb{R}$ and $c\mathcal{F} = \{x \mapsto cf(x) \colon \forall f \in \mathcal{F}\}$,*

$$\mathcal{R}_m(c\mathcal{F}) = |c|\mathcal{R}_m(\mathcal{F}).$$

*5) Define $g\mathcal{F} = \{x \mapsto f(x)g(x) \colon \forall f \in \mathcal{F}\}$. Assume $||\mathcal{F}||_\infty \coloneqq \sup_{f \in \mathcal{F}} ||f||_\infty < \infty$, we have*

$$\mathcal{R}_m(g\mathcal{F}) \leq 2(\mathcal{R}_m[\mathcal{F}] + \frac{||g||_\infty}{\sqrt{m}})(||\mathcal{F}||_\infty + ||g||_\infty).$$

*Proof.* 1) - 4) are standard results; see Theorem 12 in Bartlett & Mendelson (2002).

For 5), note that

$$fg = \frac{1}{4}(f + g)^2 - \frac{1}{4}(f - g)^2.$$

3) gives

$$\mathcal{R}_m(\mathcal{F} \pm g) \leq \mathcal{R}_m[\mathcal{F}] + \frac{||g||_\infty}{\sqrt{m}}$$

Further, note that $\phi(x) = x^2$ is $2(||F||_\infty + ||g||_\infty)$-Lipschitz on interval $[-||F||_\infty - ||g||_\infty, ||F||_\infty + ||g||_\infty]$. Applying 2) and then 1) and 4) gives

$$\mathcal{R}_m(g\mathcal{F}) \leq 2(||\mathcal{F}||_\infty + ||g||_\infty)(\mathcal{R}_m(\mathcal{F}) + \frac{||g||_\infty}{\sqrt{m}}).$$

□

Our results require bounding the Rademacher complexity $\mathcal{R}_m(\mathcal{P}_j\Phi)$ of the Steinalized features, $\mathcal{P}_j\Phi = \{w \mapsto \mathcal{P}_{x^j}\phi(\boldsymbol{x}, w) \colon \boldsymbol{x} \in \mathcal{X}\}$. The following result bounds the Rademacher complexity of the Steinalized set using the complexity of the original feature set and its gradient set.

**Lemma C.2.** *Define $\Phi = \{\boldsymbol{w} \mapsto \phi(\boldsymbol{x}, \boldsymbol{w}) \colon \forall \boldsymbol{x} \in \mathcal{X}\}$ and $\nabla_j\Phi = \{w \mapsto \nabla_{x^j}\phi(\boldsymbol{x}, \boldsymbol{w}) \colon \forall \boldsymbol{x} \in \mathcal{X}\}$. Then*

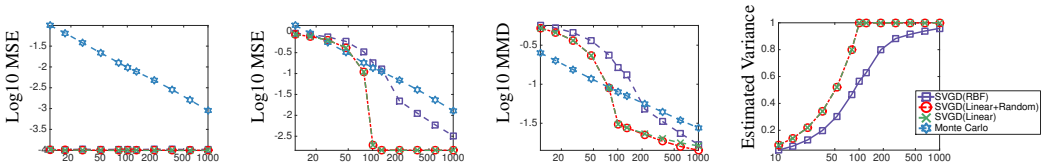
$$\mathcal{R}_m(\mathcal{P}_j\Phi) \leq ||\nabla_{\boldsymbol{x}_\ell} \log p||_\infty \mathcal{R}_m(\Phi) + \mathcal{R}_m(\nabla_j\Phi),$$

*where $||\nabla_{\boldsymbol{x}_\ell} \log p||_\infty = \sup_{\boldsymbol{x} \in \mathcal{X}} |\nabla_{\boldsymbol{x}_\ell} \log p(\boldsymbol{x})|$.*

# D Empirical Experiments

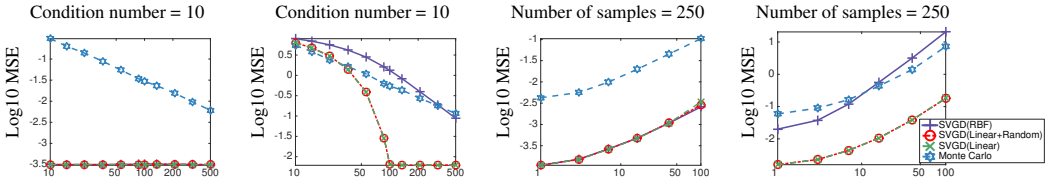
Our results show that linear features allow us to obtain accurate estimates of the first and second moments for Gaussian-like distributions, while random features can obtain a good overall distributional approximation with high probability. To test these theoretical observations empirically, we design a "linear+random" kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \alpha(1 + \boldsymbol{x}^\top \boldsymbol{x}') + \beta \sum_{\ell=d+2}^{n} \phi(\boldsymbol{x}, \boldsymbol{w}_\ell)\phi(\boldsymbol{x}', \boldsymbol{w}_\ell),$$




(a) $\mathbb{E}(x)$ (b) $\mathbb{E}(x^2)$ (c) MMD (d) Variance

Figure 1: Results on standard Gaussian distribution ($d = 100$). (a)-(b) show the MSE when using the obtained particles to estimate the mean and second order moments of each dimension, averaged across the dimensions. (c) shows the maximum mean discrepancy between the particle distribution and true distribution. (d) shows the average values of the estimated variance (the true variance is 1).




(a) $\mathbb{E}(x)$ (b) $\mathbb{E}(x^2)$ (c) $\mathbb{E}(x)$ (d) $\mathbb{E}(x^2)$

Figure 2: (a)-(b) Results on random 100 dimensional non-spherical Gaussian distributions whose covariance matrix has a conditional number of $\lambda_{\max}/\lambda_{\min} = 10$. (c)-(d) The performance on random non-spherical Gaussian distributions with different conditional numbers. Results averaged on 20 random models.

where we take $\alpha = 1/(d+1)$ and $\beta = 1/(n-d-1)$ in our experiments. In the case when there are fewer particles than dimension plus one ($n \leq d+1$), we have $k(\boldsymbol{x}, \boldsymbol{x}') = 1 + \boldsymbol{x}^\top \boldsymbol{x}'$, which only include the linear features, and when $n > (d+1)$, additional random features are added, so that the total number of features matches the number of particles.

We take $\phi(\boldsymbol{x}, \boldsymbol{w})$ to be the random cosine feature in (6) to approximate the Gaussian RBF kernel. Note that in our method, the random parameters $\{\boldsymbol{w}_\ell\}$ are drawn in the beginning and fixed across the iterations of the algorithm, but we adopt the bandwidth $h$ across the iterations using the median trick. We compare exact Monte Carlo with SVGD with different kernels, including the standard Gaussian RBF kernel, the linear kernel $k(\boldsymbol{x}, \boldsymbol{x}') = 1 + \boldsymbol{x}^\top \boldsymbol{x}'$, and the linear+random kernel defined above.

**Gaussian Models** We start with verifying our theory on a simple standard Gaussian distribution $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}, 0, I)$ with $d = 100$ dimensions. In Figure 1, we can see that all SVGD methods estimate the mean parameters exceptionally well (Figure 1(a)). Variance estimation is more difficult for SVGD in general, but both the Linear+Random and Linear kernels perform well as the theory predicts: the errors drop quickly as $n$ approaches $d+1$ (the minimum particle size needed to recover mean and covariance matrices), and only the numerical error is left when $n > d+1$.

To examine the variance estimation more closely, we show in Figure 1(d) the value of the estimated variance (averaged across the dimensions) on the same 100-dimensional standard Gaussian distribution. We find that all the variants of SVGD tend to underestimate the variance when there is insufficient number of particles (in particular, when $n < d+1$), but the kernels that include linear features give (near) exact estimation once $n \geq d+1$.

Figure 2 shows a similar plot for 100-dimensional non-spherical Gaussian distributions when the conditional number of the covariance matrix varies. In particular, we set $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x};\ \boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} \sim \mathrm{Unif}([-3, 3])$ and $\Sigma = I + \alpha \Lambda \Lambda^\top$, with the elements of $\Lambda$ drawn from $\mathcal{N}(0, 1)$ and $\alpha$ adjusted to make the conditional number $\lambda_{\max}/\lambda_{\min}$ of $\Sigma$ equal specific numbers. When the condition number equals 1, we should have $\Sigma = I$.

Figure 2(a)-(b) show the estimation of the first and second order moments when the conditional number equals 10, in which SVGD(linear+random) and SVGD(linear) again show a near exact

recovery after $n > d + 1$. Figure 2(c)-(d) show that as the conditional number increases, the accuracy of all the methods decreases, but SVGD(linear+random) and SVGD(linear) still significantly outperform Monte Carlo estimation. The increased errors in SVGD(linear+random) and SVGD(linear) are caused by the increase of numerical error because it is more difficult to satisfy the fixed point equation with high accuracy when the conditional number is large.

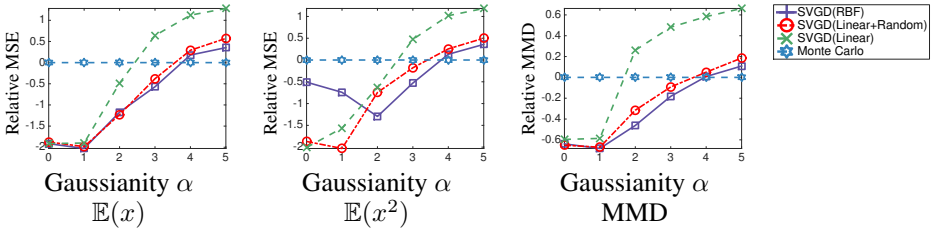



Figure 3: Results on Gaussian mixture models $p(\boldsymbol{x}) = \sum_{k=1}^{15} \mathcal{N}(\alpha \boldsymbol{\mu}_k, I)$, where $\boldsymbol{\mu}_k \sim \mathrm{Uniform}([0, 1])$ and $\alpha$ controls the Gaussianity of $p$ (when $\alpha = 0$, $p$ is standard Gaussian). All the results are the relative performance w.r.t. exact Monte Carlo sampling method with the sample size (we fix for all the methods). We fix $n = 100$ for all the methods and average the result over 20 random models.

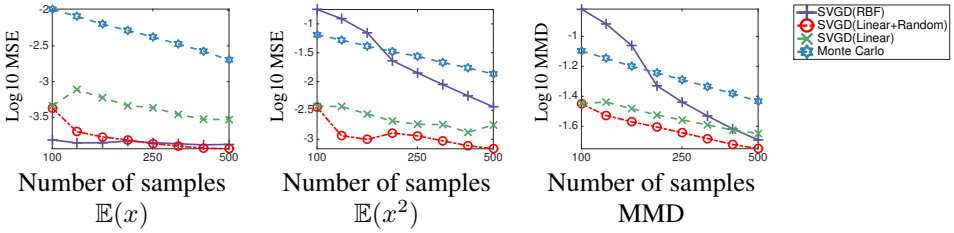



Figure 4: Results on randomly generated Gaussian-Bernoulli RBM, averaged on 20 trials.

**Gaussian Mixture Models** We consider a Gaussian mixture model with density fucntion $p(\boldsymbol{x}) = \frac{1}{15} \sum_{j=1}^{15} \mathcal{N}(\boldsymbol{x}; \alpha \boldsymbol{\mu}_j, I)$, where $\boldsymbol{\mu}_j$ is randomly drawn from $\mathrm{Unifrom}([0, 1])$, and $\alpha$ can be viewed as controlling the Gaussianity of $p(\boldsymbol{x})$: when $\alpha$ equals zero, $p(\boldsymbol{x})$ reduces to the standard Gaussian distribution, while when $\alpha$ is large, $p(\boldsymbol{x})$ would be highly multimodal with mixture components far away from each other.

Figure D shows the relative performance of SVGD with different kernels compared to exact Monte Carlo sampling. We find that SVGD methods generally outperform Monte Carlo unless $\alpha$ is very large. In Figure D(b), we can see that SVGD(Linear) outperforms SVGD(RBF) when $p$ is close to Gaussian (small $\alpha$), and performs worse than SVGD(RBF) when $p$ is highly non-Gaussian (large $\alpha$). SVGD(Linear+Random) combines the advantages of both and tends to match the best of SVGD(Linear) and SVGD(RBF) in all the range of $\alpha$.

**Gaussian-Bernoulli RBM** Gaussian-Bernoulli RBM is a hidden variable model consisting of a continuous observable variable $\boldsymbol{x} \in \mathbb{R}^d$ and a binary hidden variable $\boldsymbol{h} \in \{\pm 1\}^{d'}$ with probability

$$p(\boldsymbol{x}, \boldsymbol{h}) \propto \sum_{\boldsymbol{h} \in \{\pm\}^{d'}} \exp(\boldsymbol{x}^\top B \boldsymbol{h} + \boldsymbol{b}^\top \boldsymbol{x} + \boldsymbol{c}^\top \boldsymbol{h} - \frac{1}{2} ||\boldsymbol{x}||_2^2),$$

where we randomly draw $\boldsymbol{b}$ and $\boldsymbol{c}$ from $\mathcal{N}(0, I)$, and the elements of $B$ from $\mathrm{Uniform}(\{\pm 0.1\})$. We use $d = 100$ observable variables and $d' = 10$ hidden variables, so $p(\boldsymbol{x})$ is effectively a Gaussian mixture with $2^{10}$ components. The results are shown in Figure D, where we find that SVGD(Linear+Random) again achieves the best performance in terms of all the evaluation metrics.