A Convergence Theory for SVGD in the Population Limit under Talagrand's Inequality T1

Adil Salim [1] **Lukang Sun** [2] **Peter Richtárik** [2]

Abstract

Stein Variational Gradient Descent (SVGD) is an algorithm for sampling from a target density which is known up to a multiplicative constant. Although SVGD is a popular algorithm in practice, its theoretical study is limited to a few recent works. We study the convergence of SVGD in the population limit, (i.e., with an infinite number of particles) to sample from a non-logconcave target distribution satisfying Talagrand's inequality T1. We first establish the convergence of the algorithm. Then, we establish a dimension-dependent complexity bound in terms of the Kernelized Stein Discrepancy (KSD). Unlike existing works, we do not assume that the KSD is bounded along the trajectory of the algorithm. Our approach relies on interpreting SVGD as a gradient descent over a space of probability measures.

1. Introduction

Sampling from a given target distribution π is a fundamental task of many Machine Learning procedures. In Bayesian Machine Learning, the target distribution π is typically known up to a multiplicative factor and often takes the form

$$\pi(x) \propto \exp(-F(x)), \tag{1}$$

where $F : \mathcal{X} \to \mathbb{R}$ is a L-smooth nonconvex function defined on $\mathcal{X} := \mathbb{R}^d$, and satisfying

$$\int \exp(-F(x))dx < \infty.$$

As sampling algorithms are intended to be applied to large scale problems, it has become increasingly important to understand their theoretical properties, such as their complexity, as a function of the *dimension of the problem* d, and the *desired accuracy* ε. In this regard, most of the Machine Learning literature on sampling has concentrated on understanding the complexity (in terms of d and ε) of (variants of) the Langevin algorithm, see Durmus et al. (2019); Bernton (2018); Wibisono (2018); Cheng et al. (2018); Salim & Richtárik (2020); Hsieh et al. (2018); Dalalyan (2017); Durmus & Moulines (2017); Rolland et al. (2020); Vempala & Wibisono (2019); Zou et al. (2019); Şimşekli (2017); Shen & Lee (2019); Bubeck et al. (2018); Durmus et al. (2018); Ma et al. (2019); Foster et al. (2021); Li et al. (2021).

1.1. Stein Variational Gradient Descent (SVGD)

Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016; Liu, 2017) is an alternative to the Langevin algorithm that has been applied in several contexts in Machine Learning, including Reinforcement Learning (Liu et al., 2017), sequential decision making (Zhang et al., 2018; 2019), Generative Adversarial Networks (Tao et al., 2019), Variational Auto Encoders (Pu et al., 2017), and Federated Learning (Kassab & Simeone, 2020).

The literature on theoretical properties of SVGD is scarce compared to that of Langevin algorithm, and limited to a few recent works (Korba et al., 2020; Lu et al., 2019; Duncan et al., 2019; Liu, 2017; Chewi et al., 2020; Gorham et al., 2020; Nüsken & Renger, 2021; Shi et al., 2021). In this paper, our goal is to provide a clean convergence theory for SVGD in the population limit, *i.e.*, with an infinite number of particles.

1.2. Related works

The Machine Learning literature on the complexity of sampling from a non-logconcave target distribution has mainly focused on the Langevin algorithm. For instance,[1] Vempala & Wibisono (2019) showed that Langevin algorithm reaches ε accuracy in terms of the Kullback-Leibler divergence after $\tilde{\Omega}(\frac{L^2 d}{\lambda^2 \varepsilon})$ iterations, assuming that the target distribution satis-

[1]Microsoft Research, Redmond, USA [2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Correspondence to: Adil Salim <adilsalim@microsoft.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]The example of Vempala & Wibisono (2019) is taken only for illustration purpose. Many other results were obtained for Langevin algorithm, even in nonconvex cases, see above.

fies the logarithmic-Sobolev inequality (LSI) with constant λ. In this work, we will assume Talagrand's inequality T1 with constant λ, which is milder than LSI with constant λ, and we will prove a complexity result in terms of another discrepancy called Kernelized Stein Discrepancy (KSD). Besides, a very recent work studies Langevin algorithm for a non-logconcave target distribution without assuming LSI and provides guarantees in terms of the Fisher information (Balasubramanian et al., 2022).

Most existing results on SVGD deal with the *continuous time* approximation of SVGD in the *population limit*, a Partial Differential Equation (PDE) representing SVGD with a vanishing step size and an infinite number of particles (Lu et al., 2019; Duncan et al., 2019; Liu, 2017; Nüsken & Renger, 2021; Chewi et al., 2020). In particular, Duncan et al. (2019) propose a Stein logarithmic Sobolev inequality that implies the linear convergence of this PDE. However, it is not yet understood when Stein logarithmic Sobolev inequality holds. Besides, Chewi et al. (2020) showed that the Wasserstein gradient flow of the chi-squared divergence can be seen as an approximation of that PDE, and showed linear convergence of the Wasserstein gradient flow of the chi-squared under Poincaré inequality. Other results, such as those of Lu et al. (2019); Liu (2017); Nüsken & Renger (2021), include asymptotic convergence properties of the PDE, but do not include convergence rates. In this paper, we will prove convergence rates for SVGD in discrete time.

1.2.1. Comparison to Korba et al. (2020)

The closest work to ours is Korba et al. (2020). To our knowledge, Korba et al. (2020) showed the first complexity result for SVGD in *discrete time*. This result is proven in the *population limit* and in terms of the Kernelized Stein Discrepancy (KSD), similarly to our main complexity result.

However, their complexity result relies on the assumption that the KSD is *uniformly bounded* along the iterations of SVGD, an assumption that cannot be checked prior to running the algorithm. Moreover, their complexity bound does not express *the dependence in the dimension* d explicitly. This is because the uniform bound on the KSD appears in their complexity bound. On the contrary, one of our contributions is to present a dimension-dependent complexity result under verifiable assumptions.

Besides, Korba et al. (2020) provide a bound on the distance between SVGD in the finite number of particles regime and SVGD in the population limit. This bound cannot be used to study the complexity or convergence rate of SVGD in the finite number of particles regime, see Korba et al. (2020, Proposition 7).

1.3. Contributions

We consider SVGD in the population limit, similarly to concurrent works such as Liu (2017); Korba et al. (2020); Gorham et al. (2020). Our paper intends to provide a clean analysis of SVGD, a problem stated in Liu (2017, Conclusion). To this end, we do not make any assumptions on the trajectory of the algorithm. Instead, our key assumption is that the target distribution π satisfies T1, the mildest of the Talagrand's inequalities, which holds under a mild assumption on the tails of the distribution; see Villani (2008, Theorem 22.10). Moreover, T1 is implied, for example, by the logarithmic Sobolev inequality (Villani, 2008, Theorem 22.17), with the same constant λ.

Although sampling algorithms are meant to be applied on high-dimensional problems, the question of the dependence of the complexity of SVGD in d has not been studied in concurrent works, nor has been studied the generic weak convergence of SVGD under verifiable assumptions, to our knowledge. Assuming that the T1 inequality holds, we provide

- a generic weak convergence result for SVGD (actually our result is a bit stronger: convergence holds in 1-Wasserstein distance),
- a complexity bound for SVGD in terms of the dimension d and the desired accuracy ε, under verifiable assumptions (i.e., assumptions that do not depend on the trajectory of the algorithm): $\tilde{\Omega}\left(\frac{Ld^{3/2}}{\lambda^{1/2}\varepsilon}\right)$ iterations suffice to obtain a sample μ such that $\mathrm{KSD}^2(\mu|\pi) < \varepsilon$, where L is the smoothness constant of F and λ the constant in T1 inequality.

Note that these results hold without assuming F convex. In particular, in the population limit, SVGD applied to *non-logconcave* target distributions satisfying T1 converges to the target distribution.

1.4. Paper structure

The remainder of the paper is organized as follows. In Section 2 we introduce the necessary mathematical and notational background on optimal transport, reproducing kernel Hilbert spaces and SVGD in order to be able to describe and explain our results. Section 3 is devoted to the development of our theory. Finally, in Section 4 we formulate three corollaries of our key result, capturing weak convergence and complexity estimates for SVGD. Technical proofs are postponed to the Appendix.

2. Background and Notation

2.1. Notation

For any Hilbert space H, we denote by $\langle \cdot, \cdot \rangle_H$ the inner product of H and by $\| \cdot \|_H$ its norm.

We denote by $C_0(\mathcal{X})$ the set of continuous functions from $\mathcal{X}$ to $\mathbb{R}$ vanishing at infinity and by $C^1(\mathcal{X}, \mathcal{Y})$ the set of continuously differentiable functions from $\mathcal{X}$ to a Hilbert space $\mathcal{Y}$. Given $\phi \in C^1(\mathcal{X}, \mathbb{R})$, its gradient is denoted by $\nabla \phi$, and if $\phi \in C^1(\mathcal{X}, \mathcal{X})$, the Jacobian of ϕ is denoted by $J\phi$. For every $x \in \mathcal{X}$, $J\phi(x)$ can be seen as a $d \times d$ matrix. The trace of the Jacobian, also called divergence, is denoted by $\operatorname{div} \phi$.

For any $d \times d$ matrix A, $\|A\|_{\mathrm{HS}}$ denotes the Hilbert Schmidt norm of A and $\|A\|_{\mathrm{op}}$ the operator norm of A viewed as a linear operator $A : \mathcal{X} \to \mathcal{X}$ (where $\mathcal{X}$ is endowed with the standard Euclidean inner product). Finally, δ_x is the Dirac measure at $x \in \mathcal{X}$.

2.2. Optimal transport

Consider $p \geq 1$. We denote by $\mathcal{P}_p(\mathcal{X})$ the set of Borel probability measures μ over $\mathcal{X}$ with finite p^{th} moment: $\int \|x\|^p d\mu(x) < \infty$. We denote by $L^p(\mu)$ the set of measurable functions $f : \mathcal{X} \to \mathcal{X}$ such that $\int \|f\|^p d\mu < \infty$. Note that the identity map I of $\mathcal{X}$ satisfies $I \in L^p(\mu)$ if $\mu \in \mathcal{P}_p(\mathcal{X})$. Moreover, denoting the image (or pushforward) measure of μ by a map T as $T\#\mu$, we have that if $\mu \in \mathcal{P}_p(\mathcal{X})$ and $T \in L^p(\mu)$ then $T\#\mu \in \mathcal{P}_p(\mathcal{X})$ using the transfer lemma.

For every $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, the p-Wasserstein distance between μ and ν is defined by

$$W_p^p(\mu, \nu) = \inf_{s \in \mathcal{S}(\mu, \nu)} \int \|x - y\|^p ds(x, y), \tag{2}$$

where $\mathcal{S}(\mu, \nu)$ is the set of couplings between μ and ν, *i.e.*, the set of nonnegative measures over $\mathcal{X}^2$ such that $P\#s = \mu$ (resp. $Q\#s = \nu$) where $P : (x, y) \mapsto x$ (resp. $Q : (x, y) \mapsto y$) denotes the projection onto the first (resp. the second) component. The p-Wasserstein distance is a metric over $\mathcal{P}_p(\mathcal{X})$. The metric space $(\mathcal{P}_2(\mathcal{X}), W_2)$ is called the Wasserstein space.

In this paper, we consider a target probability distribution π proportional to $\exp(-F)$, where F satisfies the following.

Assumption 2.1. The Hessian H_F is well-defined and $\exists L \geq 0$ such that $\|H_F\|_{\mathrm{op}} \leq L$.

Moreover, using $\int \exp(-F(x))dx < \infty$, F admits a stationary point.

Proposition 2.2. *Under Assumptions 2.1 , there exists $x_\star \in \mathcal{X}$ for which $\nabla F(x_\star) = 0$, i.e., F admits a stationary point.*

To specify the dependence in the dimension of our complexity bounds, we will initialize the algorithm from a Gaussian distribution centered at a stationary point. Such a stationary point can be found by gradient descent on F for instance.

The task of sampling from π can be formalized as an optimization problem. Indeed, define the Kullback-Leibler (KL) divergence as

$$\mathrm{KL}(\mu|\pi) \coloneqq \int \log\left(\frac{d\mu}{d\pi}(x)\right) d\mu(x), \tag{3}$$

if μ admits the density $\frac{d\mu}{d\pi}$ with respect to π, and $\mathrm{KL}(\mu|\pi) \coloneqq +\infty$ else. Then, $\mathrm{KL}(\mu|\pi) \geq 0$ and $\mathrm{KL}(\mu|\pi) = 0$ if and only if $\mu = \pi$. Therefore, assuming $\pi \in \mathcal{P}_2(\mathcal{X})$, the optimization problem

$$\min_{\mu \in \mathcal{P}_2(\mathcal{X})} \mathcal{F}(\mu), \tag{4}$$

where

$$\mathcal{F}(\mu) \coloneqq \mathrm{KL}(\mu|\pi),$$

admits a unique solution: the distribution π. We will see in Section 3 that SVGD can be seen as a gradient descent algorithm to solve (4).

Indeed, the Wasserstein space can be endowed with a differential structure. In particular, when it is well defined, the Wasserstein gradient of the functional $\mathcal{F}$ denoted by $\nabla_W \mathcal{F}(\mu)$ is an element of $L^2(\mu)$ and satisfies $\nabla_W \mathcal{F}(\mu) = \nabla \log\left(\frac{d\mu}{d\pi}\right)$.

2.2.1. Functional inequalities

The analysis of sampling algorithm in the case where F is nonconvex often goes through functional inequalities.

Definition 2.3 (Logarithmic Sobolev Inequality (LSI))**.** The distribution π satisfies the Logarithmic Sobolev Inequality if there exists $\lambda > 0$ such that for all $\mu \in \mathcal{P}_2(\mathcal{X})$,

$$\mathcal{F}(\mu) \leq \frac{2}{\lambda} \|\nabla_W \mathcal{F}(\mu)\|^2_{L^2(\mu)}.$$

LSI is a popular assumption in the analysis of Langevin algorithm in the case when F is not convex see *e.g.* Vempala & Wibisono (2019).

Definition 2.4 (Talagrand's Inequality Tp)**.** Let $p \geq 1$. The distribution π satisfies the Talagrand's Inequality Tp if there exists $\lambda > 0$ such that for all $\mu \in \mathcal{P}_p(\mathcal{X})$, we have $W_p(\mu, \pi) \leq \sqrt{\frac{2\mathcal{F}(\mu)}{\lambda}}$.

We now claim that T1 is milder than LSI. Indeed, using $W_1(\mu, \pi) \leq W_2(\mu, \pi)$, T2 implies T1 with the same constant λ. Moreover, using Villani (2008, Theorem 22.17), LSI implies T2 with the same constant λ. In conclusion, LSI $\Rightarrow$ T2 $\Rightarrow$ T1, with the same constant λ.

Besides, if F is λ-strongly convex, then π satisfies LSI with constant λ. A bounded perturbation of π in the latter case would also satisfies LSI with a constant independent of the dimension (Villani, 2008, Remark 21.5).

Finally, to get the exponential convergence of SVGD in continuous time, another inequality called Stein-LSI was proposed in Duncan et al. (2019). Stein-LSI is an assumption on both the kernel and the target distribution, and it implies LSI. Obtaining reasonable sufficient conditions for Stein-LSI to hold is an open problem, but there are simple cases where it cannot hold (Duncan et al., 2019, Lemma 36). In particular, Stein-LSI never holds under the assumptions that we will make in this paper to study SVGD in discrete time, see Korba et al. (2020, Section 11.3).

Our key assumption on π is that it satisfies the Talagrand's inequality T1 (Villani, 2008, Definition 22.1).

Assumption 2.5. The target distribution π satisfies T1.

We will use Assumption 2.5 to recursively control the KSD by the KL divergence along the iterations of the algorithm.

The target distribution π satisfies T1 if and only if there exist $a \in \mathcal{X}$ and $\beta > 0$ such that

$$\int \exp(\beta \|x - a\|^2) d\pi(x) < \infty, \tag{5}$$

see Villani (2008, Theorem 22.10). Therefore, Assumption 2.5 is essentially an assumption on the tails of π. In particular, $\pi \in \mathcal{P}_2(\mathcal{X})$.

2.3. Reproducing Kernel Hilbert Space

We consider a kernel k associated to a Reproducing Kernel Hilbert Space (RKHS) denoted by $\mathcal{H}_0$. We denote by $\Phi : \mathcal{X} \to \mathcal{H}_0$ the so-called feature map $\Phi : x \mapsto k(\cdot, x)$. The product space $\mathcal{H}_0^d$ is also a Hilbert space denoted $\mathcal{H} := \mathcal{H}_0^d$. We make the following assumption on the kernel k.

Assumption 2.6. There exists $B > 0$ such that the inequalities

$$\|\Phi(x)\|_{\mathcal{H}_0} \leq B,$$

and

$$\|\nabla \Phi(x)\|_{\mathcal{H}}^2 = \sum_{i=1}^{d} \|\partial_i \Phi(x)\|_{\mathcal{H}_0}^2 \leq B^2$$

hold for all $x \in \mathcal{X}$. Moreover, $\nabla \Phi : \mathcal{X} \to \mathcal{H}$ is continuous.

Assumption 2.6 is satisfied by the Gaussian kernel for example, with B independent of d using a scaling argument. Assumption 2.6 states that $\Phi : \mathcal{X} \to \mathcal{H}_0$ is bounded, Lipschitz and C^1. This is satisfied by many classical kernels used in practice. Note that $k(x, x) = \|\Phi(x)\|_{\mathcal{H}_0}^2$ and that $\operatorname{div}_1 \nabla_2 k(x, a) = \langle \nabla \Phi(x), \nabla \Phi(a) \rangle_{\mathcal{H}}$ (in particular, $\operatorname{div}_1 \nabla_2 k(x, x) = \|\nabla \Phi(x)\|_{\mathcal{H}}^2$). Hence, $\nabla \Phi$ is continuous iff $x \mapsto \operatorname{div}_1 \nabla_2 k(x, x)$ and $x \mapsto \operatorname{div}_1 \nabla_2 k(x, a)$ are continuous for every $a \in \mathcal{X}$.

Under Assumption 2.6, $\mathcal{H} \subset L^2(\mu)$ for every probability distribution on $\mathcal{X}$, and the inclusion map $\iota_\mu : \mathcal{H} \to L^2(\mu)$ is continuous. We denote by $P_\mu : L^2(\mu) \to \mathcal{H}$ its adjoint defined by the relation: for every $f \in L^2(\mu)$, $g \in \mathcal{H}$,

$$\langle f, \iota_\mu g \rangle_{L^2(\mu)} = \langle P_\mu f, g \rangle_{\mathcal{H}}. \tag{6}$$

Then, P_μ can be expressed as a convolution with k (Carmeli et al., 2010, Proposition 3):

$$P_\mu f(x) = \int k(x, y) f(y) d\mu(y), \tag{7}$$

or $P_\mu f = \int \Phi(y) f(y) d\mu(y)$ where the integral converges in norm.

2.4. Stein Variational Gradient Descent

2.4.1. The population limit

Stein Variational Gradient Descent (SVGD) is an algorithm to sample from $\pi \propto \exp(-F)$. SVGD proceeds by maintaining a set of N particles over $\mathbb{R}^d$, whose empirical distribution μ_n^N at time n aims to approximate π as $n \to \infty$, see Liu & Wang (2016). The SVGD algorithm is presented above.

Algorithm 1 Stein Variational Gradient Descent (Liu & Wang, 2016)

Initialization: a set $x_0^1, \dots, x_0^N \in \mathcal{X}$ of N particles, a kernel k, a step size $\gamma > 0$.
for $n = 0, 1, 2, \dots$ **do**
  for $i = 1, 2, \dots, N$ **do**

$$x_{n+1}^i = x_n^i - \frac{\gamma}{N} \sum_{j=1}^{N} k(x_n^i, x_n^j) \nabla F(x_n^j) - \nabla_2 k(x_n^i, x_n^j)$$

  end for
end for

Denoting by μ_n^N the empirical distribution of $x_n^1, \dots, x_n^N$, *i.e.*,

$$\mu_n^N := \frac{1}{N} \sum_{i=1}^{N} \delta_{x_n^i},$$

the SVGD update can be written

$$\begin{aligned} x_{n+1}^i &= x_n^i - \gamma \int k(x_n^i, y) \nabla F(y) - \nabla_2 k(x_n^i, y) d\mu_n^N(y) \\ &= \left(I - \gamma \int k(\cdot, y) \nabla F(y) - \nabla_2 k(\cdot, y) d\mu_n^N(y) \right)(x_n^i). \end{aligned}$$

Therefore, SVGD performs the update

$$\mu_{n+1}^N = \left(I - \gamma \int \Phi(y)\nabla F(y) - \nabla\Phi(y) d\mu_n^N(y)\right) \#\mu_n^N,$$

at the level of measures. We call *population limit* the regime where, formally, $N = \infty$. Mathematically, this corresponds to the assumption that μ_0 has a density (which can be seen as intuitively seen $\mathbb{E}\lim_{N\to\infty}\mu_0^N$) which belongs to $C_0(\mathcal{X})$. In this case, we shall see in our analysis that μ_n has a density for every n. To summarize, in the population limit, SVGD performs the same update:

$$\mu_{n+1} = (I - \gamma h_{\mu_n})\#\mu_n, \tag{8}$$

where

$$h_\mu(x) \coloneqq \int k(x,y)\nabla F(y) - \nabla_y k(x,y) d\mu(y)$$

or

$$h_\mu \coloneqq \int \Phi(y)\nabla F(y) - \nabla\Phi(y) d\mu(y),$$

and where μ_n has a density.

Finally, note that the SVGD algorithm was originally derived in Liu & Wang (2016) from its population limit. The authors first introduced the SVGD update in the population limit, and then, the SVGD algorithm (Algorithm 1) is obtained from the population limit by approximating the expectations by empirical means.

Our point of view on SVGD in the population limit. We now provide the intuition behind our results on SVGD.

> *In the population limit, SVGD can be seen as a Riemannian gradient descent, thanks to the following two reasons.*

First, in a Riemannian interpretation of the Wasserstein space (Villani, 2008), for every $\mu \in \mathcal{P}_2(\mathcal{X})$, the map $\exp_\mu : \phi \mapsto (I+\phi)\#\mu$ can be seen as the exponential map at μ. In the population limit, SVGD (8) can be rewritten as

$$\mu_{n+1} = \exp_{\mu_n}(-\gamma h_{\mu_n}).$$

Second, $-h_\mu$ can be seen as the negative gradient of $\mathcal{F}$ at μ under a certain metric. Indeed, using integration by parts, $h_\mu = P_\mu \nabla_W \mathcal{F}(\mu)$, see *e.g.* Korba et al. (2020); Duncan et al. (2019). Therefore, for every $g \in \mathcal{H}$, $\langle h_\mu, g\rangle_{\mathcal{H}} = \langle \nabla_W \mathcal{F}(\mu), g\rangle_{L^2(\mu)}$, hence h_μ can be seen as a Wasserstein gradient of $\mathcal{F}$ under the inner product of $\mathcal{H}$.

The Kernelized Stein Discrepancy (KSD) is a natural discrepancy between probability distributions that was introduced prior to SVGD (Liu et al., 2016; Chwialkowski et al., 2016) to compare probablity measures. Indeed, if the RKHS $\mathcal{H}$ is rich enough (Liu et al., 2016; Chwialkowski et al., 2016; Oates et al., 2019), an assumption that we shall always make in this paper, then

$$\mathrm{KSD}(\mu|\pi) = 0 \Longrightarrow \mu = \pi.$$

The KSD is intimately related to SVGD, and the KSD naturally appears in the original derivation of SVGD (Liu & Wang, 2016). The KSD is defined as the square root of the Stein Fisher Information (Duncan et al., 2019) I_{Stein}:

$$I_{\mathrm{Stein}}(\mu|\pi) \coloneqq \|h_\mu\|_{\mathcal{H}}^2, \quad \mathrm{KSD}(\mu|\pi) \coloneqq \|h_\mu\|_{\mathcal{H}}^2. \tag{9}$$

In this paper, we study the complexity of SVGD in terms of the KSD. To understand better the topology of the KSD and compare it to common topologies in the space of probability measures, we refer to Gorham & Mackey (2017).

3. Analysis of SVGD

In this section, we analyze SVGD in the infinite number of particles regime. Recall that in this regime, SVGD is given by $\mu_0 \in C_0(\mathcal{X})$ and

$$\mu_{n+1} = (I - \gamma h_{\mu_n})\#\mu_n,$$

where

$$h_\mu \coloneqq \int \nabla F(x)\Phi(x) - \nabla\Phi(x) d\mu(x).$$

3.1. A fundamental inequality

We start by stating a fundamental inequality satisfied by $\mathcal{F}$ for any update of the form

$$\mu_{n+1} = (I - \gamma g)\#\mu_n, \tag{10}$$

where $g \in \mathcal{H}$.

Proposition 3.1. *Let Assumptions 2.1 and 2.6 hold true. Let $\alpha > 1$ and choose $\gamma > 0$ such that $\gamma\|g\|_{\mathcal{H}} \leq \frac{\alpha-1}{\alpha B}$. Then,*

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) - \gamma\langle h_{\mu_n}, g\rangle_{\mathcal{H}} + \frac{\gamma^2 K}{2}\|g\|_{\mathcal{H}}^2, \tag{11}$$

where $K = (\alpha^2 + L)B$.

Inequality (11) is a property of *the functional* $\mathcal{F}$, and not a property of the SVGD algorithm. Inequality (11) plays the role of a *Taylor inequality* for the functional $\mathcal{F}$, where h_{μ_n} is the Wasserstein gradient of $\mathcal{F}$ at μ_n under the metric induced by $\mathcal{H}$. Proposition 3.1 is a slight generalization of Korba et al. (2020, Proposition 5), and is not our main contribution, therefore we only sketch its proof in the Appendix.

3.2. Main result

Applying recursively the Taylor inequality Proposition 3.1 with $g = h_{\mu_n}$, we obtain the following descent property for SVGD, which is our main theoretical result. The proof of this result can be found in the Appendix.

Theorem 3.2 (Descent lemma)**.** *Let Assumptions 2.1, 2.5 and 2.6 hold true. Let $\alpha > 1$. If*

$$\gamma \leq (\alpha - 1)\times \left(\alpha B^2 \left(1 + \|\nabla F(0)\| + L\int \|x\| d\pi(x) + L\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}}\right)\right)^{-1} \tag{12}$$

or

$$\gamma \leq (\alpha - 1)\times \left(\alpha B^2 \left(1 + 2L\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}} + L\int \|x - x_\star\| d\mu_0(x))\right)\right)^{-1} \tag{13}$$

then

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) - \gamma\left(1 - \frac{\gamma B(\alpha^2 + L)}{2}\right)\mathrm{KSD}^2(\mu_n|\pi). \tag{14}$$

If $\mathcal{F}(\mu_0) < \infty$, then, using Theorem 3.2, $(\mathcal{F}(\mu_n))_n$ is nonincreasing and μ_n has a density w.r.t. Lebesgue measure for every n (since $\mathcal{F}(\mu_n) < \infty$).

In the language of the gradient descent algorithms, Theorem 3.2 is called a descent property. It can be seen as a discrete time analogue of dissipation properties obtained for the PDE modeling SVGD in continuous time in the population limit (Duncan et al., 2019; Korba et al., 2020).

Unlike Korba et al. (2020, Proposition 5) and Liu (2017, Theorem 3.3), we do not assume that $\sup_n \mathrm{KSD}(\mu_n|\pi) < \infty$ or that $\gamma \leq \mathrm{KSD}(\mu_n|\pi)^{-1}$ to obtain our descent property. The step size γ is bounded by a constant. Iterating Theorem 3.2, we obtain convergence results as corollaries in the next section.

4. Convergence and Complexity

4.1. Convergence

We now show that Theorem 3.2 implies weak convergence and convergence in W_1.

Corollary 4.1 (Weak convergence)**.** *Let Assumptions 2.1, 2.5 and 2.6 hold true. Let $\alpha > 1$. If $\gamma < \frac{2}{B(\alpha^2+L)}$, and γ further satisfies either (12) or (13), then $\mu_n \to_{n\to+\infty} \pi$ weakly and $W_1(\mu_n, \pi) \to 0$.*

Proof. Using Theorem 3.2 and iterating,

$$\mathcal{F}(\mu_n) \leq \mathcal{F}(\mu_0) - \gamma\left(1 - \frac{\gamma B(\alpha^2 + L)}{2}\right)\sum_{k=0}^{n-1}\mathrm{KSD}^2(\mu_k|\pi).$$

Therefore, $\mathcal{F}(\mu_n)$ is uniformly bounded. For every $n \geq 1$,

$$\gamma\left(1 - \frac{\gamma B(\alpha^2 + L)}{2}\right)\sum_{k=0}^{n-1}\mathrm{KSD}^2(\mu_k|\pi) \leq \mathcal{F}(\mu_0).$$

Consequently, $\sum_{n=0}^{+\infty}\mathrm{KSD}^2(\mu_n|\pi) < \infty$. Therefore $\mathrm{KSD}(\mu_n|\pi) \to_{n\to+\infty} 0$.

Moreover, using Assumption 2.5 and (5), for every $a \in \mathcal{X}$, $\int \exp(\langle a, x\rangle) d\pi(x) < \infty$. Therefore, using Dupuis & Ellis (2011, Lemma 1.4.3), (μ_n) is both tight and uniformly integrable. Consider a subsequence of $(\mu_{\phi(n)})$ converging weakly to some $\mu_\star$. We shall prove that $\mu_\star = \pi$.

First, using Assumption 2.1 and Assumption 2.6, $x \mapsto \nabla F(x)\Phi(x) - \nabla\Phi(x) \in \mathcal{H}$ is continuous and

$$\begin{aligned}
&\|\nabla F(x)\Phi(x) - \nabla\Phi(x)\|_{\mathcal{H}} \\
&\|\nabla F(x)\Phi(x)\|_{\mathcal{H}} + \|\nabla\Phi(x)\|_{\mathcal{H}} \\
=& \|\nabla F(x)\|\,\|\Phi(x)\|_{\mathcal{H}_0} + \|\nabla\Phi(x)\|_{\mathcal{H}} \\
\leq& B\left(\|\nabla F(x)\| + 1\right) \\
\leq& B\left(\|\nabla F(0)\| + L\|x\| + 1\right).
\end{aligned}$$

Moreover, as a subsequence, $(\mu_{\phi(n)})$ is also uniformly integrable and also converges weakly to $\mu_\star$. Therefore, using Villani (2003, Theorem 7.12) with $p = 1$, $\mathbb{E}_{x\sim\mu_{\phi(n)}}\left(\nabla F(x)\Phi(x) - \nabla\Phi(x)\right)$ converges to $\mathbb{E}_{x\sim\mu_\star}\left(\nabla F(x)\Phi(x) - \nabla\Phi(x)\right)$ in $\mathcal{H}$. In other words, $h_{\mu_{\phi(n)}}$ converges to $h_{\mu_\star}$ in $\mathcal{H}$. Taking the norm, $\mathrm{KSD}(\mu_{\phi(n)}|\pi) \to \mathrm{KSD}(\mu_\star|\pi)$ along the subsequence. Recalling that $\mathrm{KSD}(\mu_n|\pi) \to 0$ we obtain $\mathrm{KSD}(\mu_\star|\pi) = 0$, which implies $\mu_\star = \pi$.

In conclusion, $\mu_n \to_{n\to+\infty} \pi$ weakly. Moreover, the convergence also happens in W_1 because (μ_n) is uniformly integrable, see (Villani, 2003, Theorem 7.12). □

In summary, under T1 and some smoothness assumptions but *without convexity of the potential*, SVGD in the population limit converges to the target distribution.

One can be surprised to see that SVGD converges without convexity assumption on F, but this is actually natural if one thinks about the gradient descent interpretation of SVGD. Indeed, SVGD in the population limit is a gradient descent on the KL divergence, which is

- "smooth" if we restrict the descent directions to a RKHS (*i.e.*, it satisfies a Taylor inequality Proposition 3.1),
- coercive (*i.e.*, sublevel sets are tight) Dupuis & Ellis (2011, Lemma 1.4.3),
- and has a single stationary point which is its global minimizer (the KSD is the norm of the gradient of KL

in our interpretation, and the KSD is equal to zero only at the optimum).

One can show that, over $\mathbb{R}^d$, gradient descent applied to a smooth coercive function with a single stationary point converges to the global minimizer. The situation here is similar.

4.2. Complexity

Next, we provide a $\mathcal{O}(1/n)$ convergence rate for the empirical mean of the iterates μ_n in terms of the squared KSD. This result is obtained from our descent lemma (Theorem 3.2).

Corollary 4.2 (Convergence rate)**.** *Let Assumptions 2.1, 2.5 and 2.6 hold true. Let $\alpha > 1$. If $\gamma < \frac{2}{B(\alpha^2+L)}$, and γ further satisfies either (12) or (13), then*

$$I_{\mathrm{Stein}}(\bar{\mu}_n|\pi) \leq \frac{2\mathcal{F}(\mu_0)}{n\gamma}, \tag{15}$$

where $\bar{\mu}_n = \frac{1}{n}\sum_{k=0}^{n-1}\mu_k$.

Note that this convergence rate is given in terms of the uniform mixture of $\mu_0, \dots, \mu_{n-1}$. Similar mixtures appear in the analysis of Langevin algorithm (see *e.g.* Durmus et al. (2019)). Note also that the convergence rate in Corollary 4.2 is similar to the convergence rate of the squared norm of the gradient in the gradient descent algorithm applied to a smooth function (Nesterov, 2013).

Proof. Using Theorem 3.2, $\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) - \frac{\gamma}{2}\mathrm{KSD}^2(\mu_n|\pi)$, and by iterating, we get

$$0 \leq \mathcal{F}(\mu_n) \leq \mathcal{F}(\mu_0) - \frac{\gamma}{2}\sum_{k=0}^{n-1}\|h_{\mu_k}\|^2.$$

Rearranging the terms, and using the convexity of the squared norm,

$$\|h_{\bar{\mu}_n}\|^2 = \left\|\frac{1}{n}\sum_{k=0}^{n-1}h_{\mu_k}\right\|^2 \leq \frac{1}{n}\sum_{k=0}^{n-1}\|h_{\mu_k}\|^2 \leq \frac{2\mathcal{F}(\mu_0)}{n\gamma}.$$

□

From the last result, we can characterize the iteration complexity of SVGD.

Corollary 4.3 (Complexity)**.** *Let Assumptions 2.1, 2.5 and 2.6 hold true. Let $\alpha > 1$. If $\gamma \leq \min(\frac{2}{B(\alpha^2+L)}, \frac{\alpha-1}{\alpha K})$, where*

$$K \coloneqq B^2\left(1 + 2L\sqrt{\frac{2}{\lambda}}\sqrt{F(x_\star) + \frac{d}{2}\log\left(\frac{L}{2\pi}\right)} + \sqrt{Ld}\right),$$

and if $\mu_0 = \mathcal{N}(x_\star, \frac{1}{L}I)$, then

$$n = \tilde{\Omega}\left(\frac{Ld^{3/2}}{\lambda^{1/2}\varepsilon}\right)$$

iterations of SVGD suffice to output $\mu \coloneqq \bar{\mu}_n$ such that $I_{\mathrm{Stein}}(\mu|\pi) \leq \varepsilon$.

To our knowledge, Corollary 4.3 provides the first *dimension-dependent* complexity result for SVGD. Its proof can be found in the appendix. The dependence of the T1 constant λ in the dimension d is subject to active research in optimal transport theory (Villani, 2008, Remark 22.11) and is out of the scope of this paper. Yet, using Villani (2008, Theorem 22.10, Equation 22.16), λ can be taken as

$$1/\lambda = \min_{a\in\mathcal{X},\beta>0}\frac{1}{\beta^2}\left(1 + \log\int\exp(\beta\|x-a\|^2)d\pi(x)\right).$$

Note that the output μ of the algorithm is a mixture of the iterates: $\mu = \bar{\mu}_n$. Besides, optimizing the complexity over α leads to involved calculations that do not change the overall complexity. To see this, note that the larger the step size γ, the smaller the complexity. But, even if the step size $\gamma = \min(\frac{2}{BL}, 1/K)$ were allowed, the overall complexity would be the same.

5. Conclusion

We proved that under T1 inequality and some smoothness assumptions on the kernel and the potential of the target distribution but without convexity, SVGD in the population limit converges weakly and in 1-Wasserstein distance to the target distribution. Moreover, we showed that SVGD reaches ε accuracy in terms of the squared Kernelized Stein Discrepancy after $\tilde{\Omega}\left(\frac{d^{3/2}}{\varepsilon}\right)$ iterations.

A possible extension of our work is to study SVGD under functional inequalities other than T1, such as (Bolley & Villani, 2005, Corollary 2.6 (i)) (which is weaker than T1) or the Poincaré inequality (in the form of Villani (2008, Theorem 22.25 (iii))). We claim that our approach can be extended to study SVGD in these settings.

Finally, an important and difficult open problem in the analysis of SVGD is to characterize its complexity with a finite number of particles, i.e. with discrete measures. In this regime, we lose the interpretation of SVGD as a gradient descent in the space of probability measures, because the KL divergence w.r.t. the target distribution is infinite. However, we believe that our clean analysis in the population limit makes a first step towards this open problem.

References

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, M. Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo. *arXiv preprint arXiv:2202.05214*, 2022.

Bernton, E. Langevin Monte Carlo and JKO splitting. In *Conference on Learning Theory (COLT)*, pp. 1777–1798, 2018.

Bolley, F. and Villani, C. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

Bubeck, S., Eldan, R., and Lehec, J. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory (COLT)*, pp. 300–323, 2018.

Chewi, S., Gouic, T. L., Lu, C., Maunu, T., and Rigollet, P. SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2098–2109, 2020.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pp. 2606–2615, 2016.

Dalalyan, A. S. 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, 2017.

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

Dupuis, P. and Ellis, R. S. *A weak convergence approach to the theory of large deviations*, volume 902. John Wiley & Sons, 2011.

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

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

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

Foster, J., Lyons, T., and Oberhauser, H. The shifted ode method for underdamped langevin mcmc. *arXiv preprint arXiv:2101.03446*, 2021.

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

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

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2878–2887, 2018.

Kassab, R. and Simeone, O. Federated generalized bayesian learning via distributed stein variational gradient descent. *arXiv preprint arXiv:2009.06419*, 2020.

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

Li, R., Zha, H., and Tao, M. Sqrt (d) dimension dependence of langevin monte carlo. *arXiv preprint arXiv:2109.03839*, 2021.

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

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

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

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

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

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

Nesterov, Y. E. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

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

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. Convergence rates for a class of estimators based on stein's method. *Bernoulli*, 25(2):1141–1159, 2019.

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

Rolland, P., Eftekhari, A., Kavis, A., and Cevher, V. Double-loop unadjusted Langevin algorithm. In *International Conference on Machine Learning (ICML)*, pp. 8169–8177, 2020.

Salim, A. and Richtárik, P. Primal dual interpretation of the proximal stochastic gradient Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3786–3796, 2020.

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

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

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

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

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

Villani, C. *Topics in optimal transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Society, 2003.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

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

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

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

Zou, D., Xu, P., and Gu, Q. Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2936–2945, 2019.

Appendix

A. Proof of Proposition 2.2

First, we prove that F is coercive, *i.e.*, for every $C > 0$, the set $S = \{x \in \mathcal{X} : F(x) \leq C\}$ is compact. Since F is continuous, S is closed. It remains to prove that S is bounded. Assume, by contradiction, that S is unbounded. Then, there exists a sequence (x_n) of points in $\mathcal{X}$ such that $F(x_n) \leq C$, $\|x_n\| \to +\infty$ and $B(x_n) \cap B(x_m) = \emptyset$ for every $n \neq m$, where $B(x)$ denotes the unit ball centered at x.

Let $n \geq 0$. Using the smoothness of F (Assumption 2.1), for every $x \in B(x_n)$,

$$F(x) \leq F(x_n) + \langle \nabla F(x_n), x - x_n \rangle + \frac{L}{2}.$$

Denote by V the volume of the unit ball centered at x, *i.e.* its Lebesgue measure. The positive number V does not depend on x. Then

$$\begin{aligned}
\int_{B(x_n)} \exp(-F(x))dx &\geq \int_{B(x_n)} \exp\left(-F(x_n) - \langle \nabla F(x_n), x - x_n \rangle - \frac{L}{2}\right) dx \\
&= V \exp\left(-F(x_n) - \frac{L}{2}\right) \int_{B(x_n)} \exp\left(\langle \nabla F(x_n), x_n - x \rangle\right) \frac{dx}{V} \\
&= V \exp\left(-F(x_n) - \frac{L}{2}\right) \int_{B(0)} \exp\left(\langle \nabla F(x_n), u \rangle\right) \frac{du}{V} \\
&\geq V \exp\left(-F(x_n) - \frac{L}{2}\right) \exp\left(\int_{B(0)} \langle \nabla F(x_n), u \rangle \frac{du}{V}\right) \\
&= V \exp\left(-F(x_n) - \frac{L}{2}\right) \\
&\geq V \exp\left(-C - \frac{L}{2}\right),
\end{aligned}$$

where we used Jensen's inequality for the uniform distribution over $B(0)$, thanks to the convexity of $t \mapsto \exp(t)$. Finally,

$$\int \exp(-F(x))dx \geq \sum_{n=0}^{\infty} \int_{B(x_n)} \exp(-F(x))dx \geq \sum_{n=0}^{\infty} V \exp\left(-C - \frac{L}{2}\right) = +\infty,$$

which means that $\exp(-F)$ is not integrable. This contradicts the definition of F and therefore, S is bounded.

Next, since the set S is compact, F is coercive, and hence F admits a stationary point. Indeed, F is continuous over the compact set $\{x \in \mathcal{X} : F(x) \leq 1\}$, and therefore, F admits a minimizer, $x_\star$, over this set. Moreover, this point $x_\star$ is a stationary point *i.e.*, $\nabla F(x_\star) = 0$ (note that the point $x_\star$ is actually a global minimizer of F).

B. Proof of Proposition 3.1

Let $\phi_t = I - tg$ for $t \in [0, \gamma]$ and $\rho_t = (\phi_t)\#\mu_n$. Note that $\rho_0 = \mu_n$ and $\rho_\gamma = \mu_{n+1}$. First, for every $x \in \mathcal{X}$,

$$\|g(x)\|^2 = \sum_{i=1}^{d} \langle k(x, .), g_i \rangle_{\mathcal{H}_0}^2 \leq \|k(x, .)\|_{\mathcal{H}_0}^2 \|g\|_{\mathcal{H}}^2 \leq B^2 \|g\|_{\mathcal{H}}^2, \tag{16}$$

and

$$\begin{aligned}\|Jg(x)\|_{\mathrm{HS}}^2 &= \sum_{i,j=1}^{d} \left|\frac{\partial g_i(x)}{\partial x_j}\right|^2 \\ &= \sum_{i,j=1}^{d} \langle \partial_{x_j} k(x,.), g_i\rangle_{\mathcal{H}_0}^2 \\ &\leq \sum_{i,j=1}^{d} \|\partial_{x_j} k(x,.)\|_{\mathcal{H}_0}^2 \|g_i\|_{\mathcal{H}_0}^2 \\ &= \|\nabla k(x,.)\|_{\mathcal{H}}^2 \|g\|_{\mathcal{H}}^2 \\ &\leq B^2 \|g\|_{\mathcal{H}}^2. \end{aligned} \tag{17}$$

Hence,

$$\|tJg(x)\|_{\mathrm{op}} \leq \|tJg(x)\|_{\mathrm{HS}} \leq \gamma B \|g\|_{\mathcal{H}} \leq \frac{\alpha - 1}{\alpha} < 1, \tag{18}$$

using our assumption on the step size γ. Inequality (18) proves that ϕ_t is a diffeomorphism for every $t \in [0, \gamma]$. Moreover,

$$\|(J\phi_t(x))^{-1}\|_{\mathrm{op}} \leq \sum_{k=0}^{\infty} \|tJg(x)\|_{\mathrm{op}}^k \leq \sum_{k=0}^{\infty} \left(\frac{\alpha - 1}{\alpha}\right)^k = \alpha. \tag{19}$$

Using the density of the pushforward formula,

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

Moreover, $\det(J\phi_t(x))^{-1} \leq \alpha^d$ for every $x \in \mathcal{X}$ using (19). Besides, if $\mu_n \in C_0(\mathcal{X})$ then $\mu_n \circ \phi_t^{-1} \in C_0(\mathcal{X})$ using that ϕ_t is a diffeomorphism. Therefore, $\rho_t \in C_0(\mathcal{X})$ as the product of a $C_0(\mathcal{X})$ function with a bounded function. In particular, $\mu_{n+1} \in C_0(\mathcal{X})$. By induction, $\mu_k \in C_0(\mathcal{X})$ for every k.

Using Villani (2003, Theorem 5.34), the velocity field ruling the time evolution of ρ_t is $w_t \in L^2(\rho_t)$ defined by $w_t(x) = -g(\phi_t^{-1}(x))$. Denote $\varphi(t) = \mathcal{F}(\rho_t)$. Using a Taylor expansion,

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^{\gamma} (\gamma - t)\varphi''(t)dt. \tag{20}$$

We now identify each term. First, $\varphi(0) = \mathcal{F}(\mu_n)$ and $\varphi(\gamma) = \mathcal{F}(\mu_{n+1})$. Using the reproducing property, one can show that

$$\varphi'(0) = -\langle h_{\mu_n}, g\rangle_{\mathcal{H}}. \tag{21}$$

Moreover, one can show that $\varphi''(t) = \psi_1(t) + \psi_2(t)$, where

$$\psi_1(t) = \mathbb{E}_{x\sim\rho_t}\left[\langle w_t(x), H_F(x)w_t(x)\rangle\right] \quad \text{and} \quad \psi_2(t) = \mathbb{E}_{x\sim\rho_t}\left[\|Jw_t(x)\|_{\mathrm{HS}}^2\right]. \tag{22}$$

Recall that $w_t = -g \circ (\phi_t)^{-1}$. The first term $\psi_1(t)$ is bounded using the transfer lemma, Assumption 2.1 and Inequality (16):

$$\psi_1(t) = \mathbb{E}_{x\sim\mu_n}\left[\langle g(x), H_V(\phi_t(x))g(x)\rangle\right] \leq L\|g\|_{L^2(\mu_n)}^2 \leq LB^2\|g\|_{\mathcal{H}}^2.$$

For the second term $\psi_2(t)$, using the chain rule, $-Jw_t \circ \phi_t = Jg(J\phi_t)^{-1}$. Therefore,

$$\|Jw_t \circ \phi_t(x)\|_{\mathrm{HS}}^2 \leq \|Jg(x)\|_{\mathrm{HS}}^2 \|(J\phi_t)^{-1}(x)\|_{\mathrm{op}}^2 \leq \alpha^2 B^2 \|g\|_{\mathcal{H}}^2,$$

using (17) and (19). Combining each of the quantity in the Taylor expansion (20) gives the desired result.

C. Proof of Theorem 3.2

We start with a Lemma.

Lemma C.1. *Let Assumptions 2.1, 2.5 and 2.6 hold true. Then, for every $\mu \in \mathcal{P}_2(\mathcal{X})$, we have*

$$\|h_\mu\|_{\mathcal{H}} \leq B\left(1 + \|\nabla F(0)\| + L\int \|x\| d\pi(x)\right) + BL\sqrt{\frac{2\mathcal{F}(\mu)}{\lambda}}$$

and

$$\|h_\mu\|_{\mathcal{H}} \leq B\left(1 + L\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}} + L\sqrt{\frac{2\mathcal{F}(\mu)}{\lambda}} + L\int \|x - x_\star\| d\mu_0(x)\right).$$

Proof. Using Assumption 2.6

$$\begin{aligned}
\|h_\mu\|_{\mathcal{H}} &= \|\mathbb{E}_{x\sim\mu}\left(\nabla F(x)\Phi(x) - \nabla\Phi(x)\right)\|_{\mathcal{H}} \\
&\leq \mathbb{E}_{x\sim\mu}\|\nabla F(x)\Phi(x) - \nabla\Phi(x)\|_{\mathcal{H}} \\
&\leq \mathbb{E}_{x\sim\mu}\|\nabla F(x)\Phi(x)\|_{\mathcal{H}} + \mathbb{E}_{x\sim\mu}\|\nabla\Phi(x)\|_{\mathcal{H}} \\
&= \mathbb{E}_{x\sim\mu}\|\nabla F(x)\|\,\|\Phi(x)\|_{\mathcal{H}} + \mathbb{E}_{x\sim\mu}\|\nabla\Phi(x)\|_{\mathcal{H}} \\
&\leq B\left(\mathbb{E}_{x\sim\mu}\|\nabla F(x)\| + 1\right).
\end{aligned}$$

Using Assumption 2.1 and Proposition 2.2, $\|\nabla F(x)\| \leq \|\nabla F(0)\| + L\|x\|$. Therefore, using the triangle inequality for the metric W_1,

$$\begin{aligned}
\|h_\mu\|_{\mathcal{H}} &\leq B\left(1 + \|\nabla F(0)\| + L\int \|x\| d\mu(x)\right) \\
&= B\left(1 + \|\nabla F(0)\| + LW_1(\mu, \delta_0)\right) \\
&\leq B\left(1 + \|\nabla F(0)\| + LW_1(\pi, \delta_0)\right) + BLW_1(\mu, \pi).
\end{aligned}$$

We obtain the first inequality using Assumption 2.5: $W_1(\mu,\pi) \leq \sqrt{\frac{2\mathcal{F}(\mu)}{\lambda}}$.

To prove the second inequality, recall that $\|h_\mu\|_{\mathcal{H}} \leq B\left(\mathbb{E}_{x\sim\mu}\|\nabla F(x)\| + 1\right)$. Using Assumption 2.1 and Proposition 2.2, $\|\nabla F(x)\| = \|\nabla F(x) - \nabla F(x_\star)\| \leq L\|x - x_\star\|$. Therefore, using the triangle inequality for the metric W_1,

$$\begin{aligned}
\int \|x - x_\star\| d\mu(x) = W_1(\mu, \delta_{x_\star}) &\leq W_1(\mu,\pi) + W_1(\pi,\mu_0) + W_1(\mu_0, \delta_{x_\star}) \\
&\leq \sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}} + \sqrt{\frac{2\mathcal{F}(\mu)}{\lambda}} + W_1(\mu_0, \delta_{x_\star}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|h_\mu\|_{\mathcal{H}} &\leq B\left(1 + L\int \|x - x_\star\| d\mu(x)\right) \\
&\leq B\left(1 + L\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}} + L\sqrt{\frac{2\mathcal{F}(\mu)}{\lambda}} + LW_1(\mu_0, \delta_{x_\star})\right). \qquad (23)
\end{aligned}$$

□

Besides, Proposition 3.1 can be applied to SVGD by setting $g = h_{\mu_n} \in \mathcal{H}$. In this case, we obtain the following descent property if the step size is small enough.

Lemma C.2. *Let Assumptions 2.1 and 2.6 hold true. Let $\alpha > 1$ and choose $\gamma > 0$ such that $\gamma\|h_{\mu_n}\|_{\mathcal{H}} \leq \frac{\alpha-1}{\alpha B}$. Then,*

$$\mathcal{F}(\mu_{n+1}) \leq \mathcal{F}(\mu_n) - \gamma\left(1 - \frac{\gamma K}{2}\right)\|h_{\mu_n}\|_{\mathcal{H}}^2, \qquad (24)$$

where $K = (\alpha^2 + L)B$.

Contrary to Inequality (11), Inequality (24) is a property of *the SVGD algorithm.*

Having established Proposition 3.1 and Lemmas C.2 and C.1, we are now ready to formulate and prove our main Theorem 3.2.

Proof. We now prove by induction the first implication of Theorem 3.2: (12) $\Rightarrow$ (14). First, if $\gamma > 0$ satisfies (12), then, using Lemma C.1, $\gamma\|h_{\mu_0}\|_{\mathcal{H}} \leq \frac{\alpha-1}{\alpha B}$. Therefore, using Lemma C.2,

$$\mathcal{F}(\mu_1) \leq \mathcal{F}(\mu_0) - \gamma\left(1 - \frac{\gamma K}{2}\right)\|h_{\mu_0}\|_{\mathcal{H}}^2,$$

i.e., Inequality (14) holds with $n = 0$. Now, assume that the condition (12) implies Inequality (14) for every $n \in \{0, \dots, N-1\}$ and let us prove it for $n = N$. First, $\mathcal{F}(\mu_N) \leq \mathcal{F}(\mu_0)$. Letting $A := B\left(1 + \|\nabla F(0)\| + L\int \|x\| d\pi(x)\right)$, this implies

$$A + BL\sqrt{\frac{2\mathcal{F}(\mu_N)}{\lambda}} \leq A + BL\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}}.$$

Therefore, if $\gamma > 0$ satisfies (12), then $\gamma\|h_{\mu_N}\|_{\mathcal{H}} \leq \frac{\alpha-1}{\alpha B}$. To see this, using Lemma C.1 we obtain

$$\gamma\|h_{\mu_N}\|_{\mathcal{H}} \leq \gamma\left(A + BL\sqrt{\frac{2\mathcal{F}(\mu_N)}{\lambda}}\right) \leq \gamma\left(A + BL\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}}\right) \leq \frac{\alpha-1}{\alpha B}.$$

Therefore, using Lemma C.2, the condition (12) implies Inequality (14) at step $n = N$:

$$\mathcal{F}(\mu_{N+1}) \leq \mathcal{F}(\mu_N) - \gamma\left(1 - \frac{\gamma K}{2}\right)\|h_{\mu_N}\|_{\mathcal{H}}^2.$$

Finally, it remains to recall that $\|h_{\mu_N}\|_{\mathcal{H}}^2 = \mathrm{KSD}^2(\mu_N|\pi)$. The proof of the second implication of Theorem 3.2, (13) $\Rightarrow$ (14), is similar. □

D. Proof of Corollary 4.3

Using Corollary 4.2, if

$$\gamma \leq \min\left((\alpha-1)\left(\alpha B^2\left(1 + 2L\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}} + L\int \|x - x_\star\| d\mu_0(x)\right)\right)^{-1}, \frac{2}{B(\alpha^2 + L)}\right),$$

then,

$$\mathrm{KSD}^2(\bar{\mu}_n|\pi) \leq \frac{2\mathcal{F}(\mu_0)}{n\gamma}.$$

Using Vempala & Wibisono (2019, Lemma 1), $\mathcal{F}(\mu_0) \leq F(x_\star) + \frac{d}{2}\log\left(\frac{L}{2\pi}\right)$. Besides,

$$\int \|x - x_\star\| d\mu_0(x) = \mathbb{E}_{X\sim\mu_0}\|X - x_\star\| = \frac{1}{\sqrt{L}}\mathbb{E}_{X\sim\mu_0}\|\sqrt{L}(X - x_\star)\|,$$

and using the transfer lemma and Cauchy-Schwartz inequality,

$$\int \|x - x_\star\| d\mu_0(x) = \frac{1}{\sqrt{L}}\mathbb{E}_{Y\sim\mathcal{N}(0,I)}\|Y\| \leq \frac{1}{\sqrt{L}}\left(\mathbb{E}_{Y\sim\mathcal{N}(0,I)}\|Y\|^2\right)^{1/2} = \sqrt{\frac{d}{L}}.$$

Therefore,

$$\begin{aligned}
&(\alpha-1)\left(\alpha B^2\left(1+2L\sqrt{\frac{2\mathcal{F}(\mu_0)}{\lambda}}+L\int\|x-x_\star\|d\mu_0(x)\right)\right)^{-1}\\
\geq&(\alpha-1)\left(\alpha B^2\left(1+2L\sqrt{\frac{2}{\lambda}}\sqrt{F(x_\star)+\frac{d}{2}\log\left(\frac{L}{2\pi}\right)}+\sqrt{Ld}\right)\right)^{-1}\\
=&\tilde{\Omega}\left(\frac{1}{\frac{L\sqrt{d}}{\sqrt{\lambda}}+\sqrt{Ld}}\right),
\end{aligned}$$

and

$$\gamma^{-1}=\tilde{\mathcal{O}}\left(\frac{L\sqrt{d}}{\sqrt{\lambda}}+\sqrt{Ld}+L\right)=\tilde{\mathcal{O}}\left(\frac{L\sqrt{d}}{\sqrt{\lambda}}\right).$$

Since $\mathcal{F}(\mu_0)=\tilde{\mathcal{O}}(d)$,

$$\frac{\mathcal{F}(\mu_0)}{\gamma}=\tilde{\mathcal{O}}\left(\frac{Ld^{3/2}}{\sqrt{\lambda}}\right).$$

Let $\varepsilon>0$. To output the mixture $\bar{\mu}_n$ such that $\mathrm{KSD}^2(\bar{\mu}_n|\pi)<\varepsilon$, it suffices to ensure that $\frac{2\mathcal{F}(\mu_0)}{n\gamma}<\varepsilon$. Therefore, $n=\frac{2\mathcal{F}(\mu_0)}{\gamma\varepsilon}=\tilde{\Omega}\left(\frac{Ld^{3/2}}{\varepsilon\sqrt{\lambda}}\right)$ iterations suffice.