Convergence of Stein Variational Gradient Descent under a Weaker Smoothness Condition

Lukang Sun
AI Inititative
KAUST

Avetik Karagulyan
AI Inititative
KAUST

Peter Richtárik
AI Inititative
KAUST

Abstract

Stein Variational Gradient Descent (SVGD) is an important alternative to the Langevin-type algorithms for sampling from probability distributions of the form $\pi(x) \propto \exp(-V(x))$. In the existing theory of Langevin-type algorithms and SVGD, the potential function V is often assumed to be L-smooth. However, this restrictive condition excludes a large class of potential functions such as polynomials of degree greater than 2. Our paper studies the convergence of the SVGD algorithm in population limit for distributions with (L_0, L_1)-smooth potentials. This relaxed smoothness assumption was introduced by Zhang et al. (2019a) for the analysis of gradient clipping algorithms. With the help of trajectory-independent auxiliary conditions, we provide a descent lemma establishing that the algorithm decreases the KL divergence at each iteration and prove a complexity bound for SVGD in the population limit in terms of the Stein Fisher information.

1 Introduction

Bayesian methods are widely implemented in various inference tasks that emerge in computational statistics (Neal, 1992; Roberts and Tweedie, 1996), machine learning (Grenander and Miller, 1994; Şimşekli, 2017), inverse problems (Zhou et al., 2020; Durmus et al., 2018) and model selection (Feroz and Skilling, 2013; Cai et al., 2021). Often Bayesian methods boil down to approximate integration problems which are solved using Markov Chain Monte-Carlo algorithms (Robert and Casella, 1999). In practice, the target distribution π defined on $\mathbb{R}^d$ is often absolutely continuous w.r.t. the Lebesgue measure and we have access to its density up to a normalization constant. Throughout the paper we will use the same notation for probability distributions and their corresponding densities. In Bayesian statistics the target distribution is generally given by

$$\pi(x) = \frac{1}{Z} e^{-V(x)},$$

where Z is the normalization constant and $V : \mathbb{R}^d \to \mathbb{R}$ is called potential function. Here, we usually do not know the value of Z. The approximate sampling methods are proceed as follows. For every $\varepsilon > 0$, the goal is to construct a distribution μ that approximates the target π in some probability distance dist:

$$\mathsf{dist}(\mu, \pi) < \varepsilon.$$

In this paper, we will study the case when dist is the Kullback-Leibler divergence:

$$\begin{aligned} \mathrm{KL}(\mu \mid \pi) &:= \int \log\left(\frac{\mu(x)}{\pi(x)}\right) \mu(\mathrm{d}x) \\ &= \mathbb{E}_{X\sim\mu}\big[V(X)\big] - H(\mu). \end{aligned} \tag{1}$$

Here, $H(\mu) := -\int \log(\mu(x))\mu(\mathrm{d}x)$ is the negative entropy. The sampling problem can also be seen as a minimization problem in the space of probability measures (see e.g. (Liu and Wang, 2016; Wibisono, 2018; Durmus et al., 2019)). Indeed, consider the functional $\mathcal{F}(\cdot) := \mathrm{KL}(\cdot \mid \pi)$. This functional is non-negative and it admits its minimum value 0 only for the target distribution $\pi = \arg\min_\mu \mathcal{F}(\mu)$. The connection of sampling and optimization has been repeatedly leveraged in the previous literature of Langevin sampling. Various algorithms such as Langevin Monte-Carlo (e.g. Dalalyan (2017a); Wibisono (2019); Durmus et al. (2019)), Underdamped Langevin Algorithm (e.g. Ma et al. (2019); Chatterji et al. (2018)) have often been influenced by the known optimization methods. Conversely, another line of research has studied the application of sampling algorithms to solve optimization problems (see e.g. (Raginsky et al., 2017)). In this paper, we will study another algorithm called Stein Variational Gradient Descent (SVGD), which is designed as a gradient descent algorithm for the KL divergence in the space of probability measures.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1.1 SVGD

The LMC algorithm, treats the functional $\mathrm{KL}(\cdot|\pi)$ as a composite functional described in (1). For details, we refer the reader to Appendix A (see also (Wibisono, 2018)). Unlike the LMC, the Stein Variational Gradient Descent (SVGD) algorithm applies gradient descent directly to $\mathrm{KL}(\cdot|\pi)$ (see Section 2.2 for the complete definition). SVGD is an important alternative to the Langevin algorithm and already has been used extensively in different settings of machine learning, such as variational auto-encoders (Pu et al., 2017), reinforcement learning (Liu et al., 2017), sequential decision making (Zhang et al., 2018, 2019b), generative adversarial networks (Tao et al., 2019) and federated learning (Kassab and Simeone, 2022).

The seminal work of Liu and Wang (2016) introduced SVGD as a sampling method. Since then several variants of SVGD have been proposed. Here is a non-exclusive list: random batch method SVGD (Li et al., 2020), matrix kernel SVGD (Wang et al., 2019), Newton version SVGD (Detommaso et al., 2018), stochastic SVGD (Gorham et al., 2020) and mirrored SVGD (Shi et al., 2021). However, the theoretical understanding of SVGD is still limited to its population limit or the so called infinite-particle regime. See Section 4 for a discussion on the finite-particle regime. The work by Liu (2017) was the first that proved a convergence result of the SVGD in the population limit. Later, Duncan et al. (2019) studied the geometry related to the SVGD and proposed a scheme to choose the kernel. In (Korba et al., 2020), a descent lemma was established for the SVGD in population limit in terms of Kullback-Leibler divergence. The drawback of this result was that the analysis relied on the path information of the SVGD which is unknown beforehand. Salim et al. (2021) improved the work of Korba et al. (2020) and provided a clean analysis for the convergence. They assumed π satisfies Talagrand's T_1 (see Section 2) inequality which essentially replaced the initial trajectory condition. This implied a complexity bound for the SVGD in terms of the desired accuracy ϵ and dimension d. To the best of our knowledge, this is the state of the art result on SVGD.

1.2 Contributions

The main contribution of the paper relies on its weaker set of assumptions, that allow to treat a larger class of probability distributions which includes densities with polynomial potentials. We enlarge the class of probability distributions two-fold.

Smoothness: The L-smoothness assumption is very common in the sampling literature (see e.g., Durmus and Moulines (2017); Dalalyan (2017b); Dalalyan and Karagulyan (2019); Durmus et al. (2019); Vempala and Wibisono (2019); Shen and Lee (2019); Korba et al. (2020); Salim et al. (2021)). It is formulated as the L-Lipschitz continuity of the gradient. That is the Hessian $\nabla^2 V(\cdot)$ is well-defined on $\mathbb{R}^d$ and

$$\left\|\nabla^2 V(x)\right\|_{op} \leq L, \quad \forall x \in \mathbb{R}^d, \tag{2}$$

where $L > 0$ and $\|\cdot\|_{op}$ is the operator norm defined in Section 2.1. Though people have made great progress towards the understanding of these algorithms, the L-smoothness condition is quite restrictive. In fact from L-smoothness condition we can easily deduce that V has at most quadratic growth rate. In particular, the large class of polynomial potentials whose order is higher than 2 does not satisfy this condition. For the LMC algorithm, several papers have proposed different methods to relax or remove this assumption (Lehec, 2021; Brosse et al., 2019; Hutzenthaler et al., 2012; Sabanis, 2013; Erdogdu and Hosseinzadeh, 2021; Chewi et al., 2021; Balasubramanian et al., 2022).

However, to the best of our knowledge there is no result for the SVGD under a relaxed smoothness assumption. In this paper we use (L_0, L_1) smoothness (see Assumption (L_0, L_1) for the definition). The latter was initially proposed by Zhang et al. (2019a), for the gradient clipping algorithm. The new condition with parameters $(L, 0)$ is equivalent to the L-smoothness and therefore, it is indeed a weaker assumption. Important example of such functions are the higher order polynomials (see Section 5 for details).

Functional inequalities: The analysis of Langevin algorithms often relies on the strong convexity of the potential function V. A line of work has proposed different methods to relax or bypass this assumption. One possible solution is to slightly modify the original algorithm (e.g. (Dalalyan et al., 2022; Karagulyan and Dalalyan, 2020)). Another approach relies on functional inequalities such as Poincaré or logarithmic Sobolev inequalities (e.g. (Vempala and Wibisono, 2019; Chewi et al., 2020; Ahn and Chewi, 2021)). However, the verification of these inequalities is not straightforward.

Unlike the Langevin algorithm, the SVGD algorithm does not require convexity in any form. However, functional inequalities also serve for the analysis of the SVGD but with a different purpose. Salim et al. (2021) have proved a complexity result in dimension and precision error for the SVGD algorithm performed on the targets that satisfy Talagrand's T_1 inequality (see (6) for details). This assumption replaces the bound on the trajectory that was initially proposed by Korba et al. (2020). Despite the major improvement, they still assumed the L-smoothness of the potential, which as mentioned previously does not cover the polynomials.

In this regard, we propose to replace the Talagrand's inequality with the generalized T_p (see Assumption $(\mathrm{T}_p, \mathrm{S})$ for the definition). Due to its general form, this new condition includes inequalities that are easy to verify (see Appendix E).

See also the Table 1 for a visual representation of the literature review.

1.3 Paper structure

The paper is organized as follows. In Section 2, we present the mathematical setting of our problem. We introduce the basic notions, describe the SVGD algorithm in population limit and list the necessary assumptions. The first part of Section 3 is devoted to the main result, which is a descent lemma for $\mathrm{KL}(\cdot \mid \pi)$ functional on the Wasserstein space. The second part contains the complexity results and discussions. In Section 4 we briefly discuss the finite particle regime. Section 5 presents examples of distributions from practical problems that satisfy our assumptions. Section 6 concludes the results of the paper and discusses possible future work. Finally in the Appendix, the reader may find the postponed proofs, intermediary technical lemmas and supplementary discussions.

2 Mathematical setting of the problem

2.1 Notations

The space of d-dimensional real vectors is denoted by $\mathbb{R}^d$, while the set of non-negative real numbers is denoted by $\mathbb{R}^+$. We will use uppercase and lowercase Latin letters for, respectively, random and deterministic vectors. Our target distribution π is defined on $\mathbb{R}^d$, which is enhanced with the Euclidean norm. The notation $\|\cdot\|$ will correspond to the ℓ_2 norm on $\mathbb{R}^d$ unless specified. Also, we will assume that π has the p-th moment, that is $\pi \in \mathcal{P}_p(\mathbb{R}^d)$. The Jacobian of a vector valued function $h(\cdot) = (h_1(\cdot), \dots, h_d(\cdot))^\top$ is a $d \times d$ matrix defined as

$$Jh(x) := \left(\partial_{x_i} h_j\right)_{i=1,j=1}^{d,d}.$$

The divergence of the vector valued function h is the trace of its Jacobian:

$$\operatorname{div} h(x) := \sum_{i=1}^{d} \partial_{x_i} h_i(x).$$

The Hessian of a real valued function $U : \mathbb{R}^d \to \mathbb{R}$ is defined as the following square matrix:

$$\nabla^2 h(x) := \left(\frac{\partial^2 h}{\partial_{x_i}\partial_{x_j}}(x)\right)_{i=1,j=1}^{d,d}.$$

For any Hilbert space $\mathcal{H}$, we denote by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ the inner product of $\mathcal{H}$ and by $\|\cdot\|_{\mathcal{H}}$ its norm. Let $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$. The set of couplings $\Gamma(\mu,\nu)$ is the set of all joint distributions defined on $\mathbb{R}^d \times \mathbb{R}^d$ having μ and ν as its marginals. The Wasserstein-p distance between two probability measures is defined as

$$W_p(\mu,\nu) := \inf_{\eta\in\Gamma(\mu,\nu)} \left[\int \|x-y\|^p \eta(\mathrm{d}x,\mathrm{d}y)\right]^{1/p}.$$

The Kullback-Leibler divergence is defined as

$$\mathrm{KL}(\mu \mid \nu) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu(x)}{\nu(x)}\right)\mu(\mathrm{d}x), & \text{if } \mu \ll \nu; \\ +\infty, & \text{otherwise.} \end{cases}$$

We will use the spectral and the Hilbert-Schmidt norms for matrices. For $M \in \mathbb{R}^{d\times d}$ they are respectively defined as

$$\|M\|_{op} := \sqrt{\lambda_{\max}(M^\top M)},$$
$$\|M\|_{HS} := \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} M_{i,j}^2}.$$

Here $\lambda_{\max}$ corresponds to the largest eigenvalue.

2.2 SVGD in the population limit

Let us present briefly Reproducing Kernel Hilbert Spaces (RKHS) and some of its essential properties. We refer the reader to (Steinwart and Christmann, 2008)[Chapter 4] for a detailed introduction. Let the map $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel and let $\mathcal{H}_0$ be its corresponding RKHS. This means that $\mathcal{H}_0$ consists of real-valued maps from $\mathbb{R}^d$ to $\mathbb{R}$, including the feature maps $\Phi(x) := k(x,\cdot) \in \mathcal{H}_0$, and the reproducing property is satisfied:

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_0}.$$

See the discussion on the reproducing property in Appendix E.1. Let $\mathcal{H}$ be the space of the d-dimensional maps $\{(f_1,\dots,f_d)^\top \mid f_i \in \mathcal{H}_0, i = 1,\dots,d\}$. For two vector functions $f = (f_1,\dots,f_d)^\top$ and $g = (g_1,\dots,g_d)^\top$ from $\mathcal{H}$, we define the scalar product as

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

Suppose that we have a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ and $\mathcal{H}_0$ is its corresponding RKHS. As described above, we construct the Hilbert space $\mathcal{H} = \mathcal{H}_0^d$. Our goal is to construct an iterative algorithm, that has its iterates in the space of probability measures $\mathcal{P}_p(\mathbb{R}^d)$. Each iterate is defined as a pushforward measure from the previous one in a way that it minimizes the KL distance the most. To do so, for every $\psi \in \mathcal{H}$ and $\gamma > 0$ let us define the operator

$$T_\gamma(x) := x - \gamma\psi(x).$$

The operator ψ will serve us as the direction or the perturbation, while as γ is the step-size. The goal is to choose the direction in which the KL error descends the most. Thus, for every μ, the optimal choice of ψ is the solution of the following problem:

$$g_\mu := \arg\max_{\|\psi\|_{\mathcal{H}}\le 1} \left\{-\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathrm{KL}\left(T_\gamma \# \mu \mid \pi\right)\Big|_{\gamma=0}\right\}. \quad (3)$$

Paper	Method	Smoothness	Other conditions	Can deal with higher-order polynomials?	Complexity to get ε error	Criterion
Vempala and Wibisono (2019)	LMC	L-smoothness	λ_{LS}-Log-Sobolev	✗	$\widetilde{\mathcal{O}}\left[\frac{d}{\lambda_{LS}^2 \varepsilon}\right]$	KL
Balasubramanian et al. (2022)	LMC	L-smoothness	-	✗	$\widetilde{O}\left[\frac{L^2 d^2}{\varepsilon^2}\right]$	FI
Chewi et al. (2021)	LMC	Hölder continuity	α-tails + Modified Log-Sobolev	✗	$\widetilde{O}\left[\frac{d^{(2/\alpha)(1+1/s)-1/s}}{\varepsilon^{1/s}}\right]$	KL
Chatterji et al. (2020)	LMC	(L,α)-Hölder continuity + Gaussian smoothing	Log-concavity	✗	$\widetilde{\mathcal{O}}\left[\frac{d^{(5-3\alpha)/2}}{\varepsilon^{4/(1+\alpha)}}\right]$	W_2
Korba et al. (2020)	SVGD	L-smoothness	Trajectory bounds	✗	$\mathcal{O}\left[\frac{d}{\varepsilon}\sqrt{C}\right]$	I_{Stein}
Salim et al. (2021)	SVGD	L-smoothness	T_1 inequality with constant λ_{T}	✗	$\widetilde{\mathcal{O}}\left[\frac{d^{3/2}}{\lambda_{\mathrm{T}}^{1/2}\varepsilon}\right]$	I_{Stein}
This paper	SVGD	(L_0,L_1) smoothness	T_p inequality with constant λ_{T}	✓	$\mathcal{O}\left[(\lambda_{\mathrm{T}}^{\frac{p}{2}}\varepsilon)^{-1}(pd)^{\frac{(p+1)(p+2)}{4}}\right]$	I_{Stein}
This paper	SVGD	(L_0,L_1) smoothness	Generalized $(\mathrm{T}_p,\mathrm{S})$	✓	$\mathcal{O}\left[\frac{\lambda_{BV}^p (pd)^{p+1}}{\varepsilon}\right]$	I_{Stein}

Table 1: This summarizes several previous results on the LMC and SVGD algorithms. Here $\tilde{\mathcal{O}}$ corresponds to the order without the log-polynomial factors. One has to bear in mind, that the complexities of the SVGD and LMC cannot be compared to each other as they use different error metrics. The complexity of (Korba et al., 2020)[Corollary 6] contains the trajectory bound constant C, which may depend on the dimension d. In the last row of the table, the $(\mathrm{T}_p,\mathrm{S})$ is satisfied for the function $\mathrm{S}(r)=\lambda_{BV}(r^{1/p}+(r/2)^{1/2p})$, where λ_{BV} is a constant that may depend on the dimension (see Section 2.3).

The functional objective depends linearly on the perturbation function ψ. Indeed, Liu and Wang (2016) have proved that

$$-\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathrm{KL}\left(T_\gamma\#\mu \mid \pi\right)\Big|_{\gamma=0} = \int_{\mathbb{R}^d}\big(-V(x)\psi(x) + \operatorname{div}\psi(x)\big)\mu(\mathrm{d}x), \tag{4}$$

where $\#$ is the push-forward of μ by T_γ. If ψ satisfies mild conditions, then the right-hand side of (4) is equal to zero if and only if the measures μ and π coincide (Stein, 1972). This property motivates to define a discrepancy measure called Stein discrepancy as follows:

$$I_{\text{Stein}}(\mu \mid \pi) := \max_{\|\psi\|_{\mathcal{H}}\leq 1}\left\{\int_{\mathbb{R}^d}\big(-V(x)\psi(x) + \operatorname{div}\psi(x)\big)\mu(\mathrm{d}x)\right\}^2.$$

Liu et al. (2016) have proved that the solution of the optimization problem (3) is given explicitly by

$$g_\mu(\cdot) = -\int_{\mathbb{R}^d}\left[\nabla\log\pi(x)k(x,\cdot) + \nabla_x k(x,\cdot)\right]\mu(\mathrm{d}x)$$

and $\|g_\mu\|_{\mathcal{H}} = \sqrt{I_{\text{Stein}}(\mu \mid \pi)}$. For the proof of this inequality, we refer the reader to the remark in Appendix B.1. In our case, I_{Stein} is often referred to as the squared Kernelized Stein Discrepancy (KSD) or the Stein-Fisher information. Under certain conditions (see Appendix E.4) on the target distribution π and the kernel k, the convergence in KSD implies weak convergence:

$$I_{\text{Stein}}(\mu_n \mid \pi) \underset{n\to\infty}{\longrightarrow} 0 \implies \mu_n \underset{n\to\infty}{\longrightarrow} \pi.$$

Remark 1. *Integration by parts yields the following formula for g_μ:*

$$g_\mu(\cdot) = \int_{\mathbb{R}^d}\nabla\log\left(\frac{\mu(x)}{\pi(x)}\right)k(x,\cdot)\mu(\mathrm{d}x).$$

This latter equality is the alternative definition of the optimal direction g_μ given by Korba et al. (2020). The proof of the remark can be found in Appendix E.5. Thus, we have determined the descent direction. In order to formulate the algorithm let us initialize our algorithm at $\mu_0 \in \mathcal{P}_p(\mathbb{R}^d)$. The $(i+1)$-th iterate μ_{i+1} is obtained by transferring μ_i with the map $I-\gamma g_{\mu_i}$:

$$\mu_{i+1} = (I-\gamma g_{\mu_i})\#\mu_i. \tag{SVGD}$$

In case when $k \in L_p(\mu)$, then $\mathcal{H}_0 \subset L_p(\mu)$ (see (Steinwart and Christmann, 2008)[Theorem 4.26]). This guarantees

that every iteration of the SVGD has a finite p-th moment, that is for every $i \in \mathbb{N}$, $\mu_i \in \mathcal{P}_p(\mathbb{R}^d)$. The rest of the section is devoted to the assumptions that we will use later in the analysis.

2.3 Assumptions

In order to perform SVGD we need the target π, in particular the gradient of its potential V and the kernel $k(\cdot,\cdot)$. Below we present the four main assumptions on π,V and k, that we use in the analysis. The first assumption is a relaxed form of smoothness condition for ∇V.

We study the SVGD algorithm under the so called (L_0, L_1) smoothness, which has been borrowed from optimization literature. In (Zhang et al., 2019a), the authors studied the convergence of the clipped gradient descent under this condition. The assumption goes as follows.

Assumption (L_0, L_1)**.** *The Hessian $\nabla^2 V$ of $V = -\log \pi$ is well-defined and $\exists L_0, L_1 \geq 0$ s.t.*

$$\left\|\nabla^2 V(x)\right\|_{op} \leq L_0 + L_1 \|\nabla V(x)\|, \tag{5}$$

for any $x \in \mathbb{R}^d$.

First let us notice that when $L_1 = 0$ the inequality (5) becomes the usual smoothness condition (2). On the other hand, let us consider the generalized normal distribution (also known as exponential power distribution) which is given by its density $\pi(x) \propto \exp(-\|x-a\|^\beta)$, where $\beta \geq 2$. In this case, Hessian $\nabla^2 V$ is not bounded by a constant in terms of the operator norm. However, it is easy to verify that V satisfies Assumption (L_0, L_1), which yields that this assumption is indeed a relaxation of the L-smoothness.

The next two assumptions are intended to replace the trajectory condition (Korba et al., 2020; Duncan et al., 2019) for polynomial potentials. The latter limits the applicability of the algorithm as these results cannot guarantee the convergence of the algorithm before actually implementing it. This condition was replaced by Talagrand's T_1 inequality in the recent work by Salim et al. (2021), where complexity bounds are proved for L-smooth potentials. Talagrand's T_p inequality goes as follows:

$$W_p(\mu, \pi) \leq \sqrt{\frac{2\,\mathrm{KL}(\mu \mid \pi)}{\lambda_{\mathrm{T}}}}, \quad \forall \mu \in \mathcal{P}_p(\mathbb{R}^d). \tag{6}$$

Since the Wasserstein-p distance is increasing with respect to the order p, T_p implies T_1. In general, it is hard to verify if the distribution satisfies the T_p inequality. However, one may check that T_2 is true for m-strongly log-concave distributions with $\lambda_{\mathrm{T}} = m$. According to Bakry-Émery theory, sharp estimates on this constant are available due to perturbation arguments, such as the well-known Holley-Stroock method (see Steiner (2021)). When it comes to bounding the Stein-Fisher information for (L_0, L_1) potentials, one needs to bound the p-th Wasserstein error with $p > 2$. Thus, it imposes having a functional inequality involving W_p and KL. The generalized T_p inequality is a general form of this condition.

Definition 1 (Generalized T_p inequality)**.** *Let $p \geq 1$. The distribution π satisfies the generalized T_p inequality if there exists an increasing function $\mathrm{S} : \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, we have $W_p(\mu, \pi) \leq \mathrm{S}(\mathrm{KL}(\mu|\pi))$, where W_p is the Wasserstein-p distance.*

Assumption $(\mathrm{T}_p, \mathrm{S})$**.** *The target distribution π satisfies the generalized T_p inequality for some increasing function $\mathrm{S} : \mathbb{R}^+ \to \mathbb{R}^+$.*

If we take $\mathrm{S}(r) \equiv \sqrt{2r/\lambda_{\mathrm{T}}}$,for every $r \in \mathbb{R}^+$, then we retrieve the classical T_p inequality. Thus, we indeed generalize T_p inequality with this assumption. An important example of Assumption $(\mathrm{T}_p, \mathrm{S})$ is a consequence of Bolley and Villani (2005)[Corollary 2.3]. They prove that if

$$\int_{\mathbb{R}^d} \exp(s\|x - x_0\|^p)\pi(\mathrm{d}x) < +\infty,$$

for some $x_0 \in \mathbb{R}^d$ and $s > 0$, then Assumption $(\mathrm{T}_p, \mathrm{S})$ is satisfied for $\mathrm{S}(r) = \lambda_{BV}(r^{1/p} + (r/2)^{1/2p})$, where λ_{BV} is a constant that may depend on the dimension (see Appendix E.2 for details). Therefore, in this case the assumption is essentially a condition on the tails of the target distribution π. In particular, for the generalized normal distribution the verification of this bound is straightforward.

The third assumption is chosen to essentially restrict ourselves for potentials with (at most) polynomial growth. Mathematically, it is expressed as follows:

Assumption $(\mathrm{poly}, \mathrm{Q})$**.** *For some $p > 0$, there exists a polynomial with positive coefficients such that $\mathrm{ord}(\mathrm{Q}) = p$ and the following inequality is true:*

$$\left\|\nabla V(x)\right\| \leq \mathrm{Q}(\|x\|).$$

We want to emphasize that this assumption is not very restrictive as the vast majority of continuous distributions that appear in practice satisfy it. Using Taylor formula, one may check that L-smooth potentials satisfy this condition with $\mathrm{Q}(r) = Lr + \|\nabla V(0)\|$. It is important to notice that the constant p is the same for the assumptions $(\mathrm{T}_p, \mathrm{S})$ and $(\mathrm{poly}, \mathrm{Q})$. This will allow us to treat the polynomials of order p. For detailed examples of distributions that satisfy our set of assumptions please refer to Section 5.

We conclude our list of assumptions with a bound on the kernel $k(\cdot,\cdot)$.

Assumption $(\ker, B)$**.** *There exists $B > 0$ such that $\|k(x,\cdot)\|_{\mathcal{H}_0} \leq B$ and*

$$\|\nabla_x k(x,\cdot)\|_{\mathcal{H}} = \left(\sum_{i=1}^d \|\partial_{x_i} k\,(x,\cdot)\|^2_{\mathcal{H}_0}\right)^{\frac{1}{2}} \leq B,$$

for all $x \in \mathbb{R}^d$.

Due to the reproducing property, these conditions are equivalent to $k(x,x) \le B$ and $\partial^2_{x_i,x_i} k(x,x) \le B^2$. Here, the first and the second partial derivatives are operated, respectively, on the first and the second variables of $k(\cdot,\cdot)$. Based on this criterion, one may check that the multiquadratic kernel $k(x,y) := (c^2 + \|x-y\|^2)^\beta$ for some $c > 0 > \beta > -1$ satisfies the Assumption $(\ker, B)$. See Appendix E.1 for details on the reproductive property.

3 Main results

In this section we present our main results. We start with a proposition that bounds the difference between the value of the objective at two consecutive iterations. The proof of the proposition can be found in Appendix B.1

Proposition 1. *Suppose that Assumptions* $(\ker, B)$ *and* (L_0, L_1) *are satisfied. Let* $\alpha > 1$ *and choose*

$$\gamma \le \frac{(\alpha-1)\cdot \min\{1, 1/L_1\}}{\alpha B \|g_{\mu_n}\|_{\mathcal{H}}}. \tag{7}$$

Then

$$\begin{aligned}\mathrm{KL}\left(\mu_{n+1} \mid \pi\right) - \mathrm{KL}\left(\mu_n \mid \pi\right) \le -\gamma\Big[1 - \frac{\gamma}{2}B^2 \\ \times \left(\alpha^2 + (e-1)A_n\right)\Big] I_{\mathrm{Stein}}(\mu_n \mid \pi),\end{aligned} \tag{8}$$

where $A_n := L_0 + L_1 \mathbb{E}_{X\sim\mu_n}[\|\nabla V(X)\|]$.

Proposition 1 may be seen as a descent lemma on the KL divergence. Let us develop the condition (7). The following lemma is from Salim et al. (2021) and provides us with an upper bound on $\|g_{\mu_n}\|_{\mathcal{H}}$.

Lemma 1. *If Assumption* $(\ker, B)$ *is satisfied, then*

$$\|g_{\mu_n}\|_{\mathcal{H}} = I_{\mathrm{Stein}}(\mu_n \mid \pi)^{\frac{1}{2}} \le B\left(\mathbb{E}_{X\sim\mu_n}\left[\|\nabla V(X)\|\right] + 1\right)$$

for all $n \in \mathbb{N}$.

Thus, we can get replace $\|\cdot\|_{\mathcal{H}}$ by expectation over μ_k in the bound for γ in (7):

$$\gamma \le \frac{(\alpha-1)}{\max\{1, L_1\}}\left[\alpha B^2\left(\mathbb{E}_{X\sim\mu_n}\left[\|\nabla V(X)\|\right] + 1\right)\right]^{-1}. \tag{9}$$

On the other hand, in order to guarantee that the functional $\mathcal{F}$ is decreasing with each iteration of the algorithm, we would like the right-hand side of (8) to be negative. The latter yields the following condition on the step-size:

$$\begin{aligned}\gamma &< \frac{2}{B^2}\left(\alpha^2 + (e-1)A_n\right)^{-1} \\ &= \frac{2}{B^2}\left(\alpha^2 + (e-1)\left(L_0 + L_1\mathbb{E}_{X\sim\mu_n}[\|\nabla V(X)\|]\right)\right)^{-1}.\end{aligned} \tag{10}$$

Both in (9) and (10) we have dependence on $\mathbb{E}_{X\sim\mu_n}[\|\nabla V(X)\|]$. The following lemma will allow us to establish an upper bound on this quantity and which will help us to get rid of trajectory dependent bounds on the step-size.

Lemma 2. *Suppose that the potential function* V *satisfies the Assumptions* $(\mathrm{T}_p, \mathrm{S})$ *and* $(\mathrm{poly}, \mathrm{Q})$ *for some constants* p *and a polynomial* Q. *Then,*

$$\mathbb{E}_{X\sim\mu_n}[\|\nabla V(X)\|] \le \mathrm{Q}\left(\mathrm{S}\big(\mathrm{KL}(\mu_n \mid \pi)\big) + W_p(\pi, \delta_0)\right).$$

The important implication of Lemma 2 is that both upper bounds on the step-size are inversely proportional to $\mathrm{KL}(\mu_n \mid \pi)$. Let us now state the main theorem.

Theorem 1 (Descent lemma)**.** *Let the target distribution* π *and its potential function* V *satisfy the Assumptions* $(\ker, B)$, (L_0, L_1), $(\mathrm{T}_p, \mathrm{S})$, *and* $(\mathrm{poly}, \mathrm{Q})$. *Define* $C_0 := \mathrm{Q}\left(\mathrm{S}\big(\mathrm{KL}(\mu_0 \mid \pi)\big) + W_p(\pi, \delta_0)\right)$ *and suppose that for some* $\alpha > 1$ *the step-size* γ *satisfies*

$$\gamma < \frac{\alpha - 1}{\alpha B^2\big(\alpha^2 + (e-1)(\max(L_0, L_1, 1) + \max(L_1, 1)C_0)\big)}. \tag{11}$$

Then for every $n = 0, 1, \ldots$ *the following inequality is true:*

$$\mathrm{KL}\left(\mu_{n+1} \mid \pi\right) - \mathrm{KL}\left(\mu_n \mid \pi\right) \le -\frac{\gamma}{2} I_{\mathrm{Stein}}(\mu_n \mid \pi).$$

The proof is postponed to Appendix B.2. It is shown in Mattingly et al. (2002); Hutzenthaler et al. (2012) that ULA (unadjusted Langevin algorithm) is not stable if ∇V grows super-linearly, i.e. $\liminf_{\|x\|\to\infty} \frac{\|\nabla V(x)\|}{\|x\|} \to \infty$ (Mattingly et al. (2002, Lemma 6.3) showed a counterexample with $V(x) = \frac{x^4}{4}, x \in \mathbb{R}$). However, the above theorem shows that, at least in the population limit, SVGD could be stable even when ∇V grows super-linearly. For an illustration of this point, see the experiments in Appendix F.

Corollary 1 (Convergence)**.** *Let the assumptions of Theorem 1 be satisfied. Then following statements are true.*

1. $I_{\mathrm{Stein}}(\mu_n \mid \pi)$ *converges to zero, when* $n \to \infty$.
2. *The average Stein-Fisher error of the first* n *iterates converges to 0, with* $O(1/n)$ *rate:*

$$\frac{1}{n}\sum_{i=0}^{n-1} I_{\mathrm{Stein}}\left(\mu_i \mid \pi\right) \le \frac{2}{n\gamma}\mathrm{KL}(\mu_0 \mid \pi).$$

This yields in particular that the series with the general term $I_{\mathrm{Stein}}(\mu_n \mid \pi)$ is convergent, that is $I_{\mathrm{Stein}}(\mu_n \mid \pi) \to 0$. When π is distantly dissipative and the k is the inverse multiquadratic kernel, then convergence in I_{Stein} yields weak convergence of the sequence μ_n to the target π (see Appendix E.4). Proofs of the corollary can be found in Appendix B.3. If we bound the initial KL error, then for every

precision error $\varepsilon > 0$, Corollary 1 provides us with a stopping rule. Based on this scheme, the next theorem estimates the complexity of the algorithm for some particular choices of S and μ_0.

Theorem 2. *Let assumptions* $(\ker, B)$, (L_0, L_1), *and* $(\mathrm{poly}, \mathrm{Q})$ *hold and let* $\mu_0 = \mathcal{N}(0, \mathrm{I}_d)$. *Then in order to have* ε *average Stein-Fisher error it is sufficient to perform* n *iterations of the SVGD, where*

- $n = \mathcal{O}\left(\varepsilon^{-1} Q(1)^3 \max(L_1, 1) \lambda_{BV}^p (pd)^{p+1}\right)$, *if the target* π *satisfies Assumption* $(\mathrm{T}_p, \mathrm{S})$ *with* $\mathrm{S}(r) = \lambda_{BV}(r^{1/p} + (r/2)^{1/2p})$;
- $n = \mathcal{O}\left(\varepsilon^{-1} \mathrm{Q}(1)^{\frac{p+2}{2}} \max(L_1, 1) \lambda_{\mathrm{T}}^{-\frac{p}{2}} (pd)^{\frac{(p+1)(p+2)}{4}}\right)$, *if the target* π *satisfies Talagrand's* T_p *inequality* (6) *with constant* λ_{T}.

The proof is postponed to Appendix C.1. We observe that in both cases we achieve polynomial convergence of the algorithm in terms of the dimension and the precision. As mentioned in Section 2, in the first setting, $(\mathrm{T}_p, \mathrm{S})$ essentially becomes a tail condition, which can be easily verified for a wide class of log-polynomial densities. The Talagrand's T_p is harder to verify in the general case. However, unlike λ_{T}, the constant λ_{BV} may be dimension dependent (see Appendix E.2). Therefore, when $p \leq 2$ Talagrand's inequality yields better convergence. In particular, when $p = 1$, we recover the complexity $\mathcal{O}(d^{3/2})$ which coincides with the result from (Salim et al., 2021). On the other hand, when p is large, the algorithm is likely to perform significantly faster under the generalized T_p.

4 Finite-particle regime

Up to this point, we have only discussed the SVGD in the population limit. The latter is an iterative algorithm on the space of measures and, hence, is not applicable in practical settings. In this section we present its discretization.

The finite-particle SVGD algorithm approximates each iterate μ_k by a mixture of N Dirac distributions $\hat{\mu}_k = \frac{1}{N}\sum_i \delta_{x_i^k}$. Then at each iteration k the optimal perturbation g_{μ_k} can be approximated empirically by

$$\hat{g}_k(x) := \frac{1}{N}\sum_{i=1}^{N}\left[-k(x_i^k, x)\nabla V(x_i^k) + \nabla_{x_i^k} k\left(x_i^k, x\right)\right].$$

Using $\hat{g}_k(x)$ we can design an algorithm for finite number of particles evolving as an approximation of SVGD in population limit. The pseudo-code is presented in Algorithm 1.

Liu (2017) proved the weak convergence of the empirical distribution $\hat{\mu}_k$ to the target π, when the number of particles goes to infinity (see also (Lu et al., 2019; Nüsken and Renger, 2021)). Liu and Wang (2018) studied the algorithm

Algorithm 1 Finite-particle SVGD

Input: Initialize a set of particles $\{x_i^0\}_{i=1}^N$, step-size γ, number of iterations K
for $k = 0, 1, 2, \ldots, K-1$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    Compute: $\hat{g}_k(x_i^k)$
    $x_i^{k+1} \leftarrow x_i^k + \gamma \hat{g}_k(x_i^k)$
  end for
end for
Return: $\{x_i^K\}_{i=1}^N$

using its fixed point properties. If a finite set of particles X is a fixed point of the SVGD, then they showed that

$$\frac{1}{\mathsf{card}(\mathsf{X})}\sum_{x\in\mathsf{X}} f(x) = \mathbb{E}_{Y\sim\pi}\big[f(Y)\big],$$

where f is belongs to certain class of functions called Stein's matching set. The latter set depends on the linear span of the feature maps $k(x, \cdot)$. Their analysis allows to show that for a certain choice of kernels SVGD can exactly compute the first and second order moments for multivariate Gaussians. In a recent paper, Shi and Mackey (2022) prove convergence for sub-Gaussian distributions with L-smooth potentials under strong conditions. Although the obtained rate is very slow, their analysis might be served as a general scheme for better results.

Overall, getting convergence results for the finite-particle regime that match the empirical findings remains an open problem for future work.

5 Examples

Below, we describe several sampling problems that fall into our setting. The family of generalized Gaussian distributions satisfy our set of assumptions:

$$\pi(x) \propto \exp\left(-\frac{\|x-a\|^p}{2\sigma^p}\right).$$

This family of distributions has received considerable attention from the engineering community, due to its flexible parametric form. For instance values of $p = 2.2$ and $p = 1.6$ have been found to model the ship transit noise and the sea surface agitation noise respectively Banerjee and Agrawal (2013). Many important probabilistic quantities of GGs, such as the moments, the entropy etc., are easy to compute. It is known that that this class minimizes the entropy under a p-th absolute moment constraint (Cover and Thomas, 2006). An overview of the analytic properties of this family of distributions can be found in Dytso et al. (2018).

Another model that falls into our framework was proposed by Ginzburg and Landau (Livingstone et al., 2019; Brosse

et al., 2019). This model of phase transitions in physics is defined on a three-dimensional $d = p^3$ lattice for $p \in \mathbb{N}^*$. The potential function is given for $x = (x_{ijk})_{i,j,k\in\{1,\dots,p\}} \in \mathbb{R}^d$ by

$$V(x) = \sum_{i,j,k=1}^{p} \left\{ \frac{1-\tau}{2} x_{ijk}^2 + \frac{\tau\alpha}{2} \left\| \tilde{\nabla} x_{ijk} \right\|^2 + \frac{\tau\lambda}{4} x_{ijk}^4 \right\},$$

where $\alpha, \lambda, \tau > 0$ and $\tilde{\nabla} x_{ijk}$ is the discrete gradient defined as

$$\tilde{\nabla} x_{ijk} := \left(x_{i_+jk} - x_{ijk}, x_{ij_+k} - x_{ijk}, x_{ijk_+} - x_{ijk} \right)$$

with $i_\pm = i \pm 1 \pmod p$ and similarly for $j_\pm, k_\pm$. The inference problem in this setting is to estimate the expectation of $\pi(x) \propto \exp(-V(x))$. Based on its definition, we deduce immediately that the π satisfies our assumptions, thus the SVGD algorithm is applicable in this case.

Bayesian approach to solve the tensor PCA (Richard and Montanari, 2014) was proposed by Lesieur et al. (2017). A simple instance of this problem can be formulated as follows. Consider the model $Y = \lambda X^{\otimes k} + W$, where $Y, W \in \otimes^k \mathbb{R}^d$ are tensors, $X^{\otimes k}$ is the k-th tensor power of the vector $X \in \mathbb{R}^d$ and $\lambda > 0$. We observe the tensor Y and the goal is to infer the unknown vector X. Suppose for simplicity that the noise tensor W has i.i.d. $\mathcal{N}(0,1)$ entries. Then assuming that $X \sim \pi_0$ has a log-polynomial prior distribution, we get the following posterior:

$$\pi(x) \propto e^{-\sum_{i_1,i_2,\dots,i_k} \left(\lambda x_{i_1} x_{i_2} \dots x_{i_k} - Y_{i_1 i_2 \dots i_k}\right)^2} \cdot \pi_0(x).$$

One could use the Bayesian posterior mean to estimate the signal X. In order to do it, sampling methods can be applied. In particular, the distribution π satisfies our assumptions. Thus one could use the SVGD algorithm to sample from it. We would also like to draw the reader's attention that up to our knowledge this is the first attempt to solve the Bayesian PCA using sampling methods. It remains an interesting problem to examine in future work.

6 Conclusion and Future work

In this work, we studied the convergence of the SVGD under a relaxed smoothness condition. The latter was first proposed in the context of gradient clipping for the optimization problem. This relaxed smoothness condition allows to treat distributions with polynomial potentials. Main result consists of a descent lemma for the KL error of the SVGD algorithm. The result implies polynomial convergence for the average Stein-Fisher information error for two types of functional inequalities.

One could also try to use general analytic functions instead of polynomials. Then approximating this function by its Taylor series, one would be able to bound the Stein-Fisher information using the generalized T_p inequality. However, to the best of our knowledge, most distributions satisfy $(poly, Q)$. The latter means that $(poly, Q)$ does not impose a significant limitation. Generalizing the assumption for analytic functions would complicate the readability of the paper and would perhaps diverge from the main contribution.

Numerous applications of the SVGD and its modifications have been developed in the literature. However, the discretized algorithm does not have practical convergence guarantees in the general case. Despite the performed simple experiments, this paper is of purely theoretical interest as the algorithm operates on the space of probability distributions. We want to underline the fact that the experiments describe only the convergence of the discretized method. The theoretical analysis of the discretized regime remains an open problem to be studied in the future work.

References

Achim, A., Loza, A., Bull, D., and Canagarajah, N. (2008). Statistical modelling for wavelet-domain image fusion. *Image Fusion*, page 119.

Ahn, K. and Chewi, S. (2021). Efficient constrained sampling via the mirror-langevin algorithm. *Advances in Neural Information Processing Systems*, 34.

Amann, H. (2011). *Ordinary differential equations: an introduction to nonlinear analysis*, volume 13. Walter de gruyter.

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. (2022). Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. In *Conference on Learning Theory*, pages 2896–2923. PMLR.

Banerjee, S. and Agrawal, M. (2013). Underwater acoustic noise with generalized Gaussian statistics: Effects on error performance. In *2013 MTS/IEEE OCEANS-Bergen*, pages 1–8. IEEE.

Bhattacharya, R. (1978). Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *The Annals of Probability*, pages 541–553.

Bolley, F. and Villani, C. (2005). Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pages 331–352.

Brosse, N., Durmus, A., Moulines, É., and Sabanis, S. (2019). The tamed unadjusted Langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663.

Cai, X., McEwen, J. D., and Pereyra, M. (2021). High-dimensional Bayesian model selection by proximal nested sampling. *arXiv preprint arXiv:2106.03646*.

Chatterji, N., Diakonikolas, J., Jordan, M. I., and Bartlett, P. (2020). Langevin Monte Carlo without smoothness. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1716–1726. PMLR.

Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L., and Jordan, M. I. (2018). On the theory of variance reduction for stochastic gradient Monte Carlo. *arXiv preprint arXiv:1802.05431*.

Cheng, X. and Bartlett, P. L. (2018). Convergence of Langevin MCMC in KL-divergence. *PMLR 83*, (83):186–211.

Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. (2018). Sharp convergence rates for Langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*.

Chewi, S., Erdogdu, M. A., Li, M. B., Shen, R., and Zhang, M. (2021). Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *arXiv preprint arXiv:2112.12662*.

Chewi, S., Le Gouic, T., Lu, C., Maunu, T., Rigollet, P., and Stromme, A. (2020). Exponential ergodicity of mirror-Langevin diffusions. *Advances in Neural Information Processing Systems*, 33:19573–19585.

Cover, T. M. and Thomas, J. A. (2006). Elements of information theory. *Wiley series in telecommunications and signal processing*.

Dalalyan, A. S. (2017a). Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory (COLT)*, pages 678–689.

Dalalyan, A. S. (2017b). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676.

Dalalyan, A. S. and Karagulyan, A. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311.

Dalalyan, A. S., Karagulyan, A., and Riou-Durand, L. (2022). Bounding the error of discretized langevin algorithms for non-strongly log-concave targets. *Journal of Machine Learning Research*, 23(235):1–38.

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

Duncan, A., Nüsken, N., and Szpruch, L. (2019). On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*.

Durmus, A., Majewski, S., and Miasojedow, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46.

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

Durmus, A. and Moulines, E. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882.

Durmus, A., Moulines, E., and Pereyra, M. (2018). Efficient Bayesian computation by proximal Markov Chain Monte Carlo: when Langevin meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506.

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2018). Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on Learning Theory*, pages 793–797. PMLR.

Dytso, A., Bustin, R., Poor, H. V., and Shamai, S. (2018). Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1):1–40.

Erdogdu, M. A. and Hosseinzadeh, R. (2021). On the convergence of Langevin Monte Carlo: the interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR.

Feroz, F. and Skilling, J. (2013). Exploring multi-modal distributions with nested sampling. In *AIP Conference Proceedings*, volume 1553, pages 106–113. American Institute of Physics.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, pages 1292–1301.

Gorham, J., Raj, A., and Mackey, L. (2020). Stochastic Stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581.

Hutzenthaler, M., Jentzen, A., and Kloeden, P. E. (2012). Strong convergence of an explicit numerical method for SDEs with non-globally Lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4):1611–1641.

Karagulyan, A. and Dalalyan, A. (2020). Penalized Langevin dynamics with vanishing penalty for smooth and log-concave targets. *Advances in Neural Information Processing Systems*, 33.

Kassab, R. and Simeone, O. (2022). Federated generalized Bayesian learning via distributed Stein variational gradient descent. *IEEE Transactions on Signal Processing*.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. (2020). A non-asymptotic analysis for Stein variational gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4672–4682.

Lehec, J. (2021). The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *arXiv preprint arXiv:2101.10695*.

Lesieur, T., Miolane, L., Lelarge, M., Krzakala, F., and Zdeborová, L. (2017). Statistical and computational phase transitions in spiked tensor estimation. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 511–515. IEEE.

Li, L., Li, Y., Liu, J.-G., Liu, Z., and Lu, J. (2020). A stochastic version of Stein variational gradient descent for efficient sampling. *Communications in Applied Mathematics and Computational Science*, 15(1):37–63.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3115–3123.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, pages 276–284.

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

Liu, Q. and Wang, D. (2018). Stein variational gradient descent as moment matching. *Advances in Neural Information Processing Systems*, 31.

Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017). Stein variational policy gradient. *arXiv preprint arXiv:1704.02399*.

Livingstone, S., Faulkner, M. F., and Roberts, G. O. (2019). Kinetic energy choice in hamiltonian/hybrid monte carlo. *Biometrika*, 106(2):303–319.

Lu, J., Lu, Y., and Nolen, J. (2019). Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671.

Ma, Y.-A., Chatterji, N., Cheng, X., Flammarion, N., Bartlett, P. L., and Jordan, M. I. (2019). Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*.

Mattingly, J. C., Stuart, A. M., and Higham, D. J. (2002). Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232.

Neal, R. (1992). Bayesian learning via stochastic dynamics. *Advances in neural information processing systems*, 5.

Nüsken, N. and Renger, D. (2021). Stein variational gradient descent: many-particle and long-time asymptotics. *arXiv preprint arXiv:2102.12956*.

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. (2017). VAE learning via Stein variational gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4236–4245.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Nonconvex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory (COLT)*, pages 1674–1703.

Richard, E. and Montanari, A. (2014). A statistical model for tensor PCA. *Advances in neural information processing systems*, 27.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268.

Roberts, G. O. and Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and computing in applied probability*, 4(4):337–357.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Sabanis, S. (2013). A note on tamed Euler approximations. *Electronic Communications in Probability*, 18:1–10.

Salim, A., Sun, L., and Richtárik, P. (2021). Complexity analysis of Stein variational gradient descent under Talagrand's inequality T1. *arXiv preprint arXiv:2106.03076*.

Shen, R. and Lee, Y. T. (2019). The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2100–2111.

Shi, J., Liu, C., and Mackey, L. (2021). Sampling with mirrored Stein operators. *arXiv preprint arXiv:2106.12506*.

Shi, J. and Mackey, L. (2022). A finite-particle convergence rate for stein variational gradient descent. *arXiv preprint arXiv:2211.09721*.

Şimşekli, U. (2017). Fractional Langevin Monte Carlo: Exploring Lévy driven stochastic differential equations for Markov chain Monte Carlo. In *International Conference on Machine Learning*, pages 3200–3209. PMLR.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, volume 6, pages 583–603. University of California Press.

Steiner, C. (2021). A Feynman-Kac approach for logarithmic Sobolev inequalities. *Electronic Journal of Probability*, 26:1–19.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Tao, C., Dai, S., Chen, L., Bai, K., Chen, J., Liu, C., Zhang, R., Bobashev, G., and Carin, L. (2019). Variational annealing of GANs: A Langevin perspective. In *International Conference on Machine Learning (ICML)*, pages 6176–6185.

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8092–8104.

Wang, D., Tang, Z., Bajaj, C., and Liu, Q. (2019). Stein variational gradient descent with matrix-valued kernels. *Advances in neural information processing systems*, 32.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory (COLT)*, pages 2093–3027.

Wibisono, A. (2019). Proximal Langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*.

Winkelbauer, A. (2012). Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. (2019a). Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations (ICLR)*.

Zhang, R., Li, C., Chen, C., and Carin, L. (2018). Learning structural weight uncertainty for sequential decision-making. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1137–1146.

Zhang, R., Wen, Z., Chen, C., Fang, C., Yu, T., and Carin, L. (2019b). Scalable Thompson sampling via optimal transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 87–96.

Zhou, Q., Yu, T., Zhang, X., and Li, J. (2020). Bayesian inference and uncertainty quantification for medical image reconstruction with Poisson data. *SIAM Journal on Imaging Sciences*, 13(1):29–52.

Appendix

Contents

A Langevin Monte-Carlo from the optimization perspective

Langevin algorithms have extensively been studied by statistics and machine learning communities during the last decades. The algorithms rely on an SDE called (vanilla) Langevin dynamics:

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t,$$

where B_t is a d-dimensional Brownian motion on $\mathbb{R}^d$. The important property of this SDE is that under technical conditions the target $\pi(\cdot) \propto \exp(-V(\cdot))$ is its unique invariant distribution. Moreover, it is ergodic and $\mathrm{KL}(\mu_t \mid \pi)$ converges linearly to zero, where μ_t is the distribution of X_t (Bhattacharya, 1978). The Euler-Maruyama discretization of this SDE over the time axis results in the Langevin Monte-Carlo (LMC) algorithm:

$$\theta_{i+1} = \theta - \gamma\nabla V(\theta_i) + \sqrt{2h}\xi_{i+1}, \text{ for } i \in \mathbb{N}. \tag{12}$$

Here $\gamma > 0$ is the discretization step-size and $\xi_1, \xi_2, \dots$ is a sequence of independent standard Gaussians independent from the initial point θ_0. This algorithm initially was studied by Roberts and Tweedie (1996); Roberts and Rosenthal (1998); Roberts and Stramer (2002); Dwivedi et al. (2018) who proposed to apply Metropolis-Hastings step (MALA). The reason for this adjustment is that the constant step-size γ induces a bias and the target distribution π is no longer the invariant distribution of the discrete-time process. Later, Dalalyan (2017b) suggested to remove the Metropolis-Hastings step with a result that essentially controls the bias depending on the step-size γ. This instigated a new line of research which studied the convergence properties of the LMC in different settings (see Durmus and Moulines (2017); Cheng et al. (2018); Cheng and Bartlett (2018); Dalalyan and Karagulyan (2019); Durmus and Moulines (2019); Vempala and Wibisono (2019)).

The LMC algorithm essentially performs a Forward-Flow iteration in the space of distributions (see (Wibisono, 2018; Durmus et al., 2019)). Indeed, if we define by μ_i the distribution of the i-th iterate θ_i, then (12) is equivalent to

$$\begin{aligned}\mu_{i+1/2} &= \big(I - \gamma\nabla V(\mu_i)\big)\#\mu_i\\ \mu_{i+1} &= \mathcal{N}(0, \mathrm{I}_d) * \mu_{i+1/2},\end{aligned}$$

where the $\#$ is the pushforward measure, while $*$ is the convolution. The first equation corresponds to the gradient descent for $\mathbb{E}_{X\sim\mu}[V(X)]$ where the argument is the measure μ. The second equation can be interpreted as the exact gradient flow of the negative entropy. The combination of this two steps results in a biased algorithm (as noticed in (Roberts and Tweedie, 1996)) since the flow step is not the adjoint of the gradient step.

B Convergence results

B.1 Proof of Proposition 1

Here we follow the procedure from Korba et al. (2020). The main difference is to upper bound the term ψ_2 (see (18)) under (L_0, L_1)-smoothness condition. For the sake of brevity we will omit the index when referring to g_{μ_n} and will write simply g. Let us define $\phi_t := I - tg$ and $\rho_t := \phi_{t\#}\mu_n$ for every $t \in [0, \gamma]$. Then, applying Taylor formula to the function

$$\varphi(t) := \mathrm{KL}\left(\rho_t \mid \pi\right) \tag{13}$$

we have the following

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)\mathrm{d}t. \tag{14}$$

By the definition of the SVGD iteration, we have that $\varphi(0) = \mathrm{KL}\left(\mu_n \mid \pi\right)$ and $\varphi(\gamma) = \mathrm{KL}\left(\mu_{n+1} \mid \pi\right)$. Let us now compute the term of (14) corresponding to the first order derivative.

Lemma 3. *Suppose that Assumption* $(\ker, B)$ *holds. Then, for any* $x \in \mathbb{R}^d$ *and* $h \in \mathcal{H}$,

$$\|Jh(x)\|_{\mathrm{HS}} \le B\|h\|_{\mathcal{H}}.$$

The proof of the lemma can be found in Appendix D.3. Applying the lemma to the function g, we obtain the following:

$$\|tJg(x)\|_{op} \le \|tJg(x)\|_{\mathrm{HS}} \le tB\|g\|_{\mathcal{H}} < 1. \tag{15}$$

The latter inequality is due to the condition on the step-size γ. The bound (15) implies that ϕ_t is a diffeomorphism. Therefore, ρ_t admits a density given by the change of variables formula:

$$\rho_t(x) = \left|J\phi_t\left(\phi_t^{-1}(x)\right)\right|^{-1} \mu_n\left(\phi_t^{-1}(x)\right).$$

Changing the variable of integration and applying the transfer lemma we get the following formula for $\varphi(t)$:

$$\begin{aligned}
\varphi(t) &= \int \log\left(\frac{\rho_t(y)}{\pi(y)}\right) \rho_t(\mathrm{d}y) \\
&= \int \log\left(\frac{\mu_n(x)\left|J\phi_t(x)\right|^{-1}}{\pi\left(\phi_t(x)\right)}\right) \mu_n(\mathrm{d}x) \\
&= \int \left[\log\left(\mu_n(x)\right) + \log\left(\left|J\phi_t(x)\right|^{-1}\right) - \log\left(\pi\left(\phi_t(x)\right)\right)\right] \mu_n(\mathrm{d}x).
\end{aligned}$$

Let us then compute the time derivative of $\varphi(t)$. Taking the derivative inside and applying Jacobi's formula for matrix determinant differentiation we obtain the following equality:

$$\varphi'(t) = -\int \operatorname{tr}\left(\left(J\phi_t(x)\right)^{-1} \frac{dJ\phi_t(x)}{dt}\right) \mu_n(\mathrm{d}x) - \int \left\langle \nabla \log \pi\left(\phi_t(x)\right), \frac{d\phi_t(x)}{dt}\right\rangle \mu_n(\mathrm{d}x).$$

By definition, $d\phi_t/dt = g$. Therefore, we can use the explicit expression of ϕ_t to write:

$$\varphi'(t) = \int \operatorname{tr}\left(\left(J\phi_t(x)\right)^{-1} Jg(x)\right) \mu_n(\mathrm{d}x) + \int \left\langle \nabla V\left(\phi_t(x)\right), g(x)\right\rangle \mu_n(\mathrm{d}x). \tag{16}$$

The Jacobian at time $t = 0$ is simply equal to the identity matrix since $\phi_0 = \mathrm{I}_d$. It follows that $\operatorname{tr}\left(\left(J\phi_0(x)\right)^{-1} Jg(x)\right) = \operatorname{tr}(Jg(x)) = \operatorname{div}(g)(x)$ by the definition of the divergence operator. Using integration by parts:

$$\begin{aligned}
\varphi'(0) &= -\int [-\operatorname{div}(g)(x) - \langle \nabla \log \pi(x), g(x)\rangle] \mu_n(\mathrm{d}x) \\
&= -\int \left\langle \nabla \log\left(\frac{\mu_n}{\pi}\right)(x), g(x)\right\rangle \mu_n(\mathrm{d}x).
\end{aligned}$$

Based on the alternative definition of g_μ (see Remark 1) and the reproducing property, we have

$$\begin{aligned}
&\int \left\langle \nabla \log\left(\frac{\mu_n}{\pi}\right)(x), g(x)\right\rangle \mu_n(\mathrm{d}x) \\
&\qquad = \iint k(x,y) \left\langle \nabla \log\left(\frac{\mu_n}{\pi}\right)(x), \nabla \log\left(\frac{\mu_n}{\pi}\right)(y)\right\rangle \mu_n(\mathrm{d}x)\mu_n(\mathrm{d}y) \\
&\qquad = \iint \langle k(x,\cdot), k(y,\cdot)\rangle_{\mathcal{H}_0} \left\langle \nabla \log\left(\frac{\mu_n}{\pi}\right)(x), \nabla \log\left(\frac{\mu_n}{\pi}\right)(y)\right\rangle \mu_n(\mathrm{d}x)\mu_n(\mathrm{d}y) \\
&\qquad = \left\langle \int \nabla \log\left(\frac{\mu_n}{\pi}\right)(x) k(x,\cdot) \mu_n(\mathrm{d}x), \int \nabla \log\left(\frac{\mu_n}{\pi}\right)(y) k(y,\cdot) \mu_n(\mathrm{d}y)\right\rangle_{\mathcal{H}} \\
&\qquad = \|g\|_{\mathcal{H}}^2 .
\end{aligned}$$

Therefore,

$$\varphi'(0) = -\|g\|_{\mathcal{H}}^2 . \tag{17}$$

Next, we calculate the term of (14) that contains the second derivative using (16). First,

$$\begin{aligned}
\frac{d}{dt} \operatorname{tr}\left(\left(J\phi_t(x)\right)^{-1} Jg(x)\right) &= \operatorname{tr}\left(\frac{d}{dt}\left(J\phi_t(x)\right)^{-1} Jg(x)\right) \\
&= -\operatorname{tr}\left(\left(J\phi_t(x)\right)^{-2} \frac{d}{dt} J\phi_t(x) Jg(x)\right) \\
&= \operatorname{tr}\left(J\phi_t(x)^{-2} Jg(x) Jg(x)\right) \\
&= \operatorname{tr}\left(\left(J\phi_t(x)^{-1} Jg(x)\right)^2\right).
\end{aligned}$$

From the definition of the function ϕ_t, we know that $J\phi_t(x) = (\mathrm{I}_d + tJg)(x)$. Thus, $(J\phi_t)^{-1}$ and Jg commute. The latter yields $\operatorname{tr}\left(\left(Jg(x)\left(J\phi_t(x)\right)^{-1}\right)^2\right) = \left\|Jg(x)\left(J\phi_t(x)\right)^{-1}\right\|_{\mathrm{HS}}^2$. On the other hand, applying the chain rule on second term of (16) yields

$$\partial_t\big(\left\langle \nabla V\left(\phi_t(x)\right), g(x)\right\rangle\big) = \left\langle g(x), \nabla^2 V\left(\phi_t(x)\right) g(x)\right\rangle.$$

Summing up, we have the following:

$$\varphi''(t) = \underbrace{\int \left\|Jg(x)\left(J\phi_t(x)\right)^{-1}\right\|_{\mathrm{HS}}^2 \mu_n(\mathrm{d}x)}_{:=\psi_1(t)} + \underbrace{\int \left\langle g(x), \nabla^2 V\left(\phi_t(x)\right) g(x)\right\rangle \mu_n(\mathrm{d}x)}_{:=\psi_2(t)}. \tag{18}$$

First, we bound $\psi_1(t)$. Cauchy-Schwarz implies that

$$\left\|Jg(x)\left(J\phi_t(x)\right)^{-1}\right\|_{\mathrm{HS}}^2 \le \|Jg(x)\|_{\mathrm{HS}}^2 \left\|\left(J\phi_t(x)\right)^{-1}\right\|_{op}^2.$$

From Lemma 3, we have $\|Jg(x)\|_{\mathrm{HS}} \le B\|g\|_{\mathcal{H}}$. To bound the second term, let us recall that $\phi_t = I - tg$ and that $t \le \gamma$. Thus, the following bound is true:

$$\left\|\left(J\phi_t(x)\right)^{-1}\right\|_{op} = \left\|\left(\left(\mathrm{I}_d - tJg\right)(x)\right)^{-1}\right\|_{op} \le \sum_{i=0}^{\infty} \|tJg(x)\|_{op}^i \le \sum_{i=0}^{\infty} \|\gamma Jg(x)\|_{\mathrm{HS}}^i.$$

Recalling (7) and combining it with Lemma 3 we obtain

$$\left\|\left(J\phi_t(x)\right)^{-1}\right\|_{op} \le \sum_{i=0}^{\infty} (\gamma B\|g\|_{\mathcal{H}})^i \le \sum_{i=0}^{\infty} \left(\frac{\alpha-1}{\alpha}\right)^i = \alpha.$$

Summing up, we have that

$$\psi_1(t) \le \alpha^2 B^2 \|g\|_{\mathcal{H}}^2.$$

Next, we bound $\psi_2(t)$. By definition,

$$\psi_2(t) = \mathbb{E}_{X\sim\mu_n}\left[\left\langle g(X), \nabla^2 V\left(\phi_t(X)\right) g(X)\right\rangle\right] \le \mathbb{E}_{X\sim\mu_n}\left[\|\nabla^2 V\left(\phi_t(X)\right)\|_{op} \|g(X)\|_2^2\right].$$

Let us bound the norm of $g(x)$. The reproduction property of the RKHS yields the following:

$$\|g(x)\|_2^2 = \sum_{i=1}^{d} \langle k(x,.), g_i\rangle_{\mathcal{H}_0}^2 \le \|k(x,.)\|_{\mathcal{H}_0}^2 \|g\|_{\mathcal{H}}^2 \le B^2\|g\|_{\mathcal{H}}^2. \tag{19}$$

Therefore,

$$\psi_2(t) \le B^2\|g\|_{\mathcal{H}}^2 \mathbb{E}_{X\sim\mu_n}\left[\|\nabla^2 V\left(\phi_t(X)\right)\|_{op}\right]. \tag{20}$$

Let us bound $\mathbb{E}_{X\sim\mu_n}\left[\|\nabla^2 V\left(\phi_t(X)\right)\|_{op}\right]$. Assumption (L_0, L_1) implies the following inequality:

$$\left\|\nabla^2 V(\phi_t(x))\right\|_{op} \le L_0 + L_1\|\nabla V(\phi_t(x))\|,$$

for every $x \in \mathbb{R}^d$. To bound the term $\|\nabla V(\phi_t(x))\|$, we use the following lemma. Note Lemma 4 has a sharper constant than (Zhang et al., 2019a)[Lemma 9].

Lemma 4. *Let V be an (L_0, L_1)-smooth function and $\Delta > 0$ be a constant. For any $x, x^+ \in \mathbb{R}^d$ such that $\|x^+ - x\| \le \Delta$, we have*

$$\left\|\nabla V\left(x^+\right)\right\| \le \frac{L_0}{L_1}\left(\exp(\Delta L_1) - 1\right) + \|\nabla V(x)\|\exp(\Delta L_1).$$

We will apply Lemma 4 to $\phi_t(x)$ and $\phi_0(x)$. By definition, $\phi_t(x) - \phi_0(x) = tg(x)$ and according to inequality (19),

$$\|\phi_t(x) - \phi_0(x)\|_2 \le tB\|g\|_{\mathcal{H}}.$$

Thus, using Lemma 4 for $x = \phi_0(x)$, $x^+ = \phi_t(x)$ and $\Delta = tB\|g\|_{\mathcal{H}}$, we obtain the following:

$$\begin{aligned}\left\|\nabla^2 V(\phi_t(x))\right\|_{op} &\leq L_0 + L_1\left(\frac{L_0}{L_1}\big(\exp(tB\|g\|_{\mathcal{H}}L_1) - 1\big) + \|\nabla V\left(\phi_0(x)\right)\|\exp(tB\|g\|_{\mathcal{H}}L_1)\right)\\ &= L_0 + L_0\big(\exp(tB\|g\|_{\mathcal{H}}L_1) - 1\big) + L_1\|\nabla V\left(\phi_0(x)\right)\|\exp(tB\|g\|_{\mathcal{H}}L_1)\\ &= \big(L_0 + L_1\|\nabla V\left(x\right)\|\big)\exp(tBL_1\|g\|_{\mathcal{H}}).\end{aligned} \tag{21}$$

Combining (20) and (21) we obtain

$$\psi_2(t) \leq B^2\|g\|_{\mathcal{H}}^2\left(L_0 + L_1\mathbb{E}_{X\sim\mu_n}\big[\,\|\nabla V(X)\|\,\big]\right)\exp(tBL_1\|g\|_{\mathcal{H}}).$$

Summing up, the bounds on ψ_1 and ψ_2 yield the following inequality:

$$\varphi''(t) \leq B^2\|g\|_{\mathcal{H}}^2\Big[\alpha^2 + \Big(L_0 + L_1\mathbb{E}_{X\sim\mu_n}\big[\,\|\nabla V(X)\|\,\big]\Big)\exp(tBL_1\|g\|_{\mathcal{H}})\Big].$$

Recall that by definition $A_n = \big(L_0 + L_1\mathbb{E}_{X\sim\mu_n}\big[\,\|\nabla V(X)\|\,\big]\big)$. Inserting the previous inequality along with (17) to (14), we get the following bound:

$$\varphi(\gamma) - \varphi(0) \leq -\gamma\|g\|_{\mathcal{H}}^2 + \Big[\frac{1}{2}\gamma^2\alpha^2B^2\|g\|_{\mathcal{H}}^2 + A_n\big(\exp(\gamma BL_1\|g\|_{\mathcal{H}}) - \gamma BL_1\|g\|_{\mathcal{H}} - 1\big)\Big].$$

One can check that $\exp(t) - t - 1 \leq (e-1)t^2/2$, when $t \in [0,1]$. Since $\gamma BL_1\|g\|_{\mathcal{H}} < 1$, we deduce

$$\varphi(\gamma) - \varphi(0) \leq \Big[-\gamma + \frac{1}{2}(\gamma^2\alpha^2B^2 + (e-1)A_n\gamma^2B^2)\Big]\|g\|_{\mathcal{H}}^2.$$

Finally, by the definition of Stein's information $\|g\|_{\mathcal{H}}^2 = I_{\mathrm{Stein}}(\mu_n \mid \pi)$. This concludes the proof.

Remark 2. *In fact, the derivation of* (17) *contains the proof of the fact that* $\|g\|_{\mathcal{H}} = \sqrt{I_{\mathrm{Stein}}(\mu_n \mid \pi)}$. *Indeed, by the definition of the SVGD (Section 2.2) the direction function* g *is chosen to minimize the descent the most. On the other hand, the Stein-Fisher information is the maximum of the Stein's linear operator which coincides with our objective. Thus, the absolute value of the objective function at* g, *that is* $|\phi'(0)|$ *equals* $\sqrt{I_{\mathrm{Stein}}(\mu_n \mid \pi)}$.

B.2 Proof of Theorem 1

Proof. Let us first prove that the sequence $\mathrm{KL}(\mu_n \mid \pi)$ is monotonically decreasing. We will use the method of mathematical induction. First let us notice that (11) implies the following system of inequalities for the step-size:

$$\begin{cases}\gamma &\leq (\alpha-1)\min\{1, 1/L_1\}\left[\alpha B^2\left(C_0+1\right)\right]^{-1};\\ \gamma &\leq B^{-2}\Big[\alpha^2 + (e-1)\Big(L_0 + L_1C_0\Big)\Big]^{-1}.\end{cases} \tag{22}$$

For $n = 0$ the first equation of the system (22) combined with Lemma 2 implies that the step-size condition is satisfied. Thus, according to Proposition 1,

$$\mathrm{KL}\left(\mu_1 \mid \pi\right) - \mathrm{KL}\left(\mu_0 \mid \pi\right) \leq -\gamma\left[1 - \frac{\gamma}{2}B^2(\alpha^2 + (e-1)(L_0 + L_1C_0))\right]I_{\mathrm{Stein}}(\mu_0 \mid \pi).$$

On the other hand, from the second inequality of (22) we get that

$$1 - \frac{\gamma}{2}B^2(\alpha^2 + (e-1)(L_0 + L_1C_0)) \geq \frac{1}{2}.$$

Therefore, $\mathrm{KL}\left(\mu_1 \mid \pi\right) \leq \mathrm{KL}\left(\mu_0 \mid \pi\right)$. Let us define by $C_n := \mathrm{Q}\left(\mathrm{S}\big(\mathrm{KL}(\mu_n \mid \pi)\big) + W_p(\pi, \delta_0)\right)$. Since Q and S are monotonically increasing for positive arguments, we obtain $C_1 \leq C_0$. Therefore,

$$\begin{cases}\gamma &\leq (\alpha-1)\min\{1, 1/L_1\}\left[\alpha B^2\left(C_1+1\right)\right]^{-1};\\ \gamma &\leq B^{-2}\Big[\alpha^2 + (e-1)\big(L_0 + L_1C_1\big)\Big]^{-1}.\end{cases}$$

We retrieve (22) where the term C_0 is replaced by C_1. Thus, we can repeat the previous arguments for μ_1 and μ_2. Similarly, we can iterate till n. Therefore we obtain the following descent bound:

$$\mathrm{KL}\left(\mu_{n+1} \mid \pi\right) - \mathrm{KL}\left(\mu_n \mid \pi\right) \leq -\gamma\left[1 - \frac{\gamma}{2}B^2(\alpha^2 + (e-1)(L_0 + L_1 C_n))\right] I_{\mathrm{Stein}}(\mu_n \mid \pi).$$

This also yields that the sequence $\mathrm{KL}(\mu_n \mid \pi)$ is decreasing. Since S and P are monotonically increasing for positive arguments, C_n is also decreasing. Thus,

$$\begin{aligned}\mathrm{KL}\left(\mu_{n+1} \mid \pi\right) - \mathrm{KL}\left(\mu_n \mid \pi\right) &\leq -\gamma\left[1 - \frac{\gamma}{2}B^2(\alpha^2 + (e-1)(L_0 + L_1 C_0))\right] I_{\mathrm{Stein}}(\mu_n \mid \pi) \\ &\leq -\frac{\gamma}{2} I_{\mathrm{Stein}}(\mu_n \mid \pi).\end{aligned}$$

The last inequality is due to the second condition on the step-size γ. □

B.3 Proof of Corollary 1

Let us start with the first statement. Summing the descent bounds of Theorem 1 for $i = 0, 1, \ldots, n-1$, we obtain the following:

$$\mathrm{KL}\left(\mu_n \mid \pi\right) - \mathrm{KL}\left(\mu_0 \mid \pi\right) \leq -\frac{\gamma}{2}\sum_{i=0}^{n-1} I_{\mathrm{Stein}}(\mu_i \mid \pi).$$

Rearranging the terms we get

$$\begin{aligned}\sum_{i=0}^{n-1} I_{\mathrm{Stein}}(\mu_i \mid \pi) &\leq \frac{2}{\gamma}(\mathrm{KL}\left(\mu_0 \mid \pi\right) - \mathrm{KL}\left(\mu_n \mid \pi\right)) \\ &\leq \frac{2}{\gamma}\mathrm{KL}\left(\mu_0 \mid \pi\right).\end{aligned}$$

This means that the series on the left-hand side is convergent and thus, its general term converges to zero. Thus the first point is proved. Dividing both sides of the previous inequality on n, we deduce the second statement.

C Complexity results

C.1 Proof of Theorem 2

Proof. From Corollary 1, we know that if (11) is satisfied, then

$$\frac{1}{n}\sum_{k=0}^{n-1} I_{\mathrm{Stein}}\left(\mu_k \mid \pi\right) \leq \frac{2\,\mathrm{KL}(\mu_0 \mid \pi)}{n\gamma}.$$

Let us bound the initial KL error.

Lemma 5. *Let Assumption* $(\mathrm{poly}, \mathrm{Q})$ *hold with some polynomial* Q *and* $\mu_0 = \mathcal{N}(0, \mathrm{I}_d)$*. We then have*

$$\mathrm{KL}(\mu_0 \mid \pi) \leq \frac{d}{2}\log\left(\frac{1}{2\Pi e}\right) + V(0) + \mathrm{Q}(1)d\sqrt{\frac{2}{\Pi}} + \frac{\mathrm{Q}(1)(2d)^{\frac{p+1}{2}}\Gamma\left(\frac{p+2}{2}\right)}{\sqrt{\Pi}(p+1)},$$

where Π *is the area of the circle of radius* 1*.*

Since $p \geq 1$, Lemma 5 implies that

$$\begin{aligned}\mathrm{KL}(\mu_0 \mid \pi) &= \mathcal{O}\left(\frac{\mathrm{Q}(1)(2d)^{\frac{p+1}{2}}\Gamma\left(\frac{p+2}{2}\right)}{\sqrt{\Pi}(p+1)}\right) \\ &= \mathcal{O}\left(\mathrm{Q}(1)\,(pd)^{\frac{p+1}{2}}\right).\end{aligned} \tag{23}$$

The second equality is due to Stirling's formula: $\Gamma(r+1) = \mathcal{O}(\sqrt{2\Pi r}(r/e)^r)$, for every $r > 0$. In order to estimate the order of the step-size, let us compute the order of C_0. By definition,

$$\begin{aligned} C_0 &= \mathrm{Q}\left(\mathrm{S}\big(\mathrm{KL}(\mu_0 \mid \pi)\big) + W_p(\pi, \delta_0)\right) \\ &\leq \mathrm{Q}\left(\mathrm{S}\big(\mathrm{KL}(\mu_0 \mid \pi)\big) + W_p(\pi, \mu_0) + W_p(\mu_0, \delta_0)\right) \\ &\leq \mathrm{Q}\left(2\mathrm{S}\big(\mathrm{KL}(\mu_0 \mid \pi)\big) + W_p(\mu_0, \delta_0)\right). \end{aligned}$$

Here we applied the triangle inequality of W_p and the Assumption $(\mathrm{T}_p, \mathrm{S})$.

Let us now consider the first point of the theorem. Using $Q(r) \leq Q(1)(r^p + 1)$ and $\mathrm{S}(r) = \lambda_{BV}(r^{1/p} + (r/2)^{1/2p}) = \mathcal{O}(\lambda_{BV} r^{1/p})$, we obtain

$$\begin{aligned} C_0 &\leq \mathrm{Q}(1) \cdot \left(\left(2\mathrm{S}\big(\mathrm{KL}(\mu_0 \mid \pi)\big) + W_p(\mu_0, \delta_0)\right)^p + 1\right) \\ &= \mathcal{O}\Big(\mathrm{Q}(1) \cdot \left(\left(2\lambda_{BV}\,\mathrm{KL}\left(\mu_0 \mid \pi\right)^{1/p} + W_p(\mu_0, \delta_0)\right)^p + 1\right)\Big) \\ &= \mathcal{O}\Big(\mathrm{Q}(1) \cdot \left((4\lambda_{BV})^p\,\mathrm{KL}\left(\mu_0 \mid \pi\right) + 2^p W_p^p(\mu_0, \delta_0) + 1\right)\Big). \end{aligned} \tag{24}$$

Applying (23) we get

$$\begin{aligned} C_0 &= \mathcal{O}\left(\mathrm{Q}(1) \cdot \left((4\lambda_{BV})^p \mathcal{O}\left(\mathrm{Q}(1)\,(pd)^{\frac{p+1}{2}}\right) + 2^p W_p^p(\mu_0, \delta_0)\right)\right) \\ &= \mathcal{O}\left(\mathrm{Q}(1)^2 (4\lambda_{BV})^p\,(pd)^{\frac{p+1}{2}} + \mathrm{Q}(1) 2^p W_p^p(\mu_0, \delta_0)\right). \end{aligned} \tag{25}$$

The following lemma is to bound the $(p+1)$-th norm of the standard multivariate Gaussian.

Lemma 6. *Let μ_0 be the standard multivariate Gaussian defined on $\mathbb{R}^d$. Then, for every integer $m \geq 2$*

$$\int_{\mathbb{R}^d} \|x\|^m \mu_0(\mathrm{d}x) \leq \frac{(2d)^{\frac{m}{2}}}{\sqrt{\Pi}} \Gamma\left(\frac{m+1}{2}\right).$$

If we write down the definition of the Wasserstein distance, then Lemma 6 for $m = p$ yields

$$W_p^p(\mu_0, \delta_0) = \int_{\mathbb{R}^d} \|x\|^p \mu_0(\mathrm{d}x) \leq \frac{(2d)^{\frac{p}{2}} \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\Pi}} = \mathcal{O}\left((pd)^{\frac{p}{2}}\right). \tag{26}$$

This implies

$$\begin{aligned} C_0 &= \mathcal{O}\left(\mathrm{Q}(1)^2 (4\lambda_{BV})^p\,(pd)^{\frac{p+1}{2}} + \mathrm{Q}(1) 2^p\,(pd)^{\frac{p}{2}}\right) \\ &= \mathcal{O}\left(\mathrm{Q}(1)^2 (4\lambda_{BV})^p\,(pd)^{\frac{p+1}{2}}\right). \end{aligned}$$

Now let us look back at (11). This inequality yields that Corollary 1 is true for

$$\begin{aligned} \gamma &= \mathcal{O}\left(\left\{B^2(\alpha^2 + (e-1)(L_0 + L_1 \mathrm{Q}(1)^2 (4\lambda_{BV})^p\,(pd)^{\frac{p+1}{2}}))\right\}^{-1}\right) \\ &= \mathcal{O}\left(\left\{B^2 \max(L_1, 1) \mathrm{Q}(1)^2 \lambda_{BV}^p\,(pd)^{\frac{p+1}{2}}\right\}^{-1}\right). \end{aligned}$$

This implies the following equality and concludes the proof of the first point:

$$\begin{aligned} n = \frac{\mathrm{KL}(\mu_0 \mid \pi)}{2\gamma\varepsilon} &= \mathcal{O}\left(\frac{1}{\varepsilon}\left[Q(1)\,(pd)^{\frac{p+1}{2}}\right]\left[B^2(\alpha^2 + (e-1)(L_0 + L_1 \mathrm{Q}(1)^2 (4\lambda_{BV})^p\,(pd)^{\frac{p+1}{2}}))\right]\right) \\ &= \mathcal{O}\left(B^2 \max(L_1, 1) \frac{Q(1)^3 \lambda_{BV}^p (pd)^{p+1}}{\varepsilon}\right). \end{aligned}$$

For the second point, we will do the same analysis. In the case of T_p inequality, the target satisfies Assumption $(\mathrm{T}_p, \mathrm{S})$ with $\mathrm{S}(r) = \sqrt{2r/\lambda_{\mathrm{T}}}$, for some constant $\lambda_{\mathrm{T}} > 0$. Then, similar to (24), we obtain the following complexity for C_0:

$$\begin{aligned} C_0 &\leq \mathrm{Q}(1) \cdot \left(\left(2\mathrm{S}\big(\mathrm{KL}(\mu_0 \mid \pi)\big) + W_p(\mu_0, \delta_0)\right)^p + 1\right) \\ &= \mathrm{Q}(1) \cdot \left(\left(2\sqrt{2\,\mathrm{KL}\left(\mu_0 \mid \pi\right)/\lambda_{\mathrm{T}}} + W_p(\mu_0, \delta_0)\right)^p + 1\right) \\ &\leq \mathrm{Q}(1) \cdot \left((32/\lambda_{\mathrm{T}})^{\frac{p}{2}}\,\mathrm{KL}\left(\mu_0 \mid \pi\right)^{\frac{p}{2}} + 2^p W_p^p(\mu_0, \delta_0) + 1\right). \end{aligned}$$

Applying (23)

$$C_0 = \mathcal{O}\left(\mathrm{Q}(1)\cdot\left((32/\lambda_{\mathrm{T}})^{\frac{p}{2}}\mathcal{O}\left(\mathrm{Q}(1)^{\frac{p}{2}}\,(pd)^{\frac{p(p+1)}{4}}\right)+2^pW_p^p(\mu_0,\delta_0)\right)\right)$$
$$= \mathcal{O}\left(\mathrm{Q}(1)^{\frac{p+2}{2}}(32/\lambda_{\mathrm{T}})^{\frac{p}{2}}\,(pd)^{\frac{p(p+1)}{4}}+\mathrm{Q}(1)2^pW_p^p(\mu_0,\delta_0)\right).$$

Applying (26), we conclude

$$C_0 = \mathcal{O}\left(\mathrm{Q}(1)^{\frac{p+2}{2}}\lambda_{\mathrm{T}}^{-\frac{p}{2}}\,(pd)^{\frac{p(p+1)}{4}}\right).$$

Thus, in order to have ε average I_{Stein} error it is sufficient to perform n iterations, where

$$n=\frac{\mathrm{KL}(\mu_0\mid\pi)}{\gamma\varepsilon}=\mathcal{O}\left(\frac{(pd)^{\frac{p+1}{2}}}{\varepsilon}\cdot B^2\big(\alpha^2+(e-1)(\max(L_0,L_1,1)+\max(L_1,1)C_0)\big)\right)$$
$$=\mathcal{O}\left(B^2\max(L_1,1)\frac{\mathrm{Q}(1)^{\frac{p+2}{2}}\lambda_{\mathrm{T}}^{-\frac{p}{2}}\,(pd)^{\frac{(p+1)(p+2)}{4}}}{\varepsilon}\right).$$

This concludes the proof.

□

D Proofs of the lemmas

D.1 Proof of Lemma 1

Proof. Denote $\Phi(x):=k(x,\cdot)\in\mathcal{H}$. Then by the definition of the Stein discrepancy we have the following:

$$I_{\mathrm{Stein}}(\mu_n\mid\pi)^{\frac{1}{2}}=\big\|\mathbb{E}_{X\sim\mu_n}\big[(\nabla V(X)\Phi(X)-\nabla\Phi(X))\big]\big\|_{\mathcal{H}}$$
$$\leq\mathbb{E}_{X\sim\mu_n}\big[\|\nabla V(X)\Phi(X)-\nabla\Phi(X)\big]\|_{\mathcal{H}}.$$

Applying the triangle inequality and Cauchy-Schwartz inequality, we obtain

$$I_{\mathrm{Stein}}(\mu_n\mid\pi)^{\frac{1}{2}}\leq\mathbb{E}_{X\sim\mu_n}\big[\|\nabla V(X)\Phi(X)\|_{\mathcal{H}}\big]+\mathbb{E}_{X\sim\mu_n}\big[\|\nabla\Phi(X)\|_{\mathcal{H}}\big]$$
$$=\mathbb{E}_{X\sim\mu_n}\big[\|\nabla V(X)\|\|\Phi(X)\|_{\mathcal{H}}\big]+\mathbb{E}_{X\sim\mu_n}\big[\|\nabla\Phi(X)\|_{\mathcal{H}}\big]$$
$$\leq B\left(\mathbb{E}_{X\sim\mu_n}\big[\|\nabla V(X)\|\big]+1\right).$$

□

D.2 Proof of Lemma 2

Let the polynomial Q have the following explicit form:

$$\mathrm{Q}(r)=\sum_{i=0}^{m}a_ir^{p_i}\tag{27}$$

for every $r\in\mathbb{R}$. Here $p=p_0>p_1>\ldots>p_m$ and $a_i>0$. Then Assumption $(\mathrm{poly},\mathrm{Q})$ yields

$$\mathbb{E}_{X\sim\mu_n}\big[\|\nabla V(X)\|\big]\leq\mathbb{E}_{X\sim\mu_n}\big[\mathrm{Q}(\|X\|)\big]$$
$$=\mathbb{E}_{X\sim\mu_n}\left[\sum_{i=0}^{m}a_i\|X\|^{p_i}\right]$$
$$=\sum_{i=0}^{m}a_iW_{p_i}\big(\mu_n,\delta_0\big)^{p_i},$$

where δ_0 is Dirac measure on $\mathbb{R}^d$ at point 0. Using the fact that W_r is monotonically increasing w.r.t. to r, we have the following

$$\begin{aligned}\mathbb{E}_{X\sim\mu_n}\left[\|\nabla V(X)\|\right] &\leq \sum_{i=0}^{m} a_i W_p\big(\mu_n,\delta_0\big)^{p_i}\\ &\leq \sum_{i=1}^{m} a_i\big(W_p(\mu_n,\pi)+W_p(\pi,\delta_0)\big)^{p_i}\\ &\leq \sum_{i=1}^{m} a_i\left(\mathrm{S}\big(\mathrm{KL}(\mu_n\mid\pi)\big)+W_p(\pi,\delta_0)\right)^{p_i}.\\ &\leq \mathrm{Q}\left(\mathrm{S}\big(\mathrm{KL}(\mu_n\mid\pi)\big)+W_p(\pi,\delta_0)\right).\end{aligned}$$

The third inequality is due to Assumption $(\mathrm{T}_p,\mathrm{S})$. Recalling (27) we conclude the proof.

D.3 Proof of Lemma 3

The proof is based on the reproducing property and Cauchy-Schwarz inequality in the RKHS space. Indeed,

$$\begin{aligned}\|Jh(x)\|_{\mathrm{HS}}^2 &= \sum_{i,j=1}^{d}\left|\frac{\partial h_i(x)}{\partial x_j}\right|^2\\ &= \sum_{i,j=1}^{d}\left\langle\partial_{x_j}k(x,.),h_i\right\rangle_{\mathcal{H}_0}\\ &\leq \sum_{i,j=1}^{d}\left\|\partial_{x_j}k(x,.)\right\|_{\mathcal{H}_0}^2\|h_i\|_{\mathcal{H}_0}^2\\ &= \|\nabla k(x,.)\|_{\mathcal{H}}^2\|h\|_{\mathcal{H}}^2\\ &\leq B^2\|h\|_{\mathcal{H}}^2.\end{aligned}$$

This concludes the proof.

D.4 Proof of Lemma 4

Proof. The proof is similar to the proof of (Zhang et al., 2019a)[Lemma 9]. Let us fix $x, x^+\in\mathbb{R}^d$ and let $\tau(t)$ be defined as $\tau(t)=t\left(x^+-x\right)+x$ with $t\in[0,1]$. Then we have

$$\nabla V(\tau(t)) = \int_0^t \nabla^2 V(\tau(u))\left(x^+-x\right)\mathrm{d}\tau + \nabla V(\tau(0)).$$

Let us bound the norm of $\nabla V(\tau(t))$. Applying triangle and Cauchy-Schwartz inequalities we obtain:

$$\begin{aligned}\|\nabla V(\tau(t))\| &\leq \int_0^t\left\|\nabla^2 V(\tau(u))\left(x^+-x\right)\right\|\mathrm{d}u + \|\nabla V(\tau(0))\|\\ &\leq \left\|x^+-x\right\|\int_0^t\left\|\nabla^2 V(\tau(u))\right\|\mathrm{d}u + \|\nabla V(x)\|.\end{aligned}$$

Assumption (L_0,L_1) yields

$$\begin{aligned}\|\nabla V(\tau(t))\| &\leq \Delta\int_0^t\left(L_0+L_1\|\nabla V(\tau(u))\|\right)\mathrm{d}u + \|\nabla V(x)\|\\ &= \Delta L_0 t + \|\nabla V(x)\| + \int_0^t \Delta L_1\left\|\nabla V(\tau(u))\right\|\mathrm{d}u.\end{aligned}$$

Applying Grönwall's integral inequality (see Lemma 8) to the function $\|\nabla V(\tau(\cdot))\|$ we obtain

$$\|\nabla V(\tau(t))\| = \Delta L_0 t + \|\nabla V(x)\| + \int_0^t \Delta L_1\big(\Delta L_0 u + \|\nabla V(x)\|\big)\exp\big(\Delta L_1(t-u)\big)\mathrm{d}u.$$

Inserting $t=1$, we get the following:

$$\begin{aligned}\|\nabla V(x^+)\| &\le \Delta L_0 + \|\nabla V(x)\| + \int_0^1 \Delta L_1\big(\Delta L_0 u + \|\nabla V(x)\|\big)\exp\big(\Delta L_1(1-u)\big)\mathrm{d}u\\ &= -\frac{L_0}{L_1} + \left(\frac{L_0}{L_1} + \|\nabla V(x)\|\right)\exp(\Delta L_1).\end{aligned}$$

This concludes the proof. □

The rest of the section contains the proofs of the lemmas appearing in the proof of Theorem 2. Without loss of generality, we may assume that the normalizing constant of the density of π is equal to 1. Thus, $\pi(x)=e^{-V(x)}$ and the following lemma is true.

D.5 Proof of Lemma 5

Proof. By Assumption $(\mathrm{poly},\mathrm{Q})$, we know that $\big\|\nabla V(x)\big\| \le \mathrm{Q}(\|x\|)$. Thus, applying Taylor formula

$$\begin{aligned}V(x) &= \int_0^1 \langle \nabla V(tx), x\rangle\,\mathrm{d}t + V(0)\\ &\le \int_0^1 \|\nabla V(tx)\|\,\|x\|\mathrm{d}t + V(0)\\ &\le \int_0^1 \mathrm{Q}(\|tx\|)\|x\|\mathrm{d}t + V(0).\end{aligned}$$

Since the coefficients of Q are positive, one may verify that for every $r\ge 0$

$$\mathrm{Q}(r) \le \mathrm{Q}(1)(r^p+1).$$

Therefore,

$$\begin{aligned}V(x) &\le \int_0^1 \mathrm{Q}(1)\left(\|tx\|^p+1\right)\|x\|\mathrm{d}t + V(0)\\ &\le \mathrm{Q}(1)\left(\frac{\|x\|^{p+1}}{p+1} + \|x\|\right) + V(0).\end{aligned}$$

Now, let us calculate $\mathrm{KL}(\mu_0 \mid \pi)$,

$$\begin{aligned}\mathrm{KL}(\mu_0\mid\pi) &= \int_{\mathbb{R}^d}\log\left(\frac{\mu_0}{\pi}(x)\right)\mu_0(\mathrm{d}x)\\ &= \int_{\mathbb{R}^d}\log(\mu_0(x))\mu_0(x)\mathrm{d}x + \int_{\mathbb{R}^d} V(x)\mu_0(\mathrm{d}x)\\ &\le \frac{d}{2}\log\left(\frac{1}{2\Pi e}\right) + \int_{\mathbb{R}^d}\left[\mathrm{Q}(1)\left(\frac{\|x\|^{p+1}}{p+1}+\|x\|\right)+V(0)\right]\mu_0(\mathrm{d}x).\end{aligned}$$

Combining Lemma 6 with the inequality below

$$\|x\| = \left(\sum_{i=1}^d |x_i|^2\right)^{\frac{1}{2}} \le \left(\sum_{i=1}^d |x_i|\right)^{2\cdot\frac{1}{2}} = \sum_{i=1}^d |x_i|,$$

we obtain the following:

$$\begin{aligned}\mathrm{KL}(\mu_0\mid\pi) &\le \frac{d}{2}\log\left(\frac{1}{2\Pi e}\right) + V(0) + \frac{\mathrm{Q}(1)}{p+1}\cdot\frac{(2d)^{\frac{p+1}{2}}}{\sqrt{\Pi}}\Gamma\left(\frac{p+2}{2}\right) + \mathrm{Q}(1)\int_{\mathbb{R}^d}\sum_{i=1}^d |x_i|\mu_0(\mathrm{d}x)\\ &= \frac{d}{2}\log\left(\frac{1}{2\Pi e}\right) + V(0) + \frac{\mathrm{Q}(1)}{p+1}\cdot\frac{(2d)^{\frac{p+1}{2}}}{\sqrt{\Pi}}\Gamma\left(\frac{p+2}{2}\right) + \mathrm{Q}(1)d\int_{\mathbb{R}}|r|\frac{1}{\sqrt{2\Pi}}e^{-\frac{r^2}{2}}\mathrm{d}r\\ &= \frac{d}{2}\log\left(\frac{1}{2\Pi e}\right) + V(0) + \frac{\mathrm{Q}(1)(2d)^{\frac{p+1}{2}}\Gamma\left(\frac{p+2}{2}\right)}{\sqrt{\Pi}(p+1)} + \mathrm{Q}(1)d\sqrt{\frac{2}{\Pi}}.\end{aligned}$$

This concludes the proof. □

D.6 Proof of Lemma 6

Proof. By Jensen inequality we have

$$\|x\|^m = \left(\sum_{i=1}^d |x_i|^2\right)^{\frac{m}{2}} = d^{\frac{m}{2}} \left(\frac{1}{d}\sum_{i=1}^d |x_i|^2\right)^{\frac{m}{2}} \leq d^{\frac{m-2}{2}} \sum_{i=1}^d |x_i|^m.$$

Thus, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^d} \|x\|^m \mu_0(\mathrm{d}x) &\leq \frac{1}{(2\Pi)^{d/2}} \int_{\mathbb{R}^d} d^{\frac{m-2}{2}} \sum_{i=1}^d |x_i|^m \exp\left(-\frac{\|x\|^2}{2}\right) \mathrm{d}x \\
&\leq \frac{d^{\frac{m-2}{2}}}{(2\Pi)^{d/2}} \sum_{i=1}^d \int_{\mathbb{R}^d} |x_i|^m \exp\left(-\frac{\|x\|^2}{2}\right) \mathrm{d}x \\
&\leq \frac{d^{\frac{m-2}{2}}}{\sqrt{2\Pi}} \sum_{i=1}^d \int_{\mathbb{R}} |x_i|^m \exp\left(-\frac{x_i^2}{2}\right) \mathrm{d}x_i.
\end{aligned}$$

From Winkelbauer (2012), we know the m-th central absolute moment of the standard one-dimensional Gaussian is equal to $2^{m/2}\Gamma(^{(m+1)}/_2)/\sqrt{\Pi}$. Thus, we obtain

$$\begin{aligned}
\int_{\mathbb{R}^d} \|x\|^m \mu_0(x)\mathrm{d}x &\leq \frac{d^{\frac{m-2}{2}}}{\sqrt{2\Pi}} \sum_{i=1}^d \int_{\mathbb{R}} |x_i|^m \exp\left(-\frac{x_i^2}{2}\right) \mathrm{d}x_i \\
&\leq \frac{d^{\frac{m}{2}}}{\sqrt{2\Pi}} \int_{\mathbb{R}} |r|^m \exp\left(-\frac{r^2}{2}\right) \mathrm{d}r \\
&\leq \frac{(2d)^{\frac{m}{2}}}{\sqrt{\Pi}} \Gamma\left(\frac{m+1}{2}\right).
\end{aligned}$$

□

E Miscellaneous

In this section, we present previously known auxiliary results that were mentioned in the paper. Some of the results are proved, the others refer to their papers of origin.

E.1 Reproducing property

Here we remind you of the reproducing property: let $f \in \mathcal{H}_0$, then we have $f(x) = \langle f(\cdot), k(x,\cdot)\rangle_{\mathcal{H}_0}$, if we choose $f(\cdot) = k(y,\cdot) \in \mathcal{H}_0$, then $k(y,x) = \langle k(y,\cdot), k(x,\cdot)\rangle_{\mathcal{H}_0}$.

Let us first use the reproducing property to calculate $\|k(x,\cdot)\|_{\mathcal{H}_0}$. The definition of the norm in Hilbert spaces yields:

$$\|k(x,\cdot)\|_{\mathcal{H}_0}^2 = \langle k(x,\cdot), k(x,\cdot)\rangle_{\mathcal{H}_0} = k(x,x).$$

Now, let us proceed to $\|\partial_{x_i} k(x,\cdot)\|_{\mathcal{H}_0}$. Let e_i be the i-th standard basis of $\mathbb{R}^d$. Then

$$\begin{aligned}
\langle \partial_{x_i} k(x,\cdot), \partial_{y_i} k(y,\cdot)\rangle_{\mathcal{H}_0} &= \lim_{\epsilon_1\to 0}\lim_{\epsilon_2\to 0} \left\langle \frac{k(x+\epsilon_1 e_i,\cdot) - k(x,\cdot)}{\epsilon_1}, \frac{k(y+\epsilon_2 e_i,\cdot) - k(y,\cdot)}{\epsilon_2}\right\rangle_{\mathcal{H}_0} \\
&= \lim_{\epsilon_1\to 0}\lim_{\epsilon_2\to 0} \frac{k(x+\epsilon_1 e_i, y+\epsilon_2 e_i) - k(x+\epsilon_1 e_i, y) - k(x, y+\epsilon_2 e_i) + k(x,y)}{\epsilon_1\epsilon_2} \\
&= \lim_{\epsilon_1\to 0}\lim_{\epsilon_2\to 0} \frac{(k(x+\epsilon_1 e_i, y+\epsilon_2 e_i) - k(x+\epsilon_1 e_i, y)) - (k(x, y+\epsilon_2 e_i) - k(x,y))}{\epsilon_1\epsilon_2} \\
&= \lim_{\epsilon_1\to 0} \frac{\partial_{y_i} k(x+\epsilon_1 e_i, y) - \partial_{y_i} k(x,y)}{\epsilon_1} \\
&= \partial_{x_i}\partial_{y_i} k(x,y).
\end{aligned}$$

Setting $x = y$, we obtain $\|\partial_{x_i} k(x,\cdot)\|_{\mathcal{H}_0}^2 = \partial_{x_i}\partial_{x_i} k(x,x) =: \partial^2_{x_i,x_i} k(x,x)$, where the first ∂_{x_i} is operated on the first variable of $k(\cdot,\cdot)$, the second ∂_{x_i} is operated on the second variable of $k(\cdot,\cdot)$.

E.2 The generalized T_p inequality

The following lemma is from Bolley and Villani (2005). The proof is omitted.

Lemma 7. *(Bolley and Villani, 2005, Corollary 2.3) Let $\mathbb{R}^d$ be the Euclidean space with its usual norm. Let $p \geq 1$ and let π be a probability measure on $\mathbb{R}^d$. Assume that there exist $x_0 \in \mathbb{R}^d$ and $s > 0$ such that $\int_{\mathbb{R}^d} \exp(s\|x_0 - x\|^p)\pi(\mathrm{d}x)$ is finite. Then*

$$\forall \mu \in \mathcal{P}(\mathbb{R}^d), \quad W_p(\mu, \pi) \leq \lambda_{BV} \left[\mathrm{KL}(\mu \mid \pi)^{\frac{1}{p}} + \left(\frac{\mathrm{KL}(\mu \mid \pi)}{2} \right)^{\frac{1}{2p}} \right], \tag{28}$$

where

$$\lambda_{BV} := 2 \inf_{x_0 \in X, s>0} \left(\frac{1}{s} \left(\frac{3}{2} + \log \int_{\mathbb{R}^d} \exp(s\|x_0 - x\|^p)\pi(\mathrm{d}x) \right) \right)^{\frac{1}{p}} < +\infty. \tag{29}$$

We want to underline the fact that the constant here may depend on the dimension. Below, we explicit the dimension-dependence of λ_{BV} for the general class of distributions $\pi(x) \propto \exp(-\|x\|^p)$ with $p \geq 1$. It is straightforward the condition of the lemma is satisfied for $x_0 = 0$. Now, let us compute the objective function in (29). We start with the integral term:

$$\begin{aligned}
\int_{x \in \mathbb{R}^d} \exp(s\,\|x\|^p)\pi(x)dx &= \frac{1}{Z} \int_{x \in \mathbb{R}^d} \exp(-(1-s)\,\|x\|^p)dx \\
&= \frac{1}{Z}(1-s)^{-\frac{d}{p}} \int_{x \in \mathbb{R}^d} \exp(-\|x\|^p)dx \\
&= (1-s)^{-\frac{d}{p}}.
\end{aligned}$$

Inserting the previous formula into (29), we obtain

$$\begin{aligned}
\lambda_{BV} :&= 2 \inf_{x_0 \in \mathbb{R}^d, s>0} \left(\frac{1}{s} \left(\frac{3}{2} + \log \int \exp(s\,\|x - x_0\|^p)d\pi(x) \right) \right)^{\frac{1}{p}} \\
&= 2 \inf_{s \in (0,1)} \left(\frac{3}{2s} - \frac{d\log(1-s)}{ps} \right)^{\frac{1}{p}} \\
&\geq 2 \inf_{s \in (0,1)} \left(\frac{3}{2s} + \frac{d}{p} \right)^{\frac{1}{p}} \\
&= 2d^{\frac{1}{p}}.
\end{aligned}$$

Thus, indeed the constant λ_{BV} can be dimension dependent.

E.3 Grönwall's integral inequality

The following lemma is the integral form of Grönwall inequality from (Amann, 2011, Chapter II.).

Lemma 8 (Grönwall Inequality)**.** *Assume $\phi, B : [0, T] \to \mathbb{R}$ are bounded non-negative measurable function and $C : [0, T] \to \mathbb{R}$ is a non-negative integrable function with the property that*

$$\phi(t) \leq B(t) + \int_0^t C(\tau)\phi(\tau)d\tau \quad \text{for all } t \in [0, T] \tag{30}$$

Then

$$\phi(t) \leq B(t) + \int_0^t B(s)C(s) \exp\left(\int_s^t C(\tau)d\tau \right) ds \quad \text{for all } t \in [0, T]. \tag{31}$$

E.4 Stein-Fisher information and weak convergence

We provide a sufficient condition on which $\lim_{n\to\infty} I_{Stein}(\mu_n \mid \pi)$ implies $\mu_n \to \pi$ weakly. This condition can be found in Gorham and Mackey (2017). Since Stein Fisher information $I_{Stein}(\cdot \mid \pi)$ depends on the target distribution π and the kernel $k(\cdot, \cdot)$, we need the following two properties, respectively:

1. π is distantly dissipative, that is $\kappa_0 \triangleq \liminf_{r\to\infty} \kappa(r) > 0$ with

$$\kappa(r) = \inf\left\{2\frac{\langle \nabla V(x) - \nabla V(y), x-y\rangle}{\|x-y\|^2} : \|x-y\| = r\right\}.$$

   If V is strongly convex outside a compact set, then π is distantly dissipative, for instance $V(x) = \|x\|^{2+\delta}$ with $\delta \geq 0$.

2. $k(\cdot,\cdot)$ is an inverse multiquadratic kernel, i.e., $k(x,y) = \left(c^2 + \|x-y\|^2\right)^{\beta}$ for some $c > 0$ and $\beta \in (-1, 0)$. It is easy to check that Assumption $(\ker, B)$ is satisfied.

E.5 Proof of Remark 1

Chain rule implies

$$\begin{aligned}
&\int_{\mathbb{R}^d} \nabla \log\left(\frac{\mu(x)}{\pi(x)}\right) k(x,\cdot)\mu(\mathrm{d}x) \\
&\quad = \int_{\mathbb{R}^d} \nabla \log(\mu(x)) k(x,\cdot)\mu(\mathrm{d}x) - \int_{\mathbb{R}^d} \nabla \log(\pi(x)) k(x,\cdot)\mu(\mathrm{d}x) \\
&\quad = \int_{\mathbb{R}^d} \nabla\mu(x) k(x,\cdot)\mathrm{d}x - \int_{\mathbb{R}^d} \log(\pi(x))\nabla_x k(x,\cdot)\mu(\mathrm{d}x).
\end{aligned}$$

Since $\mu(x) \to 0$, when $\|x\| \to +\infty$, the integration by parts yields

$$\begin{aligned}
&\int_{\mathbb{R}^d} \nabla\mu(x) k(x,\cdot)\mathrm{d}x - \int_{\mathbb{R}^d} \log(\pi(x))\nabla_x k(x,\cdot)\mu(\mathrm{d}x) \\
&\qquad = -\int_{\mathbb{R}^d} \left\{\mu(x)\nabla k(x,\cdot) + \log(\pi(x))\nabla_x k(x,\cdot)\right\}\mu(\mathrm{d}x) \\
&\qquad = g_\mu(\cdot).
\end{aligned}$$

This concludes the proof.

F Experiments

We run some experiments to test the stability of the SVGD on distributions with (L_0, L_1)-smooth potential functions. The model we choose is the one dimensional generalized Gaussian distributions:

$$\pi_{s,p}(x) = \frac{e^{-|\frac{x-\mu}{s}|^p}}{Z(s,p)}, \quad \text{with } Z(s,p) = \frac{2s}{p}\Gamma\left(\frac{1}{p}\right),$$

where $\Gamma(\cdot)$ is the well-known Gamma function. It can be calculated that $\mathbb{E}_{\pi_{s,p}}[x] = \mu$ and $\mathrm{Var}_{\pi_{s,p}}[X] = s\left(\Gamma(\frac{3}{p})/\Gamma(\frac{1}{p})\right)^{\frac{1}{2}}$, see Achim et al. (2008).

In the experiments, $\mu = 2$ and p varies between $2, 3, 4, 5$. We choose $s = 3\left(\Gamma(\frac{1}{p})/\Gamma(\frac{3}{p})\right)^{\frac{1}{2}}$, that is the all the variances are equal to 3. The set of initial samples $\left\{x_i^0\right\}_{i=1}^{250}$ are chosen independently from $\mathcal{N}(-10, 1)$. Then we run SVGD with the step-size $\gamma = 0.5$ for $n = 250, 500, 1000$ times respectively. The reproducing kernel is chosen to be $k(x,y) = \sqrt{\frac{1}{2\Pi}}e^{-\frac{|x-y|^2}{2}}$. The results are shown in Figure 1. We ran the experiments on Google Colab. The code can be found in `https://github.com/Iwillnottellyou/l0l1-SVGD.git`.

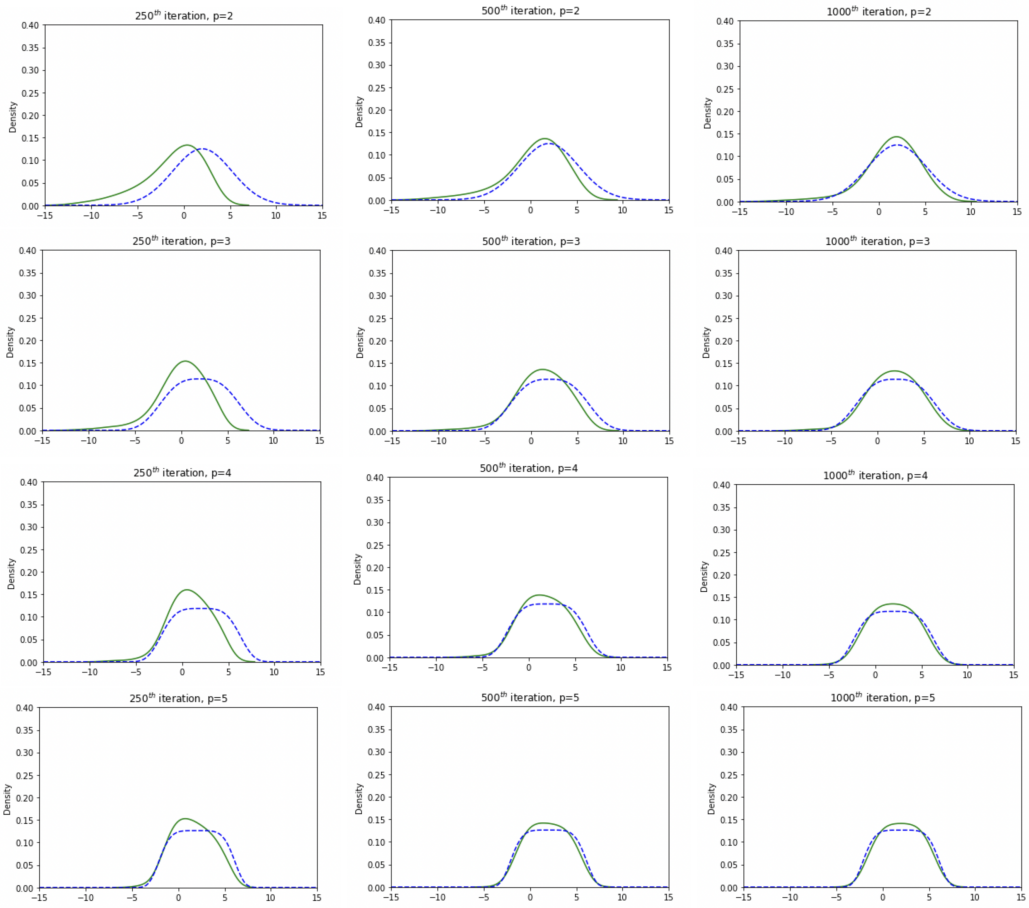

Figure 1: The performance of SVGD on $\pi_{s,p}$. The blue dashed line is the target distribution $\pi_{s,p}$, while the green line is the distribution generated by SVGD. We observe that the SVGD is stable with respect to the parameter p.