Improved Stein Variational Gradient Descent with Importance Weights

Lukang Sun
KAUST
lukang.sun@kaust.edu.sa

Peter Richtárik
KAUST
peter.richtarik@kaust.edu.sa

Abstract

Stein Variational Gradient Descent (SVGD) is a popular sampling algorithm used in various machine learning tasks. It is well known that SVGD arises from a discretization of the kernelized gradient flow of the Kullback-Leibler divergence $D_{\mathrm{KL}}(\cdot \mid \pi)$, where π is the target distribution. In this work, we propose to enhance SVGD via the introduction of *importance weights*, which leads to a new method for which we coin the name β-SVGD. In the continuous time and infinite particles regime, the time for this flow to converge to the equilibrium distribution π, quantified by the Stein Fisher information, depends on ρ_0 and π very weakly. This is very different from the kernelized gradient flow of Kullback-Leibler divergence, whose time complexity depends on $D_{\mathrm{KL}}(\rho_0 \mid \pi)$. Under certain assumptions, we provide a descent lemma for the population limit β-SVGD, which covers the descent lemma for the population limit SVGD when $\beta \to 0$. We also illustrate the advantages of β-SVGD over SVGD by experiments.

1 Introduction

The main technical task of Bayesian inference is to estimate integration with respect to the posterior distribution $\pi(x) \propto e^{-V(x)}$, where $V : \mathbb{R}^d \to \mathbb{R}$ is a potential. In practice, this is often reduced to sampling points from the distribution π. Typical methods that employ this strategy include algorithms based on Markov Chain Monte Carlo (MCMC), such as Hamiltonian Monte Carlo Neal [2011], also known as Hybrid Monte Carlo (HMC) Duane et al. [1987], Betancourt [2017], and algorithms based on Langevin dynamics Durmus and Moulines [2017], Garbuno-Inigo et al. [2020], Li and Ying [2019].

On the other hand, Stein Variational Gradient Descent (SVGD)—a different strategy suggested by Liu and Wang [2016]—is based on an interacting particle system. In the population limit, the interacting particle system can be seen as the kernelized negative gradient flow of the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(\rho \mid \pi) := \int \log\left(\tfrac{\rho}{\pi}\right)(x)\, d\rho(x); \tag{1}$$

see Liu [2017], Duncan et al. [2019]. SVGD has already been widely used in a variety of machine learning settings, including variational auto-encoders Pu et al. [2017], reinforcement learning Liu et al. [2017], sequential decision making Zhang et al. [2018, 2019], generative adversarial networks Tao et al. [2019] and federated learning Kassab and Simeone [2022]. However, current theoretical understanding of SVGD is limited to its infinite particle version Liu [2017], Korba et al. [2020], Salim et al. [2021], Sun et al. [2022], and the theory on finite particle SVGD Shi and Mackey [2022] is far from satisfactory.

Since SVGD is built on a discretization of the kernelized negative gradient flow of (1), we can learn about its sampling potential by studying this flow. In fact, a direct calculation reveals that

$$\min_{0 \le s \le t} I_{Stein}(\rho_s \mid \pi) \le \tfrac{D_{\mathrm{KL}}(\rho_0 \mid \pi)}{t}, \tag{2}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

where $I_{Stein}(\rho_s \mid \pi)$ is the norm of the kernelized Wasserstein gradient of the KL-divergence (see Definition 2.2), which is typically used to quantify how close to π are the probability distributions $(\rho_s)_{s=0}^t$ generated along this flow. In particular, if our goal is to guarantee $\min_{0\le s\le t} I_{Stein}(\rho_s \mid \pi) \le \varepsilon$, result (2) says that we need to take

$$t \ge \frac{D_{\mathrm{KL}}(\rho_0|\pi)}{\varepsilon}.$$

Unfortunately, and this is the key motivation for our work, the quantity the initial KL divergence $D_{\mathrm{KL}}(\rho_0 \mid \pi)$ can be very large. Indeed, it can be proportional to the underlying dimension, which is highly problematic in high dimensional regimes. Salim et al. [2021] and Sun et al. [2022] have recently derived an iteration complexity bound for the infinite particle SVGD method. However, similarly to the time complexity of the continuous flow, their bound depends on $D_{\mathrm{KL}}(\rho_0 \mid \pi)$.

1.1 Summary of contributions

In this paper, we design a family of continuous time flows—which we call β-SVGD flow—by combining *importance weights* with the kernelized gradient flow of the KL-divergence. Surprisingly, we prove that the time for this flow to converge to the equilibrium distribution π, that is $\min_{0\le s\le t} I_{Stein}(\rho_s \mid \pi) \le \varepsilon$ with $(\rho_s)_{s=0}^t$ generated along β-SVGD flow, can be bounded by $-\frac{1}{\varepsilon\beta(\beta+1)}$ when $\beta \in (-1, 0)$. This indicates that the importance weights can potentially accelerate SVGD. Actually, we design β-SVGD method based on a discretization of the β-SVGD flow and provide a descent lemma for its population limit version. Some experiments verify our predictions.

We summarize our contributions in the following:

- **A new family of flows.** We construct a family of continuous time flows for which we coin the name β-SVGD flows. These flows do *not* arise from a time re-parameterization of the SVGD flow since their trajectories are different, nor can they be seen as the kernelized gradient flows of the Rényi divergence.
- **Convergence rates.** When $\beta \to 0$, this returns back to the kernelized gradient flow of the KL-divergence (SVGD flow); when $\beta \in (-1, 0)$, the convergence rate of β-SVGD flows is significantly improved than that of the SVGD flow in the case $D_{\mathrm{KL}}(\rho_0 \mid \pi)$ is large. Under a Stein Poincaré inequality, we derive an exponential convergence rate of 2-Rényi divergence along 1-SVGD flow.
- **Algorithm.** We design β-SVGD algorithm based on a discretization of the β-SVGD flow and we derive a descent lemmas for the population limit β-SVGD.
- **Experiments.** Finally, we do some experiments (due to page limit, some of the experiments have been deferred to the appendix) to illustrate the advantages of β-SVGD with negative β. The simulation results on β-SVGD corroborate our theory.

1.2 Related works

The SVGD sampling technique was first presented in the fundamental work of Liu and Wang [2016]. Since then, a number of SVGD variations have been put out. The following is a partial list: Newton version SVGD Detommaso et al. [2018], stochastic SVGD Gorham et al. [2020], mirrored SVGD Shi et al. [2021], random-batch method SVGD Li et al. [2020] and matrix kernel SVGD Wang et al. [2019]. The theoretical knowledge of SVGD is still constrained to population limit SVGD. The first work to demonstrate the convergence of SVGD in the population limit was by Liu [2017], Korba et al. [2020] then derived a similar descent lemma for the population limit SVGD using a different approach. However, their results relied on the path information and thus were not self-contained, to provide a clean analysis, Salim et al. [2021] assumed a Talagrand's T_1 inequality of the target distribution π and gave the first iteration complexity analysis in terms of dimension d. Following the work of Salim et al. [2021], Sun et al. [2022] derived a descent lemma for the population limit SVGD under a non-smooth potential V.

2 Preliminaries

The target distribution is of the form $\pi(x) \propto e^{-V(x)}$. We will use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ to denote the inner product and Euclidean norm on $\mathbb{R}^d$ separately. We will use $\|\cdot\|_{op}$ and $\|\cdot\|_F$ to denote the operator

norm and Frobenius norm of a square matrix. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures with finite second moment, the Wasserstein 2-distance between $\rho, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined by

$$W_2(\rho, \mu) := \inf_{\eta \in \Gamma(\rho,\mu)} \sqrt{\int \|x - y\|^2 \, d\eta(x, y)},$$

where $\Gamma(\rho, \mu)$ is the coupling of ρ and μ. $T_{\#}\rho$ will be used to denote the push-forward distribution of ρ under the map $T : \mathbb{R}^d \to \mathbb{R}^d$.

2.1 Rényi divergence

Rényi divergence is a generalization of the Kullback-Leibler divergence.

Definition 2.1 (Rényi divergence)**.** *For two probability distributions ρ and μ on $\mathbb{R}^d$ and $\rho \ll \mu$, the Rényi divergence of positive order α is defined as*

$$\mathrm{D}_\alpha(\rho \mid \mu) := \tfrac{1}{\alpha-1} \log \left(\int \left(\tfrac{\rho}{\mu}\right)^{\alpha-1} (x) \, d\rho(x) \right), \tag{3}$$

if $\alpha \in (0, \infty)$ and $\alpha \neq 1$, if $\alpha = 1$,

$$D_{\mathrm{KL}}(\rho \mid \mu) = \mathrm{D}_\alpha(\rho \mid \mu) \mid_{\alpha=1} := \int \log \left(\tfrac{\rho}{\mu}\right)(x) \, d\rho(x). \tag{4}$$

If ρ is not absolutely continuous with respect to μ, we set $\mathrm{D}_\alpha(\rho \mid \mu) = \infty$.

Rényi divergence is non-negative, continuous and non-decreasing in terms of the parameter α; specifically, we have $D_{\mathrm{KL}}(\rho \mid \mu) = \lim_{\alpha \to 1} \mathrm{D}_\alpha(\rho \mid \mu)$. More properties of Rényi divergence can be found in a comprehensive article by Van Erven and Harremos [2014].

2.2 Background on SVGD

Stein Variational Gradient Descent (SVGD) is defined on a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_0$ with a non-negative definite reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$. The key feature of this space is its reproducing property:

$$f(x) = \langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}_0}, \qquad \forall f \in \mathcal{H}_0, \tag{5}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is the inner product defined on $\mathcal{H}_0$. Let $\mathcal{H}$ be the d-fold Cartesian product of $\mathcal{H}_0$. That is, $f \in \mathcal{H}$ if and only if there exist $f_1, \cdots, f_d \in \mathcal{H}_0$ such that $f = (f_1, \ldots, f_d)^\top$. Naturally, the inner product on $\mathcal{H}$ is given by

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^{d} \langle f_i, g_i \rangle_{\mathcal{H}_0}, \tag{6}$$

where $f = (f_1, \ldots, f_d)^\top \in \mathcal{H}$ and $g = (g_1, \ldots, g_d)^\top \in \mathcal{H}$. For more details of RKHS, the readers can refer to Berlinet and Thomas-Agnan [2011].

It is well known (see for example Ambrosio et al. [2005]) that $\nabla \log \left(\frac{\rho}{\pi}\right)$ is the Wasserstein gradient of $D_{\mathrm{KL}}(\cdot \mid \pi)$ at $\rho \in \mathcal{P}_2(\mathbb{R}^d)$. Liu and Wang [2016] proposed a kernelized Wasserstein gradient of the KL-divergence, defined by

$$g_\rho(x) := \int k(x, y) \nabla \log \left(\tfrac{\rho}{\pi}\right)(y) \, d\rho(y) \in \mathcal{H}. \tag{7}$$

Integration by parts yields

$$g_\rho(x) = -\int \left[\nabla \log \pi(y) k(x, y) + \nabla_y k(x, y)\right] d\rho(y). \tag{8}$$

Comparing the Wasserstein gradient $\nabla \log \left(\frac{\rho}{\pi}\right)$ with (8), we find that the latter can be easily approximated by

$$g_\rho(x) \approx \hat{g}_{\hat{\rho}} := -\tfrac{1}{N} \sum_{i=1}^{N} \left[\nabla \log \pi(x_i) k(x, x_i) + \nabla_{x_i} k(x, x_i)\right], \tag{9}$$

with $\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i}$ and $(x_i)_{i=1}^N$ sampled from ρ. With the above notations, the SVGD update rule

$$x_i \leftarrow x_i + \frac{\gamma}{N} \sum_{j=1}^{N} \left[\nabla \log \pi(x_j) k(x_i, x_j) + \nabla_{x_j} k(x_i, x_j)\right], \tag{10}$$

where $i = 1, \ldots, N$ and γ is the step-size, can be presented in the compact form $\hat{\rho} \leftarrow (I - \gamma \hat{g}_{\hat{\rho}})_{\#} \hat{\rho}$. When we talk about the infinite particle SVGD, or population limit SVGD, we mean $\rho \leftarrow (I - \gamma g_{\rho})_{\#} \rho$. The metric used in the study of SVGD is the Stein Fisher information or the Kernelized Stein Discrepancy (KSD).

Definition 2.2 (Stein Fisher Information). *Let $\rho \in \mathcal{P}_2(\mathbb{R}^d)$. The Stein Fisher Information of ρ relative to π is defined by*

$$I_{Stein}(\rho \mid \pi) := \iint k(x,y) \left\langle \nabla \log\left(\frac{\rho}{\pi}\right)(x), \nabla \log\left(\frac{\rho}{\pi}\right)(y) \right\rangle \, d\rho(x) \, d\rho(y). \tag{11}$$

A sufficient condition under which $\lim_{n\to\infty} I_{Stein}(\rho_n \mid \pi)$ implies $\rho_n \to \pi$ weakly can be found in Gorham and Mackey [2017], which requires: i) the kernel k to be in the form $k(x,y) = \left(c^2 + \|x-y\|^2\right)^{\theta}$ for some $c > 0$ and $\theta \in (-1, 0)$; ii) $\pi \propto e^{-V}$ to be distant dissipative; roughly speaking, this requires V to be convex outside a compact set, see Gorham and Mackey [2017] for an accurate definition. In the study of the kernelized Wasserstein gradient (8) and its corresponding continuity equation, Duncan et al. [2019] introduced the following kernelized log-Sobolev inequality to prove the exponential convergence of $D_{\mathrm{KL}}(\rho_t \mid \pi)$ along the direction (8):

Definition 2.3 (Stein log-Sobolev inequality). *We say π satisfies the Stein log-Sobolev inequality with constant $\lambda > 0$ if*

$$D_{\mathrm{KL}}(\rho \mid \pi) \le \tfrac{1}{2\lambda} I_{Stein}(\rho \mid \pi). \tag{12}$$

While this inequality can guarantee an exponential convergence rate of ρ_t to π, quantified by the KL-divergence, the condition for π to satisfy the Stein log-Sobolev inequality is very restrictive. In fact, little is known about when (12) holds.

3 Continuous time dynamics of the β-SVGD flow

In this section, we mainly focus on the continuous time dynamics of the β-SVGD flow. Due to page limitation, we leave all of the proofs to Section 9.

3.1 β-SVGD flow

In this paper, a *flow* refers to some time-dependent vector field $v_t : \mathbb{R}^d \to \mathbb{R}^d$. This time-dependent vector field will influence the mass distribution on $\mathbb{R}^d$ by the continuity equation

$$\frac{\partial \rho_t}{\partial t} + \operatorname{div}(\rho_t v_t) = 0, \tag{13}$$

readers can refer to Ambrosio et al. [2005] for more details.

Definition 3.1 (β-SVGD flow). *Given a weight parameter $\beta \in \mathbb{R}$, the β-SVGD flow is given by*

$$v_t^{\beta}(x) := -\left(\frac{\pi}{\rho_t}\right)^{\beta}(x) \int k(x,y) \nabla \log\left(\frac{\rho_t}{\pi}\right)(y) \, d\rho_t(y). \tag{14}$$

Note that when $\beta = 0$, this is the negative kernelized Wasserstein gradient (7).

Note that we can not treat β-SVGD flow as the kernelized Wasserstein gradient flow of the $(\beta+1)$-Rényi divergence. However, they are closely related, and we can derive the following theorem.

Theorem 3.2 (Main result). *Along the β-SVGD flow (14), we have*

$$\min_{t\in[0,T]} I_{Stein}(\rho_t \mid \pi) \le \frac{1}{T}\int_0^T I_{Stein}(\rho_t \mid \pi) dt \le \begin{cases} \frac{e^{\beta \mathrm{D}_{\beta+1}(\rho_0 \mid \pi)}}{T\beta(\beta+1)} & \beta > 0 \\ \frac{D_{\mathrm{KL}}(\rho_0 \mid \pi)}{T} & \beta = 0 \\ -\frac{1}{T\beta(\beta+1)} & \beta \in (-1, 0) \\ \frac{D_{\mathrm{KL}}(\pi \mid \rho_0)}{T} & \beta = -1 \\ \frac{e^{(-\beta-1)\mathrm{D}_{-\beta}(\pi \mid \rho_0)}}{|T\beta(\beta+1)|} & \beta < -1 \end{cases}. \tag{15}$$

Proof. The proof relies on a straightforward calculation that involves considering the definitions of Rényi divergence and Stein-Fisher Information. The detailed calculations and derivations will be provided in the appendix. □

Interestingly, the right hand side of (15) is kind of symmetric around $\beta = -\frac{1}{2}$ and attains its minimum $\frac{4}{T}$ at $\beta = -\frac{1}{2}$, if ρ_0 differs from π a lot. As illustrated by Example 3.3, we generally have the right hand side of (15) depends on d at least linearly when $\beta \notin (-1, 0)$.

It is somewhat unexpected to observe that the time complexity is independent of ρ_0 and π, or to be more precise, that it relies only very weakly on ρ_0 and π when $\beta \in (-1, 0)$. We wish to stress that this is *not* achieved by time re-parameterization. When $\beta \in (-1, 0)$, term $\left(\frac{\pi}{\rho_t}\right)^\beta$ in β-SVGD has an added advantage and can be seen as the acceleration and stabilization factor in front of the kernelized Wasserstein gradient of KL-divergence. Specifically, the negative kernelized Wasserstein gradient of KL-divergence $v_t^0(x)$ is the vector field that compels ρ_t to approach π, while $\left(\frac{\pi}{\rho_t}\right)^\beta(x)$ is big (roughly speaking this means x is close to the mass concentration region of ρ_t but away from the one of π), this factor will enhance the vector field at point x and force the mass around x move faster towards the mass concentration region of π; on the other hand, if $\left(\frac{\pi}{\rho_t}\right)^\beta(x)$ is small (this means x is already near to the mass concentration region of π), this factor will weaken the vector field and make the mass surrounding x stable and remain within the mass concentration region of π. This is the intuitive justification for why, when $\beta \in (-1, 0)$, the time complexity for β-SVGD flow to diminish the Stein Fisher information only depends on ρ_0 and π very weakly.

Example 3.3. *Let $\rho_0 = \mathcal{N}(0, I_d)$ and $\pi = \mathcal{N}\left(0, \frac{1}{2} I_d\right)$, then it can be calculated that $\mathrm{D}_\alpha(\rho_0 \mid \pi) \geq D_{\mathrm{KL}}(\rho_0 \mid \pi) = \frac{d}{2}\log(\frac{e}{2})$ and $\mathrm{D}_\alpha(\pi \mid \rho_0) \geq D_{\mathrm{KL}}(\pi \mid \rho_0) = \frac{d}{4}\log(\frac{4}{e})$, where $\alpha \geq 1$.*

3.2 Exponential convergence of 1-SVGD flow under the Stein Poincaré inequality

In this section, we study the 1-SVGD flow

$$v_t^1(x) := -\frac{\pi}{\rho_t}(x) \int k(x, y) \nabla \log\left(\frac{\rho_t}{\pi}\right)(y)\, d\rho_t(y), \tag{16}$$

which can be seen as the negative kernelized Wasserstein gradient flow of $D_{\mathrm{KL}}(\pi \mid \cdot)$, which is $\frac{\pi}{\rho}\nabla \log(\frac{\rho}{\pi})$. We will show that under the Stein Poincaré inequality, the 1-SVGD flow will decrese the 2-Rényi divergence exponentially fast.

Definition 3.4 (Stein Poincaré inequality)**.** *We say that π satisfies the Stein Poincaré inequality with constant $\lambda > 0$ if*

$$\int |g|^2\, d\pi \leq \frac{1}{\lambda} \iint k(x, y) \langle \nabla g(x), \nabla g(y) \rangle\, d\pi(x)\, d\pi(y), \tag{17}$$

for any smooth g with $\int g\, d\pi = 0$.

Just as Poincaré inequality is a linearized log-Sobolev inequality (see for example [Bakry et al., 2014, Proposition 5.1.3]), Stein Poincaré inequality is also a linearized Stein log-Sobolev inequality, see Section 8. Although Stein Poincaré inequality is weaker than Stein log-Sobolev inequality, the condition for it to hold is quite restrictive, as in the case of Stein log-Sobolev inequality, see the discussion in [Duncan et al., 2019, Section 6]. The following theorem is inspired by Cao et al. [2019], in which they proved the exponential convergence of Rényi divergence along Langevin dynamic under a strongly convex potential V.

Theorem 3.5. *Suppose π satisfies the Stein Poincaré inequality with constant $\lambda > 0$. Then the flow (16) satisfies*

$$\mathrm{D}_2(\rho_t \mid \pi) \leq C \cdot \mathrm{D}_2(\rho_0 \mid \pi) \cdot e^{-2\lambda t}, \tag{18}$$

where $C = \frac{e^{\mathrm{D}_2(\rho_0 \mid \pi)} - 1}{\mathrm{D}_2(\rho_0 \mid \pi)}$.

Proof. We only provide a sketch here, for more detail, please refer to the appendix. The proof is based on a direct calculation of $\frac{d}{dt}\mathrm{D}_2(\rho_t \mid \pi)$, then combining Stein Poincaré inequality, we can give an upper bound to $\frac{d}{dt}\mathrm{D}_2(\rho_t \mid \pi)$, finally by a differential inequality, we finish the proof. □

Since $\mathrm{D}_{\alpha_1}(\rho \mid \pi) \leq \mathrm{D}_{\alpha_2}(\rho \mid \pi)$ for any $0 < \alpha_1 \leq \alpha_2 < \infty$, the exponential convergence of α-Rényi divergence with $\alpha \in (0, 2)$ can be easily deduced from (18).

Corollary 3.6. *Suppose π satisfies the Stein Poincaré inequality with constant $\lambda > 0$. Then the flow (16) satisfies*

$$\mathrm{D}_\alpha(\rho_t \mid \pi) \leq C \cdot \mathrm{D}_\alpha(\rho_0 \mid \pi) \cdot e^{-2\lambda t} \tag{19}$$

for all $\alpha \in (0, 2]$, where $C = \frac{e^{\mathrm{D}_2(\rho_0 \mid \pi)} - 1}{\mathrm{D}_\alpha(\rho_0 \mid \pi)}$.

4 The β-SVGD algorithm

The β-SVGD algorithm[1] proposed here is a sampling method suggested by the discretization of the β-SVGD flow (14). Our method reverts to the traditional SVGD algorithm when $\beta = 0$.

As in SVGD, the integral term in the β-SVGD flow (14) can be approximated by (9). However, when $\beta \neq 0$, we have to estimate the extra importance weight term $\left(\frac{\pi}{\rho_t}\right)^\beta$. Due to the lack of the normalization constant of π and the curse of dimension, we can hardly to use the kernel density estimation Silverman [2018] to approximate $\frac{\pi}{\rho_t}$ accurately in high dimension. Here, we use a different approach to approximate $\frac{\pi}{\rho_t}$, known as the Stein importance weight Liu and Lee [2017]. As noted in Liu and Lee [2017], *the Stein importance weight will concentrate around the true value of* $\frac{\pi}{\rho_t}$ and its calculation does not rely on the normalization constant of π and can be scaled to high dimension. Given N points $(x_i)_{i=1}^N$ sampled from ρ_t, a non-negative definite reproducing kernel k (can be different from the one in β-SVGD) and the score function $\nabla \log(\pi) = -\nabla V$, the Stein importance weight $\hat{\boldsymbol{w}} \in \mathbb{R}_+^d$ is the solution of the following constrained quadratic optimization problem:

$$\arg\min_{\boldsymbol{w}} \left\{ \frac{1}{2} \boldsymbol{w}^\top \boldsymbol{K}_\pi \boldsymbol{w}, \quad \text{s.t.} \quad \sum_{i=1}^N w_i = 1, \quad w_i \geq 0 \right\}, \tag{20}$$

where matrix $\boldsymbol{K}_\pi := \{k_\pi(x_i, x_j)\}_{i,j=1}^N$ with entry

$$\begin{aligned} k_\pi(x_i, x_j) := k(x, y) \left\langle \nabla V(x_i), \nabla V(x_j) \right\rangle - \left\langle \nabla V(x_i), \nabla_{x_j} k(x_i, x_j) \right\rangle \\ - \left\langle \nabla V(x_j), \nabla_{x_i} k(x_i, x_j) \right\rangle + \operatorname{tr}\left(\nabla_{x_i} \nabla_{x_j} k(x_i, x_j) \right). \end{aligned} \tag{21}$$

Stein matrix $\boldsymbol{K}_\pi$ can be efficiently constructed using simple matrix operation, since $\{\nabla V(x_i)\}_{i=1}^N$ have already been computed in the SVGD update (9). It can be proved that as $N \to +\infty$, $N\hat{\boldsymbol{w}}$ will approximate $\left(\frac{\pi}{\rho_t}\right)$, see [Liu and Lee, 2017, Theorem 2.5., Theorem 3.2.]. Problem (20) can be solved efficiently by mirror descent with step-size r:

$$\omega_i^{s+1} = \frac{\omega_i^s e^{-r \sum_{j=1}^N k_\pi(x_i, x_j) \omega_j^s}}{\sum_{l=1}^n \omega_l^s e^{-r \sum_{j=1}^N k_\pi(x_l, x_j) \omega_j^s}}, \quad i = 1, 2, \ldots, N. \tag{22}$$

With matrix $\boldsymbol{K}_\pi$, the computation cost of mirror descent to find the optimum with ε-accuracy is $O(\frac{N^2}{\varepsilon})$, which is independent of dimension d. In general, N cannot be too large because the cost of one iteration of SVGD is $O(N^2 d)$, which quadratically depends on N.

Remark 4.1. *The kernel used to calculate the Stein importance weight does not have to be the same one as used in the SVGD update. However, in this paper, we set them the same for simplicity.*

Remark 4.2. *In Algorithm 1, we replace* $\left(\frac{\pi}{\rho_t}\right)^\beta (x_i)$ *by* $\left(\max\left(N\hat{\boldsymbol{w}}_i, \tau\right)\right)^\beta$*, here* τ *is a small positive number to separate* $N\hat{\boldsymbol{w}}_i$ *from* 0*. The benefits of* $(N\hat{\boldsymbol{w}})^\beta$*, just like the benefits of* $\left(\frac{\pi}{\rho_t}\right)^\beta$ *as explained in Section 3.1, are twofold: it accelerates points with small weights and stabilizes points with big weights.*

4.1 Descent property of the population limit β-SVGD

In this section, we study the convergence of the population limit β-SVGD, that is

$$x_{n+1} = x_n - \gamma \left[\left(\frac{\pi}{\rho_n} \right)^\beta (x_n) \wedge M \right] \int k(x_n, y) \nabla \log\left(\frac{\rho_n}{\pi} \right)(y) d\rho_n(y), \tag{23}$$

where $\left(\frac{\pi}{\rho_n}\right)^\beta (x_n) \wedge M = \lim_{N\to\infty} \left(\max\left(N\omega_i, \tau\right)\right)^\beta$ and $M := \frac{1}{\tau^\beta}$.

Specifically, we establish a descent lemma for it. The derivation of the descent lemma is based on the following assumptions. The first two assumptions are typically used in the analysis of SVGD related algorithms, seeLiu [2017].

[1]For simplicity, we will often just call it β-SVGD; not to be confused with the β-SVGD flow.

Algorithm 1 β-Stein Variational Gradient Descent (β-SVGD)

1: **Input:** A set of initial particles $(x_i^0)_{i=1}^N$, initial importance weight $\omega_i = 1/N$, iteration number n and step-size γ for β-SVGD update, iteration number m and step-size r for mirror descent update.
2: **for** $l = 0, 1, \ldots, n$ **do**
3: Update ω_i by mirror descent with iteration number m and step-size r.
4: Update particles with step-size γ and small gap τ:

$$x_i^{l+1} \leftarrow x_i^l + \gamma \left(\max\left(N\omega_i, \tau\right)\right)^\beta \times \sum_{j=1}^N \left[-k(x_i^l, x_j^l)\nabla_{x_j^l} V(x_j^l) + \nabla_{x_j^l} k(x_i^l, x_j^l)\right], \quad i = 1, \ldots, N$$

5: **end for**
6: **Return:** Particles $(x_i^{n+1})_{i=1}^N$.

Assumption 4.3. *The potential function V of the target distribution $\pi \propto e^{-V}$ is L-smooth, that is, $\left\|\nabla^2 V\right\|_{op} \leq L$.*

Assumption 4.4. *Kernel k is continuously differentiable and there exists $B > 0$ such that $\|k(x,.)\|_{\mathcal{H}_0} \leq B$ and $\|\nabla_x k(x,.)\|_{\mathcal{H}}^2 = \sum_{i=1}^d \|\partial_{x_i} k\,(x,.)\|_{\mathcal{H}_0}^2 \leq B^2, \qquad \forall x \in \mathbb{R}^d$.*

The next assumption depends on ρ_n and is used to control the extra term $\left(\frac{\pi}{\rho_n}\right)^\beta \wedge M$ in (23).

Assumption 4.5. *$f_n := \log\left(\frac{\pi}{\rho_n}\right) \in C^1(\mathbb{R}^d)$ and there is some constant $C_b \geq 0$, such that $\log(\|\nabla f_n(x)\|) \leq \frac{|\beta f_n(x)|}{2} + C_b, \quad \forall x \in \mathbb{R}^d$.*

The above regularity assumption on f_n is very weak; for example, it is satisfied by, but not limited to, any C^1 polynomial or any C^1 function that is not far from a polynomial in the C^1 norm.

Under the above assumptions, we have the following descent lemma.

Proposition 4.6 (Descent Lemma)**.** *Let $\beta \in (-1, 0)$, suppose Assumption 4.3, 4.4 and 4.5 hold. For any small δ, if $2\delta \leq I_{Stein}\left(\rho_n \mid \pi\right) < \infty$, we can choose M big enough depends on ρ_n, π, δ, and γ satisfies*

$$\begin{cases} 0 < \gamma \leq \frac{1}{6\left(-\beta M^{\frac{3}{2}} e^{C_b} + M\right) B I_{Stein}(\rho_n|\pi)^{\frac{1}{2}}} \\ 0 < \gamma \leq \frac{2(\beta+1)(I_{Stein}(\rho_n|\pi) - \delta)}{B^2 I_{Stein}(\rho_n|\pi)\left(LM^2 + 10(-\beta M^{\frac{3}{2}} e^{C_b} + M)^2\right)} \\ 0 < \gamma \leq \frac{\beta+1}{B^2\left(LM^2 + 10(-\beta M^{\frac{3}{2}} e^{C_b} + M)^2\right)} \end{cases}, \tag{24}$$

then we have the descent property

$$e^{\beta \mathrm{D}_{\beta+1}(\rho_{n+1}|\pi)} - e^{\beta \mathrm{D}_{\beta+1}(\rho_n|\pi)} \geq -\beta(\beta+1)\gamma\left(\frac{1}{2} I_{Stein}\left(\rho_n \mid \pi\right) - \delta\right). \tag{25}$$

Proof. The full proof is in the appendix, here we only provide a sketch. We need first to upper bound two terms: $\mathcal{I} := \log(\pi)(x) - \log(\pi)(\phi_n(x))$ and $\mathcal{II} := -\log(|\det \mathrm{D}\,\phi_n|)(x)$, where map ϕ_n is defined by Equation (23); with the bounds and Jensen inequality, we can give a lower bound to the left hand side of Equation (25); in the last, we analyze the condition γ should satisfy. □

The proof and the choice of M can be found in Section 9. Let β and δ approach 0, then we can recover the descent lemma for the population limit SVGD Liu [2017]. Salim et al. [2021] and Sun et al. [2022] analyzed the population limit SVGD under a Talagrand type inequality, however, in our case, we do not have such kind inequality for the Rényi divergence. The theoretical analysis of the finite particle SVGD and its variants is still widely open, for example, the existence and uniqueness of the stationary distribution of the finite particle SVGD are not known yet, see Liu and Wang [2018]. Shi and Mackey [2022] analyzed the finite particle SVGD, however, their bound is pessimistic. To find a reasonable analysis for the finite particle SVGD and β-SVGD is challenging and we leave this question for the future study.

5 Experiments

In this section, we use some experiments to show the benefits of the importance weights in the update of SVGD. The code can be found through `https://github.com/Iwillnottellyou/BETA-SVGD.git`. For more experiments on Bayesian Neural Network and Bayesian Logistic Regression, please refer to the appendix. In all the experiments, we choose $k(x,y) = e^{-\frac{\|x-y\|^2}{d}}$, and due to the page limitation, we will leave part of the experiments details to Section 11.

Gaussian Mixture: In this experiment, we use the obtained particles to estimate expectation $\mathbb{E}_\pi[h]$ with different test functions $h(\cdot)$. In Figure 1, we can see the clear improvement of β-SVGD over SVGD, for more test results on Gaussian Mixture, see Section 11.

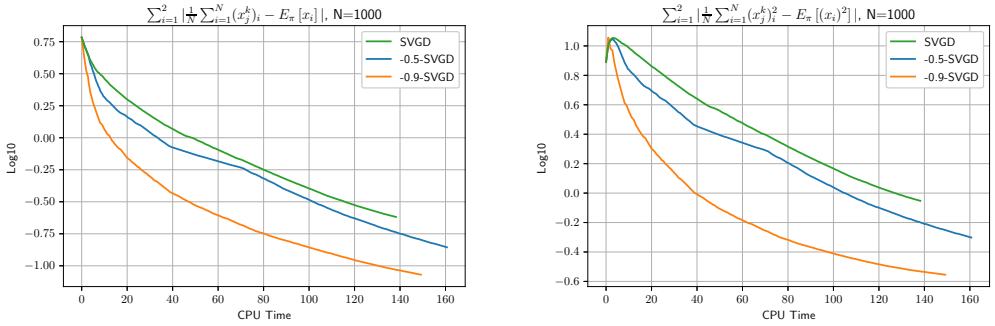

Figure 1: Gaussian Mixture with dimension $d = 2$. The target distribution is $\pi(x) = \frac{2}{5}\mathcal{N}((2,0), I_2) + \frac{1}{5}\mathcal{N}((4,0), I_2) + \frac{2}{5}\mathcal{N}((3,-3), I_2)$. Each sampled point x_j^k is of the form $\left((x_j^k)_1, (x_j^k)_2\right)$, where k denote the k-th iteration, j denote the j-th sampled point. For distribution π, we have $\mathbb{E}_\pi[x_1] = 2.8, \mathbb{E}_\pi[x_2] = -1.2$ and $\mathbb{E}_\pi\left[(x_1)^2\right] = 9.4, \mathbb{E}_\pi\left[(x_2)^2\right] = 4.6$. The initial N points are sampled from $\mathcal{N}((-2,0), I_2)$. The step-size γ for both algorithms equals 0.2. In β-SVGD, we choose the small gap $\tau = 0.01$ and we update the Stein importance weights every 20 iterations using 40 mirror descent steps with step-size $r = 0.3$. Since the function computed in the second image is x^2, it is not surprising that there is an increase in the first few iterations.

6 Conclusion

In this paper, we study how the importance weights can influence the SVGD, our theoretical analysis and the experiments reveal some previously unexplored facts. Specifically, we construct a family of continuous time flows called β-SVGD flows on the space of probability distributions, when $\beta \in (-1, 0)$, its convergence rate is independent of the initial distribution and the target distribution. Based on β-SVGD flow, we design a family of weighted SVGD called β-SVGD. β-SVGD has the similar computation complexity as SVGD, and due to the Stein importance weight, it converges faster and is more stable than SVGD in our experiments. β-SVGD generally performs better than SVGD in the iteration of the algorithms, however, due to the extra calculation of the Stein importance weights, its computation cost is higher than SVGD, so one related question is to combine more advanced constrained quadratic optimization methods into β-SVGD to improve the efficiency of the computation of the importance weights. Another related direction is to explore the influence of the importance weights in the update of MCMC algorithms, like the Langevin type methods. We leave all these related questions for the future study.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Yu Cao, Jianfeng Lu, and Yulong Lu. Exponential decay of Rényi divergence under Fokker–Planck equations. *Journal of Statistical Physics*, 176(5):1172–1184, 2019.

Gianluca Detommaso, Tiangang Cui, Youssef Marzouk, Alessio Spantini, and Robert Scheichl. A Stein variational Newton method. *Advances in Neural Information Processing Systems*, 31, 2018.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

Alfredo Garbuno-Inigo, Franca Hoffmann, Wuchen Li, and Andrew M Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM Journal on Applied Dynamical Systems*, 19(1):412–441, 2020.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

Jackson Gorham, Anant Raj, and Lester Mackey. Stochastic Stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942, 2020.

Rahif Kassab and Osvaldo Simeone. Federated generalized Bayesian learning via distributed Stein variational gradient descent. *IEEE Transactions on Signal Processing*, 2022.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33: 4672–4682, 2020.

Lei Li, Yingzhou Li, Jian-Guo Liu, Zibu Liu, and Jianfeng Lu. A stochastic version of Stein variational gradient descent for efficient sampling. *Communications in Applied Mathematics and Computational Science*, 15(1):37–63, 2020.

Wuchen Li and Lexing Ying. Hessian transport gradient flows. *Research in the Mathematical Sciences*, 6(4):1–20, 2019.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in Neural Information Processing Systems*, 30, 2017.

Qiang Liu and Jason Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Qiang Liu and Dilin Wang. Stein variational gradient descent as moment matching. *Advances in Neural Information Processing Systems*, 31, 2018.

Yang Liu, Prajit Ramachandran, Qiang Liu, and Jian Peng. Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*, 2017.

R. M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. VAE learning via Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Adil Salim, Lukang Sun, and Peter Richtárik. Complexity analysis of Stein variational gradient descent under Talagrand's inequality T1. *arXiv preprint arXiv:2106.03076*, 2021.

Jiaxin Shi and Lester Mackey. A finite-particle convergence rate for stein variational gradient descent. *arXiv preprint arXiv:2211.09721*, 2022.

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with mirrored Stein operators. *arXiv preprint arXiv:2106.12506*, 2021.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Lukang Sun, Avetik Karagulyan, and Peter Richtárik. Convergence of Stein variational gradient descent under a weaker smoothness condition. *arXiv preprint arXiv:2206.00508*, 2022.

Chenyang Tao, Shuyang Dai, Liqun Chen, Ke Bai, Junya Chen, Chang Liu, Ruiyi Zhang, Georgiy Bobashev, and Lawrence Carin. Variational annealing of GANs: A Langevin perspective. In *International Conference on Machine Learning*, pages 6176–6185. PMLR, 2019.

Tim Van Erven and Peter Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. *Advances in Neural Information Processing Systems*, 32, 2019.

Ruiyi Zhang, Chunyuan Li, Changyou Chen, and Lawrence Carin. Learning structural weight uncertainty for sequential decision-making. In *International Conference on Artificial Intelligence and Statistics*, pages 1137–1146. PMLR, 2018.

Ruiyi Zhang, Zheng Wen, Changyou Chen, and Lawrence Carin. Scalable Thompson sampling via optimal transport. *arXiv preprint arXiv:1902.07239*, 2019.

Contents

Appendix

7 Calculus

This section is devoted to provide rigorous verification for several claims in the main paper, these results are already known to readers who are familiar with Rényi divergence. We first calculate the Wasserstein gradient flow of Rényi divergence. Let ρ_t satisfies

$$\frac{\partial \rho_t}{\partial t} + \operatorname{div}(\rho_t v_t) = 0,$$

for some vector field v_t on $\mathbb{R}^d$, then when $\alpha \in (0,1) \cup (1,\infty)$, we have

$$\begin{aligned}
\frac{d}{dt} \mathrm{D}_\alpha(\rho_t \mid \pi) &= \frac{d}{dt} \frac{1}{\alpha - 1} \log\left(\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) d\rho_t(x) \right) \\
&= \frac{1}{\alpha-1} \frac{\int \frac{d}{dt}\left(\frac{\rho_t}{\pi}\right)^{\alpha}(x) d\pi(x)}{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) d\rho_t(x)} \\
&= \frac{\alpha}{\alpha-1} \frac{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) \frac{\partial \rho_t}{\partial t}(x) dx}{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) d\rho_t(x)} \\
&= -\frac{\alpha}{\alpha-1} \frac{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) \operatorname{div}(\rho_t v_t)(x) dx}{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) d\rho_t(x)} \\
&= \frac{\alpha}{\alpha-1} \frac{\int \left\langle \nabla \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x), v_t(x) \right\rangle d\rho_t(x)}{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) d\rho_t(x)} \\
&= \alpha \frac{\int \left\langle \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) \nabla \log\left(\frac{\rho_t}{\pi}\right)(x), v_t(x) \right\rangle d\rho_t(x)}{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1}(x) d\rho_t(x)} \\
&= \left\langle \frac{\alpha \left(\frac{\rho_t}{\pi}\right)^{\alpha-1} \nabla \log(\frac{\rho_t}{\pi})}{\int \left(\frac{\rho_t}{\pi}\right)^{\alpha-1} d\rho_t}, v_t \right\rangle_{\rho_t}.
\end{aligned}$$

When $\alpha = 1$, we have

$$\begin{aligned}
\frac{d}{dt} D_{\mathrm{KL}}(\rho_t \mid \pi) &= \frac{d}{dt} \int \log\left(\frac{\rho_t}{\pi}\right)(x) d\rho_t(x) \\
&= \int \frac{d}{dt} \left\{ \frac{\rho_t}{\pi}(x) \log\left(\frac{\rho_t}{\pi}\right)(x) \right\} d\pi(x) \\
&= \int \left(1 + \log\left(\frac{\rho_t}{\pi}\right)(x)\right) \frac{\partial \rho_t}{\partial t}(x) dx \\
&= -\int \left(1 + \log\left(\frac{\rho_t}{\pi}\right)(x)\right) \operatorname{div}(\rho_t v_t)(x) dx \\
&= \int \left\langle \nabla \log\left(\frac{\rho_t}{\pi}\right)(x), v_t(x) \right\rangle d\rho_t(x) \\
&= \left\langle \nabla \log\left(\frac{\rho_t}{\pi}\right), v_t \right\rangle_{\rho_t}.
\end{aligned}$$

The Wasserstein gradient of the reverse KL-divergence:

$$\begin{aligned}
\frac{d}{dt} D_{\mathrm{KL}}(\pi \mid \rho_t) &:= \frac{d}{dt} \int \log(\frac{\pi}{\rho_t})(x) d\pi(x) \\
&= -\int \frac{\frac{\partial \rho_t}{\partial t}}{\rho_t}(x) d\pi(x) \\
&= \int \operatorname{div}(\rho_t v_t)(x) \frac{\pi}{\rho_t}(x) dx \\
&= \int \left\langle -\nabla \frac{\pi}{\rho_t}(x), v_t(x) \right\rangle d\rho_t(x) \\
&= \left\langle -\nabla \frac{\pi}{\rho_t}, v_t \right\rangle_{\rho_t},
\end{aligned}$$

so it is $-\nabla \frac{\pi}{\rho_t}$.

Next, we verify that $\mathrm{D}_\alpha(\rho \mid \pi) \geq 0$. For $\alpha > 1$, we have

$$\int \left(\frac{\rho}{\pi}\right)^{\alpha-1}(x)d\rho(x) = \int \left(\frac{\rho}{\pi}\right)^{\alpha}(x)d\pi(x) \geq \left(\int \frac{\rho}{\pi}(x)d\pi(x)\right)^{\alpha} = 1$$

by the convexity of function t^α for $t \geq 0$, so

$$\mathrm{D}_\alpha(\rho \mid \pi) = \frac{1}{\alpha-1}\log\left(\int \left(\frac{\rho}{\pi}\right)^{\alpha-1}(x)d\rho(x)\right) \geq 0.$$

When $\alpha = 1$, by the convexity of function $t\log(t)$ for $t \geq 0$, we also have

$$D_{\mathrm{KL}}(\rho \mid \pi) = \int \log\left(\frac{\rho_t}{\pi}\right)(x)d\rho(x) = \int \frac{\rho_t}{\pi}(x)\log\left(\frac{\rho_t}{\pi}\right)(x)d\pi(x) \geq 0.$$

When $\alpha \in (0,1)$, function t^α for $t \geq 0$ is concave, so we first have

$$\int \left(\frac{\rho}{\pi}\right)^{\alpha-1}(x)d\rho(x) = \int \left(\frac{\rho}{\pi}\right)^{\alpha}(x)d\pi(x) \leq \left(\int \frac{\rho}{\pi}(x)d\pi(x)\right)^{\alpha} = 1,$$

finally

$$\mathrm{D}_\alpha(\rho \mid \pi) = \frac{1}{\alpha-1}\log\left(\int \left(\frac{\rho}{\pi}\right)^{\alpha-1}(x)d\rho(x)\right) \geq 0.$$

8 Stein log-Sobolev inequality implies Stein Poincaré inequality

In this section, we show that Stein Poincaré inequality is weaker than Stein log-Sobolev inequality.

Lemma 8.1 (Stein log-Sobolev implies Stein Poincaré)**.** *If π satisfies the Stein log-Sobolev inequality (12) with constant $\lambda > 0$, then it also satisfies the Stein Poincaré inequality with the same constant λ.*

Proof. Let g be bounded and $\int g d\pi = 0$. Let ϵ be small enough such that $1 + \epsilon g \geq 0$, so $\rho := \pi(1+\epsilon g)$ is a probability distribution and $\rho \ll \pi$. We need first calculate $D_{\mathrm{KL}}(\rho \mid \pi)$.

$$\begin{aligned} D_{\mathrm{KL}}(\rho \mid \pi) &= \int \log(\frac{(1+\epsilon g)\pi}{\pi})(x)(1+\epsilon g)(x)d\pi(x) \\ &= \int (1+\epsilon g)(x)\log(1+\epsilon g)(x)d\pi(x) \\ &= \int (1+\epsilon g)(x)\left(\epsilon g(x) - \frac{1}{2}\epsilon^2|g|^2(x)\right)d\pi(x) + o(\epsilon^2) \\ &= \frac{1}{2}\epsilon^2\int |g|^2(x)d\pi(x) + o(\epsilon^2), \end{aligned} \tag{26}$$

in the last step, we used $\int g d\pi = 0$. Now we calculate the right hand side of 12,

$$\begin{aligned} I_{Stein}(\rho \mid \pi) &= \iint k(x,y)\left\langle \nabla\log(\frac{\rho}{\pi})(x), \nabla\log(\frac{\rho}{\pi})(y)\right\rangle d\rho(x)d\rho(y) \\ &= \iint k(x,y)\left\langle \nabla\frac{\rho}{\pi}(x), \nabla\frac{\rho}{\pi}(y)\right\rangle d\pi(x)d\pi(y) \\ &= \iint k(x,y)\left\langle \nabla(1+\epsilon g)(x), \nabla(1+\epsilon g)(y)\right\rangle d\pi(x)d\pi(y) \\ &= \epsilon^2\iint k(x,y)\left\langle \nabla g(x), \nabla g(y)\right\rangle d\pi(x)d\pi(y). \end{aligned} \tag{27}$$

Since we have Equation (12), so

$$\frac{1}{2}\epsilon^2\int |g|^2(x)d\pi(x) + o(\epsilon^2) \leq \frac{1}{2\lambda}\epsilon^2\iint k(x,y)\left\langle \nabla g(x), \nabla g(y)\right\rangle d\pi(x)d\pi(y), \tag{28}$$

divide both side by ϵ^2 and let $\epsilon \to 0$, we have Stein Poincaré inequality

$$\int |g|^2 d\pi \leq \frac{1}{\lambda}\iint k(x,y)\left\langle \nabla g(x), \nabla g(y)\right\rangle d\pi(x)d\pi(y). \tag{29}$$

For general unbounded function g with $\int g d\pi = 0$, we can use bounded sequence to approximate it and will also have Stein Poincaré inequality 17 □

9 Missing Proofs

proof of Theorem 3.2. We first discuss the cases when $\beta > -1$. A direct calculation yields

$$\begin{aligned}\frac{d}{dt}\mathrm{D}_{\beta+1}\left(\rho_t \mid \pi\right) &= \left\langle \frac{(\beta+1)\left(\frac{\rho_t}{\pi}\right)^{\beta}\nabla\log(\frac{\rho_t}{\pi})}{\int\left(\frac{\rho_t}{\pi}\right)^{\beta}d\rho_t}, v_t^{\beta}\right\rangle_{\rho_t} \quad \text{//refer to Section 7 for more calculation details}\\ &= -\frac{\beta+1}{\int\left(\frac{\rho_t}{\pi}\right)^{\beta}d\rho_t}\iint k(x,y)\left\langle\nabla\log(\frac{\rho_t}{\pi})(x), \nabla\log(\frac{\rho_t}{\pi})(y)\right\rangle\left(\frac{\rho_t}{\pi}\right)^{\beta}\left(\frac{\pi}{\rho_t}\right)^{\beta}d\rho_t(x)d\rho_t(y)\\ &= -(\beta+1)\frac{\iint k(x,y)\left\langle\nabla\log(\frac{\rho_t}{\pi})(x), \nabla\log(\frac{\rho_t}{\pi})(y)\right\rangle d\rho_t(x)d\rho_t(y)}{\int\left(\frac{\rho_t}{\pi}\right)^{\beta}d\rho_t} \le 0,\end{aligned} \tag{30}$$

which is equivalent to

$$\frac{d}{dt}e^{\beta\mathrm{D}_{\beta+1}(\rho_t|\pi)} = -\beta(\beta+1)I_{Stein}\left(\rho_t \mid \pi\right). \tag{31}$$

Integrate the above equation for t from 0 to T, after rearrangement then we will have

$$\begin{aligned}\min_{t\in[0,T]} I_{Stein}\left(\rho_t \mid \pi\right) &\le \frac{1}{T}\int_0^T I_{Stein}(\rho_t \mid \pi)dt\\ &\le \frac{\left|e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)} - e^{\beta\mathrm{D}_{\beta+1}(\rho_T|\pi)}\right|}{T|\beta(\beta+1)|}.\end{aligned}$$

By (30), we know $\mathrm{D}_{\beta+1}\left(\rho_t \mid \pi\right)$ decreases along β-SVGD flow for any $\beta \in (-1,\infty)$. For $\beta > 0$, we have

$$\frac{\left|e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)} - e^{\beta\mathrm{D}_{\beta+1}(\rho_T|\pi)}\right|}{T|\beta(\beta+1)|} \le \frac{e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)}}{T\beta(\beta+1)}.$$

For $\beta = 0$, we use L'Hopital rule and get

$$\lim_{\beta\to 0}\frac{\left|e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)} - e^{\beta\mathrm{D}_{\beta+1}(\rho_T|\pi)}\right|}{T|\beta(\beta+1)|} = \frac{D_{\mathrm{KL}}(\rho_0 \mid \pi) - D_{\mathrm{KL}}(\rho_T \mid \pi)}{T} \le \frac{D_{\mathrm{KL}}(\rho_0 \mid \pi)}{T}.$$

For $\beta \in (-1,0)$, we have $0 \le e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)} \le e^{\beta\mathrm{D}_{\beta+1}(\rho_T|\pi)} \le 1$, so $\left|e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)} \le e^{\beta\mathrm{D}_{\beta+1}(\rho_T|\pi)}\right| \le 1$ and

$$\frac{\left|e^{\beta\mathrm{D}_{\beta+1}(\rho_0|\pi)} - e^{\beta\mathrm{D}_{\beta+1}(\rho_T|\pi)}\right|}{T|\beta(\beta+1)|} \le -\frac{1}{T\beta(\beta+1)}.$$

When $\beta < -1$, a similar calculation yields

$$\frac{d}{dt}\int\left(\frac{\rho_t}{\pi}\right)^{\beta+1}(x)d\pi(x) = \frac{d}{dt}\int\left(\frac{\pi}{\rho_t}\right)^{-\beta}(x)d\rho_t(x) = -\beta(\beta+1)I_{Stein}\left(\rho_t \mid \pi\right) \le 0,$$

after a rearrangement, we have

$$\begin{aligned}\min_{t\in[0,T]} I_{Stein}\left(\rho_t \mid \pi\right) \le \frac{1}{T}\int_0^T I_{Stein}(\rho_t \mid \pi)dt &\le \frac{\int\left(\frac{\pi}{\rho_0}\right)^{-\beta}(x)d\rho_0(x) - \int\left(\frac{\pi}{\rho_T}\right)^{-\beta}(x)d\rho_T(x)}{|T\beta(\beta+1)|}\\ &\le \frac{\int\left(\frac{\pi}{\rho_0}\right)^{-\beta}(x)d\rho_0(x)}{|T\beta(\beta+1)|}\\ &= \frac{e^{(-\beta-1)\mathrm{D}_{-\beta}(\pi|\rho_0)}}{|T\beta(\beta+1)|}.\end{aligned}$$

The case when $\beta = -1$ can be derived using L'Hopital rule. Combine all these cases, we finish the proof. □

proof of Theorem 3.5. Denoting ${\epsilon_t}^2 = \int\left(\frac{\rho_t-\pi}{\pi}\right)^2 d\pi$, $f_t = \frac{\rho_t-\pi}{\epsilon_t}$, then $\int f_t dx = 0$, $\int\frac{f_t^2}{\pi}dx = 1$, $C_t := \int\left(\frac{\rho_t}{\pi}\right)^2 d\pi = 1 + {\epsilon_t}^2$. Thus

$$
\begin{aligned}
-\frac{d}{dt}\,\mathrm{D}_2\left(\rho_t \mid \pi\right) &= 2\left\langle \nabla \log\left(\frac{\rho_t}{\pi}\right), v_t \right\rangle_{C_t^{-1}\left(\frac{\rho_t}{\pi}\right)^2 \pi} \\
&= \frac{2}{1+{\epsilon_t}^2} \iint \left\langle \nabla \log\left(\frac{\rho_t}{\pi}\right)(y), \nabla \log\left(\frac{\rho_t}{\pi}\right)(x) \right\rangle \left(\frac{\rho_t}{\pi}\right)^{-1}(y) \left(\frac{\rho_t}{\pi}\right)^2 (y) k(x,y) \left(\frac{\rho_t}{\pi}\right)(x) d\pi(x)\pi(y) \\
&= \frac{2}{1+{\epsilon_t}^2} \iint k(x,y) \left\langle \nabla\left(\frac{\rho_t}{\pi}\right)(x), \nabla\left(\frac{\rho_t}{\pi}\right)(y) \right\rangle d\pi(x) d\pi(y) \\
&= \frac{2}{1+{\epsilon_t}^2} \iint k(x,y) \left\langle \nabla\left(\frac{\rho_t}{\pi}-1\right)(x), \nabla\left(\frac{\rho_t}{\pi}-1\right)(y) \right\rangle d\pi(x) d\pi(y) \\
&= \frac{2{\epsilon_t}^2}{1+{\epsilon_t}^2} \iint k(x,y) \left\langle \nabla\left(\frac{f_t}{\pi}\right)(x), \nabla\left(\frac{f_t}{\pi}\right)(y) \right\rangle d\pi(x) d\pi(y).
\end{aligned}
$$

By Stein Poincaré inequality, we have

$$
-\iint k(x,y) \left\langle \nabla\left(\frac{f_t}{\pi}\right)(x), \nabla\left(\frac{f_t}{\pi}\right)(y) \right\rangle d\pi(x) d\pi(y) \le -\lambda \int \left|\frac{f_t}{\pi}\right|^2 (x) d\pi(x),
$$

so finally we have

$$
\begin{aligned}
\frac{d\,\mathrm{D}_2\left(\rho_t \mid \pi\right)}{dt} &= -\frac{2{\epsilon_t}^2}{1+{\epsilon_t}^2} \iint k(x,y) \left\langle \nabla\left(\frac{f_t}{\pi}\right)(x), \nabla\left(\frac{f_t}{\pi}\right)(y) \right\rangle d\pi(x) d\pi(y) \\
&\le -\frac{2{\epsilon_t}^2}{1+{\epsilon_t}^2} \lambda \int \left|\frac{f_t}{\pi}\right|^2 (x) d\pi(x) \\
&= -\frac{2\lambda{\epsilon_t}^2}{1+{\epsilon_t}^2} \\
&= -2\lambda \frac{e^{\mathrm{D}_2(\rho_t|\pi)}-1}{e^{\mathrm{D}_2(\rho_t|\pi)}} \\
&= -2\lambda\left(1-e^{-\mathrm{D}_2(\rho_t|\pi)}\right),
\end{aligned}
$$

which is equivalent to

$$
\frac{d}{dt}\log(e^{\mathrm{D}_2(\rho_t|\pi)}-1) \le -2\lambda.
$$

So

$$
\begin{aligned}
\mathrm{D}_2\left(\rho_t \mid \pi\right) &\le \log\left(1+\left(e^{\mathrm{D}_2(\rho_0|\pi)}-1\right)e^{-2\lambda t}\right) \\
&\le \left(e^{\mathrm{D}_2(\rho_0|\pi)}-1\right)e^{-2\lambda t} \\
&= \frac{e^{\mathrm{D}_2(\rho_0|\pi)}-1}{\mathrm{D}_2\left(\rho_0 \mid \pi\right)} \mathrm{D}_2\left(\rho_0 \mid \pi\right) e^{-2\lambda t}.
\end{aligned}
$$

□

proof of Corollary 3.6. By (18), when $\alpha \in (0,2)$ we have

$$
\begin{aligned}
\mathrm{D}_\alpha\left(\rho_t \mid \pi\right) &\le \mathrm{D}_2\left(\rho_t \mid \pi\right) \\
&\le \frac{e^{\mathrm{D}_2(\rho_0|\pi)}-1}{\mathrm{D}_2\left(\rho_0 \mid \pi\right)} \mathrm{D}_2\left(\rho_0 \mid \pi\right) e^{-2\lambda t} \\
&= \frac{e^{\mathrm{D}_2(\rho_0|\pi)}-1}{\mathrm{D}_2\left(\rho_0 \mid \pi\right)} \frac{\mathrm{D}_2\left(\rho_0 \mid \pi\right)}{\mathrm{D}_\alpha\left(\rho_0 \mid \pi\right)} \mathrm{D}_\alpha\left(\rho_0 \mid \pi\right) e^{-2\lambda t} \\
&= \frac{e^{\mathrm{D}_2(\rho_0|\pi)}-1}{\mathrm{D}_\alpha\left(\rho_0 \mid \pi\right)} \mathrm{D}_\alpha\left(\rho_0 \mid \pi\right) e^{-2\lambda t}.
\end{aligned} \tag{32}
$$

□

proof of Proposition 4.6. First, by monotone convergence theorem, we have

$$
\lim_{M\to+\infty} \iint \left(\frac{\rho_n}{\pi}\right)^\beta (x) \left[\left(\frac{\pi}{\rho_n}\right)^\beta (x) \wedge M\right] k(x,y) \left\langle \nabla \log\left(\frac{\rho_n}{\pi}\right)(x), \nabla \log\left(\frac{\rho_n}{\pi}\right)(y) \right\rangle d\rho_n(x) d\rho_n(y) = I_{Stein}\left(\rho_n \mid \pi\right), \tag{33}
$$

so we can choose M big enough, such that

$$\left| \iint \left(\frac{\rho_n}{\pi}\right)^{\beta}(x) \left[\left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\right] k(x,y) \left\langle \nabla \log\left(\frac{\rho_n}{\pi}\right)(x), \nabla \log\left(\frac{\rho_n}{\pi}\right)(y) \right\rangle d\rho_n(x) d\rho_n(y) - I_{Stein}\left(\rho_n \mid \pi\right) \right| \leq \delta. \tag{34}$$

In the following, we will assume M satisfying the above condition.

Denote $g_n(x) := \left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M \int k(x,y) \nabla \log\left(\frac{\rho_n}{\pi}\right)(y) d\rho_n(y)$, $\phi_n(x) := x - \gamma g_n(x)$ and $\rho_{n+1} = {\phi_n}_{\#}\rho_n$, then we have

$$\begin{aligned} e^{\beta \mathrm{D}_{\beta+1}\left(\rho_{n+1}|\pi\right)} - e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n|\pi\right)} &= e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n|{\phi_n^{-1}}_{\#}\pi\right)} - e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n|\pi\right)} \\ &= \int \left(\frac{\rho_n}{{\phi_n^{-1}}_{\#}\pi}\right)^{\beta}(x) d\rho_n(x) - \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x) d\rho_n(x) \\ &= \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x) \left(\left(\frac{\pi(x)}{{\phi_n^{-1}}_{\#}\pi(x)}\right)^{\beta} - 1\right) d\rho_n(x). \end{aligned} \tag{35}$$

Term $\left(\frac{\pi(x)}{{\phi_n^{-1}}_{\#}\pi(x)}\right)^{\beta}$ can be decomposed into two terms I and II in the following way,

$$\left(\frac{\pi(x)}{{\phi_n^{-1}}_{\#}\pi(x)}\right)^{\beta} = \left(\frac{\pi(x)}{\pi(\phi_n(x))\left|\det \mathrm{D}\,\phi_n\right|(x)}\right)^{\beta} = \exp\left(\beta\Big(\underbrace{\log(\pi)(x) - \log(\pi)(\phi_n(x))}_{\mathcal{I}} \underbrace{- \log(|\det \mathrm{D}\,\phi_n|)(x)}_{\mathcal{II}}\Big)\right), \tag{36}$$

so the next step is to upper bound term I and II separately. For term I, we have that

$$\begin{aligned} \mathcal{I} &= \log(\pi)(x) - \log(\pi)(\phi_n(x)) \\ &= V(x) - V(x - \gamma g_n(x)) \\ &= \gamma \left\langle \nabla V(x), g_n(x) \right\rangle - \int_0^{\gamma} (t - \gamma) \left\langle g_n(x), \nabla^2 V(x - t g_n(x)) g_n(x) \right\rangle dt \\ &\leq \gamma \left\langle \nabla V(x), g_n(x) \right\rangle - L \int_0^{\gamma} (t - \gamma) \left\| g_n(x) \right\|^2 dt \\ &= \gamma \left\langle \nabla V(x), g_n(x) \right\rangle + \frac{L\gamma^2}{2} \left\| g_n(x) \right\|^2. \end{aligned} \tag{37}$$

For term II, we have by Lemma 9.2 that if

$$\gamma \leq \frac{1}{6 \sup_{x \in \mathbb{R}^d} \left\| \nabla g_n(x) \right\|_F}, \tag{38}$$

then

$$\mathcal{II} \leq \gamma \operatorname{div}\left(g_n(x)\right) + 5\gamma^2 \left\| \nabla g_n(x) \right\|_F^2. \tag{39}$$

So combine (37) and (39), we have

$$\beta\left(\mathcal{I} + \mathcal{II}\right) \geq \beta\gamma \left(\left\langle \nabla V(x), g_n(x) \right\rangle + \operatorname{div}\left(g_n(x)\right) + \gamma \left(\frac{L}{2} \left\| g_n(x) \right\|^2 + 5 \left\| \nabla g_n(x) \right\|_F^2 \right) \right), \tag{40}$$

under condition (38).

Combine (40), (36), (35) and use Jensen inequality $\psi\left(\mathbb{E}\left[f(X)\right]\right) \leq \mathbb{E}\left[\psi\left(f(X)\right)\right]$ with $\psi(x) = e^x - 1$ convex and $f(x) = \beta\Big(\log(\pi)(x) - \log(\pi)(\phi_n(x)) - \log(|\det \mathrm{D}\,\phi_n|)(x)\Big)$, we have

$$
\begin{aligned}
&e^{\beta \mathrm{D}_{\beta+1}\left(\rho_{n+1}|\pi\right)} - e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n|\pi\right)} \\
&= \left(\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx\right) \frac{\int \left(\left(\frac{\pi(x)}{\phi_n^{-1}{}_{\#}\pi(x)}\right)^{\beta} - 1\right)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx}{\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx} \\
&= \left(\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx\right) \frac{\int \left(\exp\left(\beta\Big(\log(\pi)(x) - \log(\pi)(\phi_n(x)) - \log(|\det \mathrm{D}\,\phi_n|)(x)\Big)\right) - 1\right)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx}{\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx} \\
&\geq \left(\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx\right)\left\{\exp\left(\frac{\int \beta\Big(\log(\pi)(x) - \log(\pi)(\phi_n(x)) - \log(|\det \mathrm{D}\,\phi_n|)(x)\Big)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)d\rho_n(x)}{\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx}\right) - 1\right\} \\
&\geq \left(\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx\right)\exp\left(\frac{\int \beta\gamma\left(\langle \nabla V(x), g_n(x)\rangle + \operatorname{div}\left(g_n(x)\right) + \gamma\left(\frac{L}{2}\left\|g_n(x)\right\|^2 + 5\left\|\nabla g_n(x)\right\|_F^2\right)\right)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)d\rho_n(x)}{\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx}\right) \\
&- \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx \\
&\geq \left(\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx\right) \\
&\times \left\{\exp\left(\frac{\beta\gamma\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\Big(\langle \nabla V(x), g_n(x)\rangle + \operatorname{div}\left(g_n(x)\right)\Big)d\rho_n(x)}{\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx} + \beta\gamma^2\frac{\int \left(\frac{L}{2}\left\|g_n(x)\right\|^2 + 5\left\|\nabla g_n(x)\right\|_2^2\right)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx}{\int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\rho_n(x)dx}\right) - 1\right\}.
\end{aligned}
\tag{41}
$$

For term $\mathcal{III} := \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\Big(\langle \nabla V(x), g_n(x)\rangle + \operatorname{div}\left(g_n(x)\right)\Big)d\rho_n(x)$ in the last line of the above equation, we have

$$
\begin{aligned}
\mathcal{III} &= \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\left\langle \nabla V(x), \left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\int k(x,y)\nabla \log\left(\frac{\rho_n}{\pi}\right)(y)d\rho_n(y)\right\rangle d\rho_n(x) \\
&\quad - \int \left\langle \nabla\left\{\rho_n(x)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\right\}, \left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\int k(x,y)\nabla \log\left(\frac{\rho_n}{\pi}\right)(y)d\rho_n(y)\right\rangle dx \\
&= \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\left\langle \nabla V(x), \int k(x,y)\nabla \log\left(\frac{\rho_n}{\pi}\right)(y)d\rho_n(y)\right\rangle d\rho_n(x) \\
&\quad - \int \left\langle \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\nabla\rho_n(x) + \beta\rho_n(x)\left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\nabla \log(\frac{\rho_n}{\pi})(x), \left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\int k(x,y)\nabla \log\left(\frac{\rho_n}{\pi}\right)(y)d\rho_n(y)\right\rangle dx \\
&= \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\left\langle \nabla V(x), \int k(x,y)\nabla \log\left(\frac{\rho_n}{\pi}\right)(y)d\rho_n(y)\right\rangle d\rho_n(x) \\
&\quad - \int \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)M^{-1} \vee \left(\frac{\pi}{\rho_n}\right)^{\beta}(x) \wedge M\left\langle \nabla \log\left(\rho_n\right)(x), \int k(x,y)\nabla \log\left(\frac{\rho_n}{\pi}\right)(y)d\rho_n(y)\right\rangle d\rho_n(x) \\
&\quad - \beta\iint \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\left(\frac{\pi}{\rho_n}\right)^{\beta} \wedge Mk(x,y)\left\langle \nabla \log\left(\frac{\rho_n}{\pi}\right)(x), \nabla \log\left(\frac{\rho_n}{\pi}\right)(y)\right\rangle d\rho_n(x)d\rho_n(y) \\
&= -(\beta+1)\iint \left(\frac{\rho_n}{\pi}\right)^{\beta}(x)\left(\frac{\pi}{\rho_n}\right)^{\beta} \wedge Mk(x,y)\left\langle \nabla \log\left(\frac{\rho_n}{\pi}\right)(x), \nabla \log\left(\frac{\rho_n}{\pi}\right)(y)\right\rangle d\rho_n(x)d\rho_n(y) \\
&\leq -(\beta+1)\left(I_{Stein}\left(\rho_n \mid \pi\right) - \delta\right).
\end{aligned}
\tag{42}
$$

Insert (42) into (41), we have

$$
\begin{aligned}
&e^{\beta \mathrm{D}_{\beta+1}\left(\rho_{n+1} \mid \pi\right)}-e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n \mid \pi\right)} \\
&\geq\left(\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x\right) \\
&\times\left\{\exp \left(\frac{-\beta(\beta+1) \gamma\left(I_{Stein}\left(\rho_n \mid \pi\right)-\delta\right)}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}+\beta \gamma^2 \frac{\int\left(\frac{L}{2}\left\|g_n(x)\right\|^2+5\left\|\nabla g_n(x)\right\|_F^2\right)\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}\right)-1\right\} .
\end{aligned}
\tag{43}
$$

The left thing is to upper bound $\mathcal{IV} := \frac{\int\left\|g_n(x)\right\|^2\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}$ and $\mathcal{V} := \frac{\int\left\|\nabla g_n(x)\right\|_F^2\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}$.
Firstly, by the reproducing property and Assumption 4.4, we have

$$
\|s(x)\|=\sqrt{\sum_{i=1}^d\left\|s_i(x)\right\|^2}=\sqrt{\sum_{i=1}^d\left\|\left\langle s_i(\cdot), k(x, \cdot)\right\rangle_{\mathcal{H}_0}\right\|^2} \leq \sqrt{\sum_{i=1}^d B^2\left\|s_i\right\|_{\mathcal{H}_0}^2}=B\|s\|_{\mathcal{H}}=B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}},
\tag{44}
$$

where $s(x) := \int k(x, y) \nabla \log \left(\frac{\rho_n}{\pi}\right)(y) d \rho_n(y)$, and so we have $\mathcal{IV} \leq M B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}}$. Secondly, also by the reproducing property and Assumption 4.4, we have

$$
\begin{aligned}
\|\nabla s(x)\|_F &=\sqrt{\sum_{i, j=1}^d\left|\frac{\partial s_i(x)}{\partial x_j}\right|^2}=\sqrt{\sum_{i, j=1}^d\left\langle\partial_{x_j} k(x, .), s_i\right\rangle_{\mathcal{H}_0}} \leq \sqrt{\sum_{i, j=1}^d\left\|\partial_{x_j} k(x, .)\right\|_{\mathcal{H}_0}^2\left\|s_i\right\|_{\mathcal{H}_0}^2} \\
&=\sqrt{\|\nabla k(x, .)\|_{\mathcal{H}}^2\|s\|_{\mathcal{H}}^2} \leq \sqrt{B^2\|s\|_{\mathcal{H}}^2}=B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}} .
\end{aligned}
\tag{45}
$$

and

$$
\begin{aligned}
\left\|\nabla g_n(x)\right\|_F &=\left\|\nabla\left(\frac{\pi}{\rho_n}\right)^\beta(x) s(x)^{\top} 1_{\left(\frac{\pi}{\rho_n}\right)^\beta(x) \in[0, M]}(x)+\left[\left(\frac{\pi}{\rho_n}\right)^\beta(x) \wedge M\right] \nabla s(x)\right\|_F \\
&\leq\left\|\beta\left(\frac{\pi}{\rho_n}\right)^\beta(x) \nabla \log \left(\frac{\pi}{\rho_n}\right)(x) s(x)^{\top} 1_{\left(\frac{\pi}{\rho_n}\right)^\beta(x) \in[0, M]}(x)\right\|_F+\left\|\left[\left(\frac{\pi}{\rho_n}\right)^\beta(x) \wedge M\right] \nabla s(x)\right\|_F \\
&\leq-\beta\left(\frac{\pi}{\rho_n}\right)^\beta(x)\left\|\nabla \log \left(\frac{\pi}{\rho_n}\right)(x)\right\| 1_{\left(\frac{\pi}{\rho_n}\right)^\beta(x) \in[0, M]}(x) B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}}+M B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}} \\
&\leq\left(-\beta M^{\frac{3}{2}} e^{C_b}+M\right) B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}},
\end{aligned}
\tag{46}
$$

where the last line is due to Lemma 9.1, and so we have $\mathcal{V} \leq\left(-\beta M^{\frac{3}{2}} e^{C_b}+M\right) B I_{Stein}\left(\rho_n \mid \pi\right)^{\frac{1}{2}}$.

Combine the upper bound of $\mathcal{IV}$ and $\mathcal{V}$, we have

$$
\frac{\int\left(\frac{L}{2}\left\|g_n(x)\right\|^2+5\left\|\nabla g_n(x)\right\|_F^2\right)\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x} \leq\left(\frac{L}{2} M^2+5(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right) B^2 I_{Stein}\left(\rho_n \mid \pi\right),
\tag{47}
$$

and

$$
\begin{aligned}
&e^{\beta \mathrm{D}_{\beta+1}\left(\rho_{n+1} \mid \pi\right)}-e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n \mid \pi\right)} \\
&\geq\left(\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x\right) \exp \left(\frac{-\beta(\beta+1) \gamma\left(I_{Stein}\left(\rho_n \mid \pi\right)-\delta\right)}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}+\beta \gamma^2 B^2 I_{Stein}\left(\rho_n \mid \pi\right)\left(\frac{L}{2} M^2+5(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right)\right) \\
&\quad-\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x .
\end{aligned}
\tag{48}
$$

Since $\int\left(\frac{\rho_n}{\pi}\right)^\beta d \rho_n(x) \leq 1$ when $\beta \in(-1,0)$, so if set $\gamma \leq \frac{2(\beta+1)\left(I_{Stein}\left(\rho_n \mid \pi\right)-\delta\right)}{B^2 I_{Stein}\left(\rho_n \mid \pi\right)\left(L M^2+10(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right)}$, we will have $\frac{-\beta(\beta+1) \gamma\left(I_{Stein}\left(\rho_n \mid \pi\right)-\delta\right)}{\int\left(\frac{\rho_n}{\pi}\right)^\beta(x) \rho_n(x) d x}+\beta \gamma^2 B^2 I_{Stein}\left(\rho_n \mid \pi\right)\left(\frac{L}{2} M^2+5(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right) \geq 0$. Finally

we use $e^x \geq 1+x$ when $x \geq 0$ to get

$$\begin{aligned} e^{\beta \mathrm{D}_{\beta+1}\left(\rho_{n+1}|\pi\right)} - e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n|\pi\right)} &\geq -\beta(\beta+1)\left(I_{Stein}\left(\rho_n \mid \pi\right) - \delta\right) \\ &\quad + \beta\gamma^2 B^2 I_{Stein}\left(\rho_n \mid \pi\right)\left(\frac{L}{2}M^2 + 5(-\beta M^{\frac{3}{2}} e^{C_b} + M)^2\right) e^{\beta \mathrm{D}_{\beta+1}\left(\rho_n|\pi\right)} \\ &\geq -\beta(\beta+1)\gamma\left(\frac{1}{2} I_{Stein}\left(\rho_n \mid \pi\right) - \delta\right), \end{aligned} \tag{49}$$

the last line is because we choose $\gamma \leq \frac{\beta+1}{B^2\left(LM^2+10(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right)}$.

Now, we can finish the proof by giving the condition on the step-size γ:

$$\begin{cases} 0 < \gamma \leq \frac{1}{6\left(-\beta M^{\frac{3}{2}} e^{C_b}+M\right) B I_{Stein}(\rho_n|\pi)^{\frac{1}{2}}} \\ 0 < \gamma \leq \frac{2(\beta+1)(I_{Stein}(\rho_n|\pi)-\delta)}{B^2 I_{Stein}(\rho_n|\pi)\left(LM^2+10(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right)} \\ 0 < \gamma \leq \frac{\beta+1}{B^2\left(LM^2+10(-\beta M^{\frac{3}{2}} e^{C_b}+M)^2\right)} \end{cases}, \tag{50}$$

where M satisfies condition (34). □

Lemma 9.1. *Under Assumption 4.5, we have*

$$\sup_{x\in\mathbb{R}^d}\left(\frac{\pi}{\rho_n}\right)^{\beta}(x)\left\|\nabla \log\left(\frac{\pi}{\rho_n}\right)(x)\right\| 1_{\left(\frac{\pi}{\rho_n}\right)^{\beta}(x)\in[0,M]}(x) \leq C_f < \infty, \tag{51}$$

for some constant $C_f \in [0, M^{\frac{3}{2}} e^{C_b}]$.

Proof. (51) is equivalent to

$$\beta f_n(x) + \log(\|\nabla f_n(x)\|) \leq \log(C_f), \forall x \in \mathbb{R}^d \text{ such that } \beta f_n(x) \leq \log(M), \tag{52}$$

where $f_n(x) := \log\left(\frac{\pi}{\rho_n}\right)(x)$. If f_n satisfies Assumption 4.5, then we have

$$\beta f_n(x) + \log(\|\nabla f_n(x)\|) \leq \beta f_n(x) + \frac{|\beta f_n(x)|}{2} + C_b \leq \frac{3}{2}\log(M) + C_b, \tag{53}$$

for any $x \in \mathbb{R}^d$ such that $\beta f_n(x) \leq \log(M)$. So finally, we have

$$\sup_{x\in\mathbb{R}^d}\left(\frac{\pi}{\rho_n}\right)^{\beta}(x)\left\|\nabla \log\left(\frac{\pi}{\rho_n}\right)(x)\right\| 1_{\left(\frac{\pi}{\rho_n}\right)^{\beta}(x)\in[0,M]}(x) \leq C_f, \tag{54}$$

where constant C_f satisfies

$$C_f \leq M^{\frac{3}{2}} e^{C_b} < \infty. \tag{55}$$

□

Lemma 9.2. *Let B be a square matrix and $\|B\|_F = \sqrt{\sum_{ij} b_{ij}^2}$ its Frobenius norm. Let ϵ be a positive number that satisfies $0 \leq \epsilon < \frac{1}{3\|B\|_F}$. Then $I + \epsilon\left(B + B^\top\right) + \epsilon^2 BB^\top$ is positive definite, and*

$$\begin{aligned} &\epsilon \operatorname{tr}(B) - \frac{\epsilon^2}{4}\left(\frac{9\|B\|_F^2}{1-3\epsilon\|B\|_F} + 2\|B\|_F^2\right) \\ &\leq \log|\det(I+\epsilon B)| \\ &\leq \epsilon \operatorname{tr}(B) - \frac{\epsilon^2}{4}\left(\frac{9\|B\|_F^2}{1+3\epsilon\|B\|_F} + 2\|B\|_F^2\right). \end{aligned} \tag{56}$$

Therefore, take an even smaller ϵ such that $0 \leq \epsilon \leq \frac{1}{6\|B\|_F}$, we get

$$\epsilon \operatorname{tr}(B) - 5\epsilon^2 ||B||_F^2 \leq \log|\det(I+\epsilon B)| \leq \epsilon \operatorname{tr}(B) - 2\epsilon^2 ||B||_F^2.$$

Proof. We follow the proof from Liu [2017]. When $\epsilon < \frac{1}{\varrho\left(B+B^\top\right)}$, where $\varrho(\cdot)$ denotes the spectrum radius, we have

$$\varrho\left(I + \epsilon\left(B + B^\top\right) + \epsilon^2 BB^\top\right) \geq 1 - \epsilon\varrho\left(B + B^\top\right) > 0,$$

and so $I + \epsilon\left(B + B^\top\right) + \epsilon^2 BB^\top$ is positive definite. By the property of matrix determinant, we have

$$\begin{aligned}\log|\det(I+\epsilon B)| &= \frac{1}{2}\log\det\left((I+\epsilon B)(I+\epsilon B)^\top\right)\\ &= \frac{1}{2}\log\det\left(I+\epsilon\left(B+B^\top\right)+\epsilon^2 BB^\top\right)\\ &= \frac{1}{2}\log\det\left(I+\epsilon\left(B+B^\top+\epsilon BB^\top\right)\right).\end{aligned} \tag{57}$$

Let $A = B + B^\top + \epsilon BB^\top$, we can establish

$$\epsilon\operatorname{tr}(A) - \frac{\epsilon^2}{2}\frac{\|A\|_F^2}{1-\epsilon\varrho(A)} \le \log\det(I+\epsilon A) \le \epsilon\operatorname{tr}(A) - \frac{\epsilon^2}{2}\frac{\|A\|_F^2}{1+\epsilon\varrho(A)},$$

which holds for any symmetric matrix A and $0 \le \epsilon < 1/\varrho(A)$. This is because, assuming $\{\lambda_i\}$ are the eigenvalues of A,

$$\begin{aligned}\log\det(I+\epsilon A) - \epsilon\operatorname{tr}(A) &= \sum_i \left[\log(1+\epsilon\lambda_i) - \epsilon\lambda_i\right]\\ &= \sum_i\left[\int_0^1 \frac{\epsilon\lambda_i}{1+s\epsilon\lambda_i}\,\mathrm{d}s - \epsilon\lambda_i\right]\\ &= -\sum_i\int_0^1 \frac{s\epsilon^2\lambda_i^2}{1+s\epsilon\lambda_i}\,\mathrm{d}s,\end{aligned}$$

while

$$\begin{aligned}-\frac{\epsilon^2}{2}\frac{\|A\|_F}{1-\epsilon\varrho(A)} &= -\frac{1}{2}\sum_i\frac{\epsilon^2\lambda_i^2}{1-\epsilon\max_i|\lambda_i|}\\ &\le -\sum_i\int_0^1\frac{s\epsilon^2\lambda_i^2}{1+s\epsilon\lambda_i}\,\mathrm{d}s\\ &\le -\frac{1}{2}\sum_i\frac{\epsilon^2\lambda_i^2}{1+\epsilon\max_i|\lambda_i|}\\ &= -\frac{\epsilon^2}{2}\frac{\|A\|_F}{1+\epsilon\varrho(A)},\end{aligned}$$

so we have

$$-\frac{\epsilon^2}{2}\frac{\|A\|_F}{1-\epsilon\varrho(A)} \le \log\det(I+\epsilon A) - \epsilon\operatorname{tr}(A) \le -\frac{\epsilon^2}{2}\frac{\|A\|_F}{1+\epsilon\varrho(A)}. \tag{58}$$

Taking $A = B + B^\top + \epsilon BB^\top$ into Equation (58) and combine it with Equation (57), we get

$$\begin{aligned}\log|\det(I+\epsilon B)| &\ge \frac{1}{2}\log\det\left(I+\epsilon\left(B+B^\top+\epsilon BB^\top\right)\right)\\ &\ge \frac{\epsilon}{2}\operatorname{tr}\left(B+B^\top+\epsilon BB^\top\right) - \frac{\epsilon^2}{4}\frac{\left\|B+B^\top+\epsilon BB^\top\right\|_F^2}{1-\epsilon\varrho\left(B+B^\top+\epsilon BB^\top\right)}\\ &\ge \epsilon\operatorname{tr}(B) - \frac{\epsilon^2}{4}\left(\frac{9\|B\|_F^2}{1-\epsilon\varrho\left(B+B^\top+\epsilon BB^\top\right)} + 2\|B\|_F^2\right),\end{aligned}$$

similarly

$$\log|\det(I+\epsilon B)| \le \epsilon\operatorname{tr}(B) - \frac{\epsilon^2}{4}\left(\frac{9\|B\|_F^2}{1+\epsilon\varrho\left(B+B^\top+\epsilon BB^\top\right)} + 2\|B\|_F^2\right)$$

where we used the fact that $\operatorname{tr}(B) = \operatorname{tr}\left(B^\top\right)$, $\left\|BB^\top\right\|_F \le \|B\|_F^2$ and $\left\|B+B^\top+\epsilon BB^\top\right\|_F \le \|B\|_F + \left\|B^\top\right\|_F + \epsilon\left\|BB^\top\right\|_F = 3\|B\|_F$ (since $\epsilon \le \frac{1}{\|B\|_F}$). Finally we use inequality $\varrho\left(B+B^\top+\epsilon BB^\top\right) \le \varrho\left(B+B^\top\right) + \epsilon\varrho\left(BB^\top\right) \le \varrho\left(B+B^\top\right) + \sqrt{\varrho\left(BB^\top\right)}$ and

$$\begin{aligned}\varrho(B+B^\top)^2 &\le \operatorname{tr}\left(BB+BB^\top+B^\top B+B^\top B^\top\right)\\ &= \operatorname{tr}(BB) + \operatorname{tr}(B^\top B^\top) + 2\operatorname{tr}(BB^\top)\\ &\le 4\operatorname{tr}(BB^\top) \qquad\qquad //\text{since } \operatorname{tr}(BB) \le \operatorname{tr}(BB^\top)\\ &= 4\|B\|_F^2\end{aligned}$$

and $\varrho(BB^\top) \le \|B\|_F^2$, so we have

$$\varrho\left(B + B^\top + \epsilon BB^\top\right) \le 3\|B\|_F . \tag{59}$$

Combining all of these, we finally get

$$\begin{aligned} &\epsilon \operatorname{tr}(B) - \frac{\epsilon^2}{4}\left(\frac{9\|B\|_F^2}{1 - 3\epsilon\|B\|_F} + 2\|B\|_F^2\right) \\ &\le \log|\det(I + \epsilon B)| \\ &\le \epsilon \operatorname{tr}(B) - \frac{\epsilon^2}{4}\left(\frac{9\|B\|_F^2}{1 + 3\epsilon\|B\|_F} + 2\|B\|_F^2\right). \end{aligned} \tag{60}$$

□

10 Experiments on Bayesian Neural Network and Bayesian Logistic Regression

Bayesian Neural Network: In this experiment, we compare SVGD with the proposed -0.5-SVGD on Bayesian Neural Networks. Our experiment setting is similar to the one in Liu and Wang [2018]: we use neural networks with one hidden layers, and take 50 hidden units for all the datasets; all the datasets are randomly partitioned into 90% for training and 10% for testing; we use $\mathrm{RELU}(x) = \max(0, x)$ as the active function. We set the particle number $N = 100$ and the mini-batch size as 100. We set the step-size $\gamma = 10^{-4}$. In -0.5-SVGD, we renormalize the Stein matrix by a constant factor, set the mirror descent iteration number $m = 50$, step-size $r = 0.5$ and small gap $\tau = 0.25$. The Stein matrix has small dimension in these cases, which is only 100×100, so the computation of the Stein importance weights is efficient. For each dataset: Seeds, Boston house and Yacht, and each algorithm: SVGD, -0.5-SVGD, we test for 3 times (this experiment is time consuming) with iteration number $n = 500, 1000, 2000$. Table 1 shows the averaged test results, where RMSE means the root mean square deviation, LL means the log likelihood. We find -0.5-SVGD consistently improves over SVGD in terms of accuracy. For the original data obtained, see Section 11.

SEED with dimension 453				
	RMSE		LL	
Iteration	SVGD	-0.5-SVGD	SVGD	-0.5-SVGD
500	0.614621857	**0.597111981**	−1.020867456	**-0.961816462**
1000	0.550036399	**0.533375435**	−0.910755822	**-0.827453328**
2000	0.510597252	**0.49012237**	−0.807832642	**-0.726703193**
BOSTON with dimension 753				
	RMSE		LL	
Iteration	SVGD	-0.5-SVGD	SVGD	-0.5-SVGD
500	6.540663835	**6.054613357**	−3.330431563	**-3.193928609**
1000	6.163087914	**5.821032633**	−3.241251933	**-3.134616289**
2000	5.871055387	**5.541809203**	−3.164638536	**-3.07889812**
YACHT with dimension 403				
	RMSE		LL	
Iteration	SVGD	-0.5-SVGD	SVGD	-0.5-SVGD
500	9.48556172	**7.785107764**	−3.820548866	**-3.658675596**
1000	6.658104806	**5.343377953**	−3.606474782	**-3.399616321**
2000	4.812189172	**3.543966898**	−3.360595158	**-3.066375285**

Table 1: Average Test Results: the random seeds are set the same for SVGD and -0.5-SVGD in each test and the running time for -0.5-SVGD is roughly double that of SVGD.

Bayesian Logistic Regression: In this experiment, we compare the performance of SVGD and -0.5-SVGD on the Bayesian Logistic regression problem. The experiment setting is almost the same as the one in Liu and Wang [2016], for more details, please refer to Liu and Wang [2016] or Section 11. In the test, we found the Stein importance weight is close to the identical weight after only a few -0.5-SVGD iterations with relatively big step-size (specifically, the percentage of weight ω_i such that $N\omega_i < 0.1$ falls to 0 after the first few iterations of -0.5-SVGD), so the acceleration effect is not very clear in this case. However, as shown in the first image, where the step-size is relatively small, we can see a faster improvement in accuracy in the first few hundreds iterations of -0.5-SVGD. We can also see from the results that when γ is relatively large, due to the Stein importance weight, -0.5-SVGD is much more stable than SVGD.

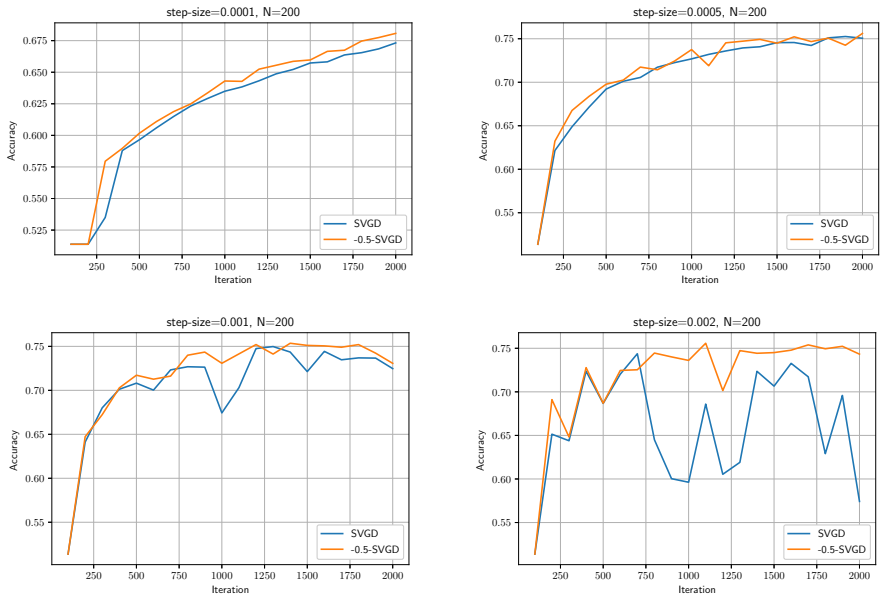

Figure 2: We test the binary Covertype dataset with 581,012 data points and 54 features ($d = 54$). We run 2000 iterations of SVGD and -0.5-SVGD with different step-size and number of particles. In each iteration of -0.5-SVGD, we run 200 steps of mirror descent with $r = 2$ to find the Stein importance weights, we set the small gap $\tau = 0.05$. The time required to run 2000 iterations of -0.5-SVGD and test the accuracy every 100 iterations is roughly double that required for SVGD.

11 More Experiments and Details

We run the experiments on a Macbook Pro (13-inch,2020) with Processor: 2.3 GHz Quad-Core Intel Core i7 and Memory: 32 GB 3733 MHz LPDDR4X.

11.1 More Experiments on Gaussian Mixture

In Figure 1, Figure 3, Figure 4 and Figure 5, we use Gaussian Mixture to test the performance of β-SVGD. We choose the reproducing kernal $k(x, y) = e^{-\frac{\|x-y\|^2}{d}}$, where d is the dimension.

11.2 More Experiments on Bayesian Logistic Regression

In Figure 6, we compare the performance of SVGD and β-SVGD with $\beta = -0.5$ in Bayesian Logistic regression problem. This Bayesian Logistic regression experiment is done in Liu and Wang [2016] to compare SVGD with several Markov Chain Monte Carlo methods, more details about this experiment can refer to Liu and Wang [2016]. As in the Gaussian Mixture experiment, we choose the reproducing kernal $k(x, y) = e^{-\frac{\|x-y\|^2}{d}}$.

11.3 More Details on Bayesian Neural Network

Table 2 shows the original test data on Bayesian Neural Network.

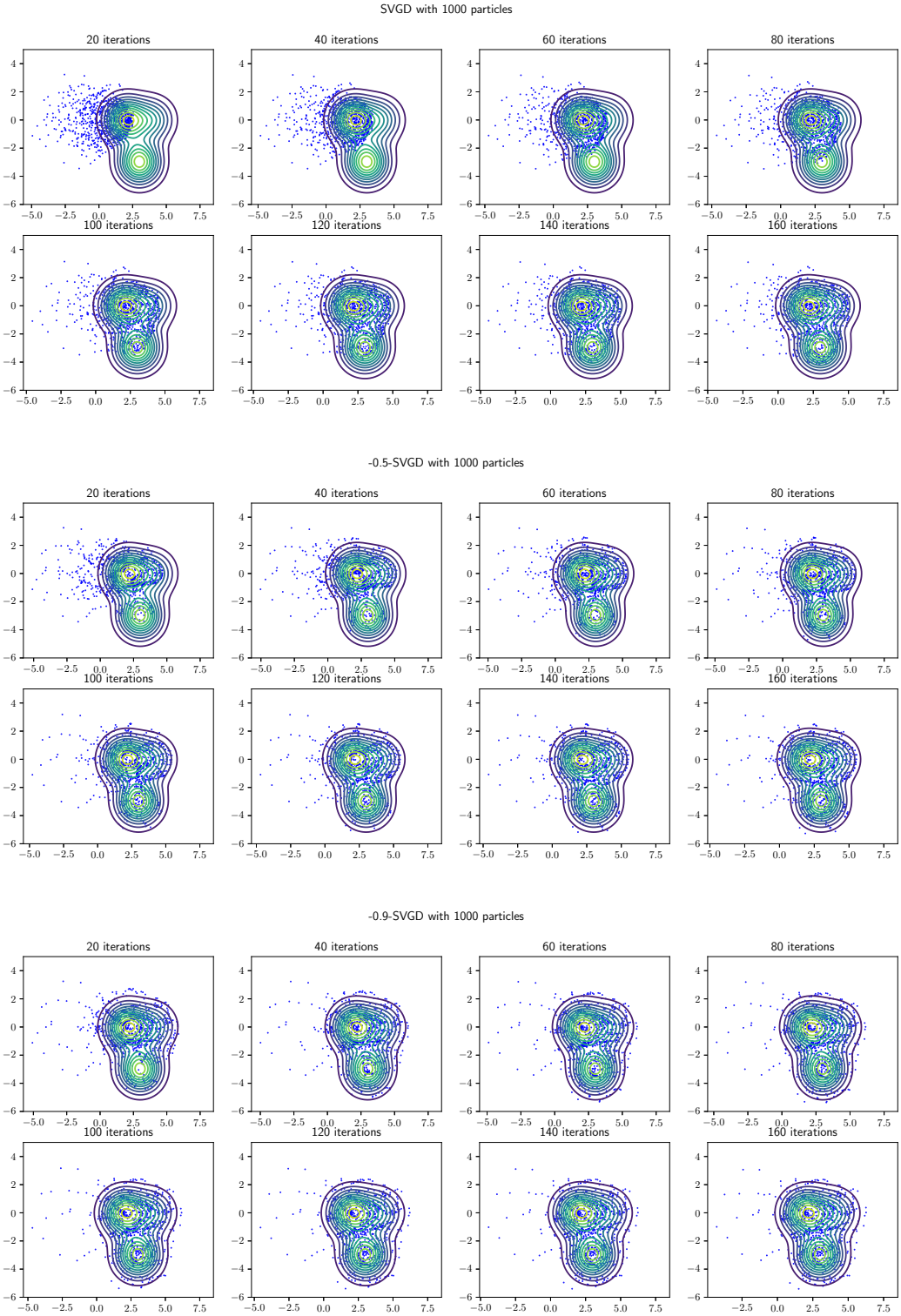

Figure 3: The same experiment setting as in Figure 1. We show how the particles move in the update of β-SVGD with $\beta = 0, -0.5, -0.9$.

0.5-SVGD						
500 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.65270335	-1.127949146	7.500990754	-3.578065184	10.90588982	-3.984105331
2nd Test	0.650713242	-1.109967755	7.560059158	-3.564446731	11.4522142	-3.995176488
3rd Test	0.648986116	-1.080414149	7.190867282	-3.544465893	12.19846953	-4.050929325
1000 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.579153432	-1.065757383	6.798066541	-3.548783886	8.355382099	-3.897892911
2nd Test	0.583035518	-1.042295985	7.032850283	-3.517889088	8.614608135	-3.854131369
3rd Test	0.580088164	-1.023501313	6.71675912	-3.46875477	8.792347324	-3.898790508
2000 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.543722837	-1.028549237	6.58145612	-3.476708161	7.460360406	-3.836464327
2nd Test	0.564119866	-1.046072614	6.894245042	-3.452564756	7.348647643	-3.790306767
3rd Test	0.565132246	-1.006777082	6.618980571	-3.42621784	7.511270881	-3.826691374
SVGD						
500 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.612381691	-1.023857439	6.536863165	-3.319276377	9.308204171	-3.802498566
2nd Test	0.614235376	-1.028596212	6.652164723	-3.361659105	9.207817711	-3.806520653
3rd Test	0.617248504	-1.010148718	6.432963616	-3.310359208	9.94066327	-3.852627381
1000 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.548803339	-0.920039422	6.129287038	-3.231650337	6.582751393	-3.601019046
2nd Test	0.558942951	-0.925617411	6.280277766	-3.267793394	6.511789693	-3.591982423
3rd Test	0.542362907	-0.886610635	6.079698939	-3.224312069	6.879773332	-3.626422876
2000 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.508115157	-0.82032888	5.824160413	-3.16041858	4.683719162	-3.352857754
2nd Test	0.521839905	-0.822264856	5.965881345	-3.179169278	4.722228614	-3.345166653
3rd Test	0.501836692	-0.780904191	5.823124403	-3.15432775	5.03061974	-3.383761067
-0.5-SVGD						
500 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.595794017	-0.966862028	5.997885458	-3.177899474	7.931862638	-3.646890816
2nd Test	0.592175621	-0.952947097	6.090872737	-3.209068156	7.467320009	-3.63783219
3rd Test	0.603366307	-0.96564026	6.075081876	-3.194818196	7.956140646	-3.691303781
1000 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.53132468	-0.825319698	5.786683985	-3.128738437	5.309869371	-3.37805003
2nd Test	0.535686912	-0.838001442	5.874849706	-3.147433109	5.136576962	-3.373933488
3rd Test	0.533114714	-0.819038846	5.801564209	-3.127677321	5.583687525	-3.446865446
2000 Iteration	SEEDS-RMSE	SEEDS-LL	BOSTON-RMSE	BOSTON-LL	YACHT-RMSE	YACHT-LL
1st Test	0.490918366	-0.729968036	5.538989749	-3.079445225	3.388128107	-3.041888645
2nd Test	0.489448573	-0.729185581	5.577819672	-3.079523837	3.361486382	-3.022368475
3rd Test	0.490000173	-0.720955963	5.508618189	-3.077725298	3.882286206	-3.134868736

Table 2: Original Test result of Bayesian Neural Network: in each test, the random seed is set the same (set it 1) for each algorithm. we can observe the effectiveness of the Stein importance weights: 0.5-SVGD is worse than SVGD and -0.5-SVGD is better than SVGD, which matches the theoretical prediction of Theorem 3.2.

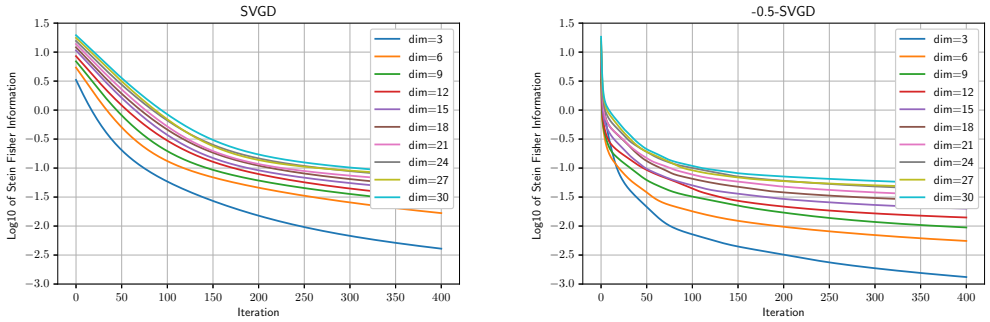

Figure 4: In this experiment, we show how the Stein Fisher information changes in the update of SVGD and -0.5-SVGD. The target distribution is $\mathcal{N}((2,\ldots,2)_d, I_d)$ and the initial points are sampled from $\mathcal{N}((0,\ldots,0)_d, I_d)$ with $N = 300$. The step-size $\gamma = 0.1$ for both algorithm and for -0.5-SVGD algorithm, we set the small gap $\tau = 0.01$ and we update the Stein importance weight in every iteration using 40 mirror descent with step-size $r = 0.3$. We can see that the Stein Fisher information drops immediately below 1 (note in the picture, the axis y is $\log_{10}$ of the Stein Fisher information) in -0.5-SVGD, while in SVGD it drops slowly.

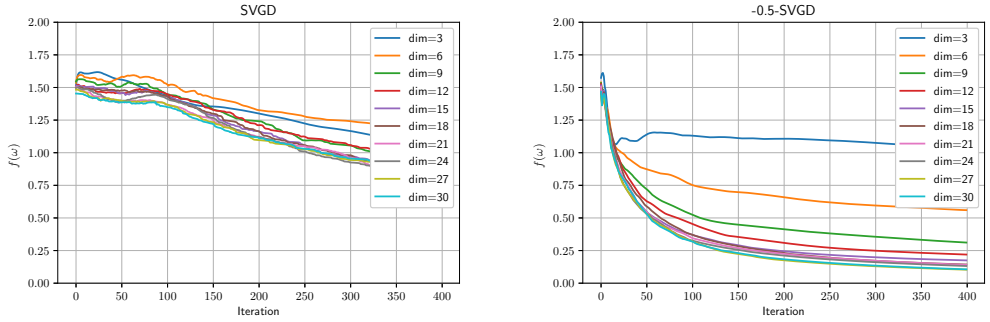

Figure 5: The experiment settings are the same as in Figure 4. We compare how the Stein importance weight changes in the update of SVGD and -0.5-SVGD (though we don't have to compute the Stein importance weight in the implementation of SVGD). The error is defined by $f(\omega^k) := \sum_{i=1}^N |w_i^k - \frac{1}{N}|$, where ω_i^k denote the Stein importance weight of point x_i^k and $N = 300$. The results suggest that in high dimensional cases, the Stein importance weight can help to accelerate the decreasing of Stein Fisher information in the beginning, then it will approach to the identical weight $\frac{1}{N}$ quickly.

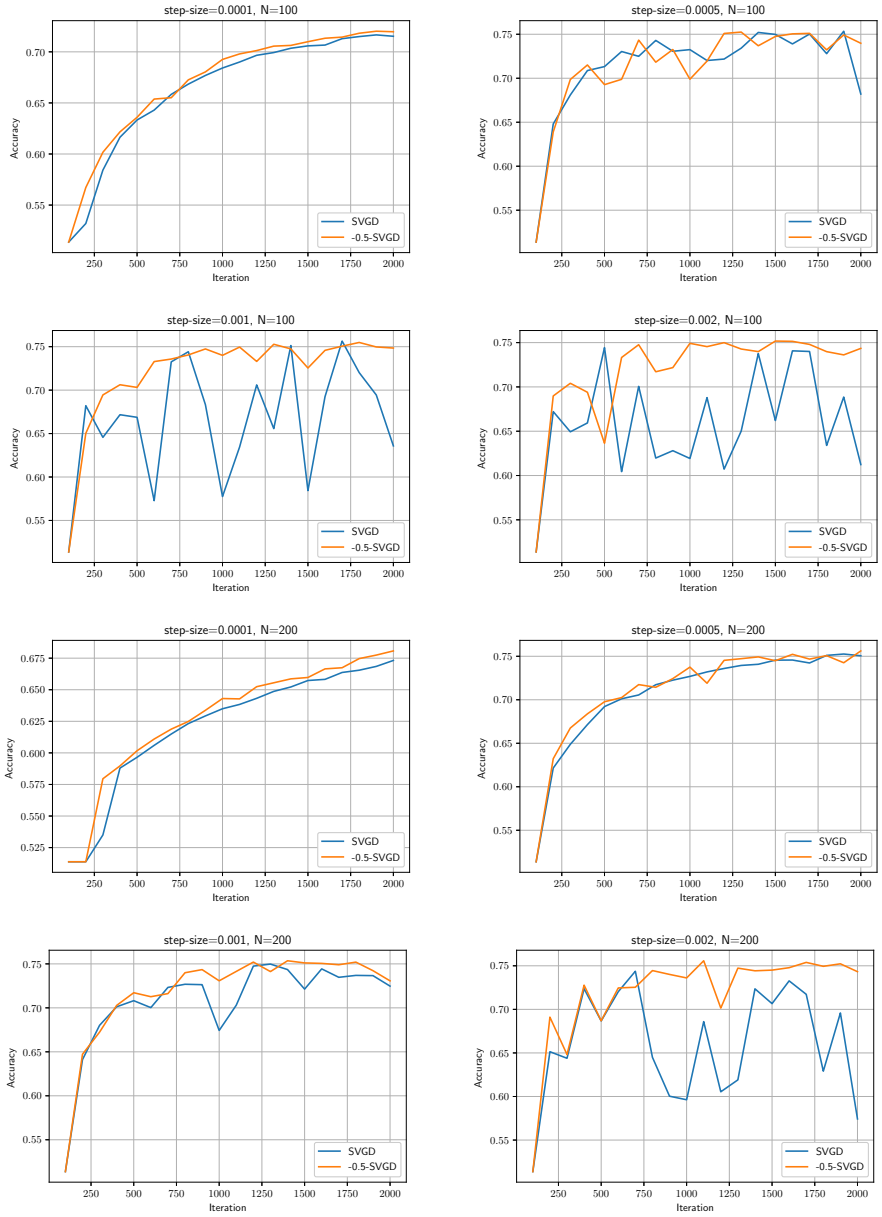

Figure 6: In this experiment, we test the binary Covertype dataset with 581,012 data points and 54 features ($d = 54$). We run 2000 iterations of SVGD and β-SVGD with different step-size and number of particles. In each iteration of -0.5-SVGD, we run 200 steps of mirror descent with step-size $r = 2$ (since the values of the entries of $\boldsymbol{K}_\pi$ in this experiment can be very big, we need to rescale the matrix by dividing a factor of 10^9 to resolve the overflow problem, so the step-size for mirror descent is chosen relatively big) to find the Stein importance weights, we set the small gap $\tau = 0.05$. The time required to run 2000 iterations of -0.5-SVGD and test the accuracy every 100 iterations is roughly double that required for SVGD. In this experiment, we found the Stein importance weight is close to the identical weight after only a few β-SVGD iterations with relatively big step-size (specifically, the percentage of weight ω_i such that $N\omega_i < 0.1$ falls to 0 after the first few iterations of -0.5-SVGD), so the acceleration effect is not very clear in this case. However, as shown in the first and fifth images, where the step-size is relatively small, we can see a faster improvement in accuracy in the first few hundreds iterations of β-SVGD. We can also see from the results that when γ is relatively large, due to the Stein importance weight, -0.5-SVGD is much more stable than SVGD.