Stein Boltzmann Sampling: A Variational Approach for Global Optimization

Gaëtan Serré, Argyris Kalogeratos, Nicolas Vayatis

Stein Boltzmann Sampling: A Variational Approach for Global Optimization

Gaëtan Serré [1] **Argyris Kalogeratos** [1] **Nicolas Vayatis** [1]

Abstract

In this paper, we introduce a new flow-based method for global optimization of Lipschitz functions, called *Stein Boltzmann Sampling* (SBS). Our method samples from the Boltzmann distribution that becomes asymptotically supported over the set of the minimizers of the function to be optimized. Candidate solutions are sampled via the *Stein Variational Gradient Descent* algorithm. We prove the asymptotic convergence of our method, introduce two SBS variants, and provide a detailed comparison with several state-of-the-art global optimization algorithms on various benchmark functions. The design of our method, the theoretical results, and our experiments, suggest that SBS is particularly well-suited to be used as a continuation of efficient global optimization methods as it can produce better solutions while making a good use of the budget.

1. Introduction

In this paper, we consider global optimization of an unknown Lipschitz continuous, a priori nonconvex, function. Optimizing an unknown function is a typical situation in real applications: hyperparameter calibration or complex system design emerge in several domains, such as biology, physics simulation, epidemiology, machine learning (e.g. (Pintér, 1991; Lee et al., 2017)). For this, sequential methods are usually employed, which means that at each iteration the algorithm uses information extracted from the previous candidate solutions to propose the new ones. Many sequential and stochastic methods has been introduced to address this problem. Recent results (Zhang et al., 2020; Davis et al., 2022; Jordan et al., 2023) showed that only stochastic algorithms can approximate optimal points of an arbitrary Lipschitz function, when considering a relaxed (but still meaningful) optimality criterion.

Sequential methods rely on two components: a sampling process to explore the search space, and a selection process to choose the next candidate solution using the information given by the previous samples. In this work, we introduce a new sequential, flow-based and deterministic method called *Stein Boltzmann Sampling* (SBS) for Lipschitz functions. Our method uses the *Stein Variational Gradient Descent* (SVGD) (Liu & Wang, 2016) method to sample from the Boltzmann distribution, which has the characteristic that tends to a distribution supported over the set of the minimizers. SVGD constructs a flow in the space of probability measures (similarly to the way a gradient flow would evolve in $\mathbb{R}^d$) that moves towards the target sampling measure. Even though our method is not a typical stochastic one (since SVGD sampling is deterministic), we prove its asymptotic convergence for any Lipschitz function using elements of the SVGD theory. We show that the SBS method achieves competitive performance on standard global optimization benchmarks versus three stochastic state-of-the-art methods. The first one, ADALIPO (Malherbe & Vayatis, 2017), is consistent over Lipschitz functions and is adapted for a very low computational budget (i.e. function evaluations at candidate minimizers). The second and third ones, CMA-ES (Hansen & Ostermeier, 1996; 2001; Hansen et al., 2003) and WOA (Mirjalili & Lewis, 2016), are two inconsistent methods but known to be very efficient in practice. Due to either early stopping conditions or time complexity, these three existing methods do not scale computationally well, hence they are not suited for when the available computational budget is low.

The contributions of this paper are as follows: First, we provide a new proof of the asymptotic convergence of SVGD, implying the consistency of our method. For the sake of completeness, we also provide proofs of all background results in the Appendix. To ensure ensure the correctness and reproducibility of the technical proofs, for some of the results (background or not), we provide links to proofs in Lean, a proof assistant (de Moura & Ullrich, 2021; mathlib Community, 2020). Then, we introduce two SBS variants:

[1]Centre Borelli, Department of Mathematics, École Normale Supérieure Paris-Saclay, CNRS, 91190, Gif-Sur-Yvette, France. Correspondence to: Gaëtan Serré <gaetan.serre@ens-paris-saclay.fr>, Argyris Kalogeratos <argyris.kalogeratos@ens-paris-saclay.fr>, Nicolas Vayatis <nicolas.vayatis@ens-paris-saclay.fr>.

one that uses particle filtering to reduce the budget needed, and a hybrid second one that uses SBS as a continuation of CMA-ES or WOA to combine their efficiency with the consistency and scalability of our method. The goal is to provide methods that make more efficient use of the computational budget, for future real-world applications. Finally, we provide a detailed comparison of our method with the three aforementioned state-of-the-art methods on several global optimization benchmarks. We also interpret the attraction and repulsion forces of SVGD in the context of global optimization.

Notations. We consider the following notations: $d \in \mathbb{N}$ is the dimension of the optimization problem; $f : \Omega \to \mathbb{R}$ is the function to optimize, its domain $\Omega \subset \mathbb{R}^d$ is a compact set; $x^* \in X^*$ is one of the global minima of f, i.e. $\forall x^*,\ f^* = f(x^*)$. Moreover, $\lambda : \mathfrak{B}^d \to \mathbb{R}_{\geq 0}$ is the standard Lebesgue measure on the Borel algebra of $\mathbb{R}^d$. Given an arbitrary function f, its support is $\mathrm{supp}(f) = \{x \in \Omega \mid f(x) \neq 0\}$. We denote by C^p the set of p-times continuously differentiable functions, and by $C_c^\infty(\Omega)$ the set of smooth functions on Ω that have compact support. Given two measurable spaces (Ω_1, Σ_1) and (Ω_2, Σ_2), a measurable function $f : \Sigma_1 \to \Sigma_2$ and a measure μ over Σ_1, let $f_{\#}\mu$ denote the pushforward measure, i.e.

$$\forall B \in \Sigma_2,\ f_{\#}\mu(B) = \mu(f^{-1}(B)).$$

2. Stein Boltzmann Sampling

2.1. The proposed method

We introduce the *Stein Boltzmann Sampling* (SBS) method. The idea is to sample from a distribution that converges asymptotically to a distribution supported over the set of minimizers X^* of an arbitrary continuous function f. We use the continuous Boltzmann distribution (BD) for this purpose.

Definition 2.1 (Continuous Boltzmann distribution)**.** Given a function $f \in C^0(\Omega, \mathbb{R})$, the Boltzmann distribution over f is induced by the probability density function $m_{f,\Omega}^{(\kappa)} : \Omega \to \mathbb{R}_{\geq 0}$ defined by:

$$m_{f,\Omega}^{(\kappa)}(x) = m^{(\kappa)}(x) = \frac{e^{-\kappa f(x)}}{\int_\Omega e^{-\kappa f(t)} \mathrm{d}t},\quad \forall \kappa \in \mathbb{R}_{\geq 0}. \quad (1)$$

A characteristic property of the BD is that it tends to distribution supported over the set of minimizers X^* as κ tends to infinity. If $\lambda(X^*) > 0$, the BD tends to a uniform distribution over X^* (see Figure 1). If X^* is finite, it tends to a sum of Dirac distribution over X^* where the weight on each minimizer depends on the local geometry of the function (Hwang, 1980). More details can be found in Section 3.1. The SBS method aims to sample from the BD with κ large enough in order for the function values at the sampled points

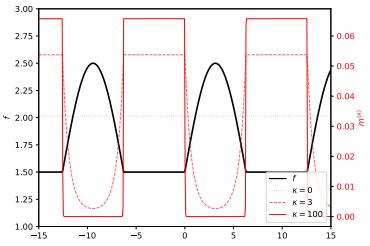


Figure 1. The Boltzmann p.d.f. becomes uniform over the set of minimizers X^* of the given function f to optimize, as κ grows, tending to infinity.

Algorithm 1 Stein Boltzmann Sampling (SBS)

Input: $f : \Omega \to \mathbb{R}$; number of vectors (particles) N; Boltzmann parameter κ; step-size ε; number of SVGD iterations n; an initial distribution $\hat{\mu}_1$ over the particles
Output: $\hat{x}$, an estimate of x^*

Sample N particles: $X_1 \leftarrow \left(x^{(1)}, \ldots, x^{(N)}\right) \sim \hat{\mu}_1^{\otimes N}$
for $i = 1$ **to** n **do**
  Compute the vector field $\phi^\star_{\hat{\mu}_i}$ (see Section 2.1)
  $X_{i+1} \leftarrow X_i + \varepsilon \phi^\star_{\hat{\mu}_i}(X_i)$ update of the particle system
  $\hat{\mu}_{i+1} \leftarrow \frac{1}{N}\sum_{j=1}^N \delta_{X_{i+1}^{(j)}}$ empirical measure over the particles
end for
$\hat{x} \leftarrow \arg\min_{1 \leq j \leq N} f(X_{n+1}^{(j)})$ the "best" particle
return $\hat{x}$

to be close to the global minimum. As $m^{(\kappa)}$ converges to a distribution supported over X^*, the approximation of f^* can be made arbitrarily accurate. However, as it is not efficient to sample from BD by estimating the intractable term $\int_\Omega e^{-\kappa f(t)} \mathrm{d}t$ using classical Monte-Carlo methods, we propose to use instead the *Stein Variational Gradient Descent* (SVGD) method. Given an initial measure μ, SVGD constructs iteratively a flow of measures that moves towards the target measure, noted as π. The update direction is given by:

$$\phi^\star_\mu = \mathbb{E}_{x \sim \mu}\left[\nabla \log \pi(x) k(\cdot, x) + \nabla_x k(\cdot, x)\right],$$

where k is the reproducing kernel of a specific RKHS $\mathcal{H}$ (see Section 3.2 for more details). In our case, π is the BD. As it appears within a gradient-log term, we do not need the normalization constant of the BD to compute $\phi^\star_\mu$. The pseudocode of the proposed SBS method can be found in Algorithm 1. Next in this section, we prove the asymptotic convergence of SBS.

2.2. Asymptotic convergence of SBS

To prove the asymptotic convergence of SBS, we need to prove that the sequence of measures constructed by SVGD, noted $(\mu_n)_{n \in \mathbb{N}}$ (see Equation (6)), converges to the measure induced by the BD, noted π. To do so, we need to study the flow of measures induced by the update direction of SVGD. Theorem 2.3 and Theorem 2.2 are known results in the literature. We provide a different proof for the latter

in Appendix A.6. Then, we introduce two lemmas that are crucial to prove the asymptotic convergence of SBS for any Lipschitz function, under absolute continuity assumptions on π and μ.

Theorem 2.2 (Time derivative of measure flow (Liu, 2017))**.** *Let $\phi : \mathbb{R}_{\geq 0} \times \Omega \to \Omega$, $\phi(t, \cdot) = \phi_t(\cdot)$ be a vector field. Let $(T_t)_{0 \leq t} : \Omega \to \Omega$ be a locally Lipschitz family of diffeomorphisms, representing the trajectories associated with the vector field ϕ_t, and such that $T_0 = I_d$. Let $\mu_t = T_{t\#}\mu$. Then, the following linear transport equation holds*

$$\begin{cases} \frac{\partial \mu_t}{\partial t} = -\nabla \cdot (\phi_t \mu_t), \forall t > 0 \\ \mu_0 = \mu \end{cases} \tag{2}$$

where $(\nabla \cdot)$ is the divergence operator, in the sense of distributions (see details in Appendix A.6). Moreover, the sequence $(\mu_n)_{n \in \mathbb{N}}$, constructed by Equation (6), is a discretized solution of the linear transport equation, considering the vector field $\phi^\star_{\mu_t}$. One can consider the resulting flow of measures

$$\Phi : \mathbb{R}_{\geq 0} \times \mathcal{P}_2(\Omega) \to \mathcal{P}_2(\Omega),$$
$$(t, \mu) \mapsto \Phi_t(\mu) = \mu_t.$$

We provide a different proof of this theorem in Appendix A.6, using optimal transport theory. This proof is more general in T but less constructive. We also prove that that sequence $(\mu_n)_{n \in \mathbb{N}}$ is an asymptotic solution of Equation (2). The latter equation has also been deeply studied in (Lu et al., 2019). This result allows to study the time-derivative of the KL-divergence between μ_t and π. Let S_μ be an integral operator associated to $\mathcal{H}$ and $\mathfrak{K}(\mu|\pi)$ a discrepancy measure between two measures μ and π called *Kernelized Stein Discrepancy* (KSD). Both objects are defined in Section 3.2. We have the following result.

Theorem 2.3 (Time-derivative of the KL-divergence (Liu, 2017))**.** *Let $(T_t)_{0 \leq t} : \Omega \to \Omega$ be a locally Lipschitz family of diffeomorphisms, representing the trajectories associated with the vector field $\phi^\star_{\mu_t} = S_{\mu_t} \nabla \log \frac{\pi}{\mu_t}$, such that $T_0 = I_d$. Let $\mu_t = T_{t\#}\mu$. Then, the time derivative of the KL-divergence between μ_t and π is given by*

$$\frac{\partial KL(\mu_t||\pi)}{\partial t} = -\mathfrak{K}(\mu_t|\pi).$$

Furthermore, as $\mathfrak{K}(\mu_t|\pi)$ is nonnegative, the KL-divergence is non-increasing along the flow of measures.

(See proof in Appendix A.7). In order to show the convergence of continuous-time SVGD, we proved that the KSD is a valid discrepancy measure. Let denote the absolutely continuity of measure μ w.r.t. π by $\mu \ll \pi$.

Lemma 2.4 (KSD valid discrepancy)**.** *Let $\mu, \pi \in \mathcal{P}_2(\Omega)$ such that $\mu \ll \pi \ll \lambda$. Then,*

$$\mu = \pi \iff \mathfrak{K}(\mu|\pi) = 0.$$

(See proof in Appendix A.8). The previous lemma directly implies that π is the unique fixed point of the flow of measures Φ.

Lemma 2.5 (Unique fixed point)**.** *Let $\pi \in \mathcal{P}_2(\Omega)$ such that $\pi \ll \lambda$. Let Φ be the flow of measures defined in Theorem 2.2. Let E be the set of measures in $\mathcal{P}_2(\Omega)$ that are absolutely continuous w.r.t. π. Then, for any $t \geq 0$, π is the unique fixed point of $(\mu : E) \mapsto \Phi_t(\mu)$.*

Since $\mathfrak{K}(\mu|\pi) = \left\|\phi^\star_\mu\right\|^2_{\mathcal{H}}$ (see Section 3.2), the proof is straightforward using the previous lemma. See complete proof in Appendix A.9. Finally, we provide a proof of the weak convergence of μ_t to π.

Theorem 2.6 (Weak convergence of SVGD)**.** *Let $\mu, \pi \in \mathcal{P}_2(\Omega)$ such that $\mu \ll \pi \ll \lambda$ and $KL(\mu||\pi) < \infty$. Let $(T_t)_{0 \leq t} : \Omega \to \Omega$ be a locally Lipschitz family of diffeomorphisms, representing the trajectories associated with the vector field $\phi^\star_{\mu_t} = S_{\mu_t} \nabla \log \frac{\pi}{\mu_t}$, such that $T_0 = I_d$. Let $\mu_t = T_{t\#}\mu$. Then, we have that*

$$\mu_t \rightharpoonup \pi.$$

See proof in Appendix A.10. The proof relies on Theorem 2.3 and Lemma 2.5; it is inspired by the proof of Theorem 2.8 in (Lu et al., 2019).

2.3. Discrete setting

In practice, SVGD is a discrete time algorithm that iteratively updates a set of particles and not a continuous measure μ. It starts by sampling a sequence of particles from a distribution μ: $X = \left(x^{(1)}, \dots, x^{(N)}\right)$, and then computes the next ones as follows:

$$X_{n+1} = X_n + \varepsilon \phi^\star_{\hat{\mu}_n}(X_n),$$
$$\text{where } \hat{\mu}_n(A) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_n^{(i)}}(A). \tag{3}$$

The previous results are sufficient to show the main theoretical result concerning SBS: its asymptotic convergence in discrete setting.

Theorem 2.7 (SBS asymptotic convergence)**.** *Let $f : \Omega \to \mathbb{R}$ be a Lipschitz function. Let $\kappa > 0$ and let π be the BD defined in Definition 2.1 associated with f and κ. Let $\hat{\mu}_0 \in \mathcal{P}_2(\Omega)$ such that $\hat{\mu}_0 \ll \lambda$, $\text{supp}(\hat{\mu}_0) = \Omega$ and $KL(\hat{\mu}_0||\pi) < \infty$. Let $\hat{\mu}_n$ be defined by Equation (3). Then, when $\varepsilon \to 0$,*

$$\left\{ f\left(X^{(n)}\right) \;\middle|\; X^{(n)} = (x^{(1)}, \dots, x^{(N)}) \sim \hat{\mu}_n^{\otimes N} \right\} \xrightarrow[\substack{\kappa \to \infty \\ N \to \infty \\ n \to \infty}]{\mathcal{L}} \{f^*\}.$$

Proof. The Rademacher's theorem states that ∇f exists almost everywhere, and therefore $\nabla \log m^{(\kappa)}$ also exists a.e. The rest of the proof is a direct application of Theorem 2.6 and Properties 3.1. ■

The implementation of SBS uses Equation (3) and estimate the gradients using finite differences. At each iteration, it updates the set of particles in the direction induced by $\phi^\star_{\hat{\mu}_n}$ by a small step size, computed using the Adam optimizer (Kingma & Ba, 2015). We choose the initial distribution $\hat{\mu}_1$ to be the uniform distribution on Ω as it maximizes the entropy (related to the exploration aspect of the method) and meets all the requirements of the SVGD theory. To better understand the previous results and objects involved, we introduce an non-exhaustive list of definitions and theoretical results related to SVGD in the next section.

3. Background

In this section, we introduce some background results related to the Boltzmann distribution (BD) and the theory related to SVGD.

3.1. Boltzmann distribution

Recall that the BD has been formally defined in Definition 2.1. The BD is a well-known distribution in statistical physics. It is used to model the distribution of the energy of a system in thermal equilibrium. The parameter κ is called the *inverse temperature*. The higher κ is, the more concentrated the mass is around the minima of f. When κ tends to infinity, the BD tends to a distribution supported over the minima of f. The BD is typically used in a discrete settings, i.e. where the number of states is finite. The continuous version can be defined using the Gibbs measure. The following properties come from (Luo, 2019). For the sake of completeness, we provide the proofs in Appendix A.1.

Properties 3.1 (Properties of the Boltzmann distribution)**.** Let $m^{(\kappa)}$ be defined as in Definition 2.1. Then, we have the following properties:

- If $\lambda(X^*) = 0$, then, $\forall x \in \Omega$,

$$\lim_{\kappa\to\infty} m^{(\kappa)}(x) = \begin{cases} \infty & \text{if } x \in X^* \\ 0 & \text{otherwise.} \end{cases}$$

- If $0 < \lambda(X^*)$, then, $\forall x \in \Omega$,

$$\lim_{\kappa\to\infty} m^{(\kappa)}(x) = \begin{cases} \frac{1}{\lambda(X^*)} & \text{if } x \in X^* \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall f \in C^0(\Omega, \mathbb{R})$,

$$\lim_{\kappa\to\infty} \int_\Omega f(x)\, m^{(\kappa)}(x)\mathrm{d}x = f^*.$$

A visual representation of the BD is given in Figure 1. One can see that, as κ increases, $m^{(\kappa)}$ becomes more and more concentrated around the minima of f. We use the BD induced by the density $m^{(\kappa)}$ (also noted $m^{(\kappa)}$ for simplicity) of Equation (1). To sample from tat distribution, we need to compute the integral $\int_\Omega e^{-\kappa f(t)}\mathrm{d}t$, which however, is likely to be intractable for a general f.

3.2. Stein Variational Gradient Descent

Sampling from an intractable distribution is a common task in Bayesian inference, where the target distribution is a posterior. Computation becomes difficult due to the presence of an intractable integral within the likelihood. The *Stein Variational Gradient Descent* (Liu & Wang, 2016) is a method that transforms iteratively an arbitrary measure μ to a target measure π. In the case of SBS, π is the BD defined in Definition 2.1, for any $\kappa > 0$. The algorithm is based on the *Stein method* (Stein, 1972). The theory of SVGD has been developed in several works over the years. Note that recently, (Korba et al., 2021) introduced a new sampling algorithm based on the same objective to SVGD, though less sensitive to the choice of the step-size. The remainder of this section highlights some key definitions and theoretical results related to SVGD.

3.2.1. Definitions

We start by defining the set of probability measures on Ω that have finite n-th moment. Let $\mathcal{P}_n(\Omega)$ denote the set of probability measures on $(\Omega, \mathcal{A})$ such that

$$\forall \mu \in \mathcal{P}_n(\Omega),\ \int_\Omega \|x\|^n \mathrm{d}\mu(x) < \infty.$$

In SVGD theory, μ and π must belong to $\mathcal{P}_2(\Omega)$, and they must be absolutely continuous w.r.t. λ, i.e.

$$\forall A \subseteq \Omega,\ \lambda(A) = 0 \implies \mu(A) = 0 \wedge \pi(A) = 0.$$

Moreover, for $KL(\mu||\pi)$ to be well-defined, μ must be absolutely continuous w.r.t. π. As the absolutely continuous relation is transitive, we can note $\mu \ll \pi \ll \lambda$. In the following, we denote the density w.r.t. λ of an arbitrary measure μ by the function $\mu : \Omega \to \mathbb{R}_{\geq 0}$.

3.2.2. Stein discrepancy

The Stein method defines the Stein operator associated to a measure μ (Liu, 2017):

$$\begin{aligned} \mathcal{A}_\mu :\ & C^1(\Omega, \Omega) \to C^0(\Omega, \mathbb{R}), \\ & \phi \mapsto \nabla \log \mu(\cdot)^\top \phi(\cdot) + \nabla \cdot \phi(\cdot), \end{aligned}$$

where (∇) is the gradient operator and $(\nabla\cdot)$ is the divergence operator. We denote this mapping by $\mathcal{A}_\mu\phi$, for any ϕ in $C^1(\Omega, \Omega)$. It also defines a class of functions, the Stein class of measures.

Definition 3.2 (Stein class of measures (Liu et al., 2016))**.** Let $\mu \in \mathcal{P}_2(\Omega)$ such that $\mu \ll \lambda$, and let $\phi : \Omega \to \Omega$. As Ω

is compact, the boundary of Ω (denoted by $\partial\Omega$) is nonempty. We say that ϕ is in the *Stein class* of μ if

$$\oint_{\partial\Omega} \mu(x)\phi(x) \cdot \vec{n}(x)\mathrm{d}S(x) = 0,$$

where $\vec{n}(x)$ is the unit normal vector to the boundary of Ω. We denote by $\mathcal{S}(\mu)$ the Stein class of μ.

The key property of $\mathcal{S}(\mu)$ is that, for any function f in $\mathcal{S}(\mu)$, the expectation of $\mathcal{A}_\mu f$ w.r.t. μ is null.

Lemma 3.3 (Stein identity (Stein, 1972))**.** *Let $\mu \in \mathcal{P}_2(\Omega)$ such that $\mu \ll \lambda$, and let $\phi \in \mathcal{S}(\mu)$. Then,*

$$\mathbb{E}_{x\sim\mu}[\mathcal{A}_\mu\phi(x)] = 0.$$

(See proof in Appendix A.2). Now, one can consider:

$$\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi\phi(x)] \text{ , where } \phi \in \mathcal{S}(\pi). \tag{4}$$

If $\mu \neq \pi$, Equation (4) would no longer be null for any ϕ in $\mathcal{S}(\pi)$. In fact, the magnitude of this expectation relates to how different μ and π are, and is used to define a discrepancy measure, known as the *Stein discrepancy* (Gorham & Mackey, 2015). The latter considers the "maximum violation of Stein's identity" given a proper set of functions $\mathcal{F} \subseteq \mathcal{S}(\pi)$:

$$\mathbb{S}(\mu, \pi) = \max_{\phi\in\mathcal{F}} \left\{\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi\phi(x)]\right\}. \tag{5}$$

Note that $\mathbb{S}(\mu, \pi)$ is not symmetric. The set $\mathcal{S}(\pi)$ might be different to $\mathcal{S}(\mu)$, and even if they are equal, inverting the densities in the expectation leads to a different result. The choice of $\mathcal{F}$ is crucial as it determines the discriminative power and tractability of the Stein discrepancy. It also has to be included in $\mathcal{S}(\pi)$. Traditionally, $\mathcal{F}$ is chosen to be the set of all functions with bounded Lipschitz norms, but this choice casts a challenging functional optimization problem. To overcome this difficulty, (Liu et al., 2016) chooses $\mathcal{F}$ to be a vector-valued RKHS, which allows to find closed-form solution to Equation (5). The Stein discrepancy restricted to that RKHS is known as *Kernelized Stein Discrepancy*.

3.2.3. Kernelized Stein Discrepancy

From now on, we consider $\mu, \pi \in \mathcal{P}_2(\Omega)$ such that π is the target measure, $\mu \ll \pi \ll \lambda$, and $\mathrm{supp}(\mu) = \Omega$. This last assumption allows us to write:

$$\forall B \subseteq \Omega, \lambda(B) = 0 \iff \mu(B) = 0.$$

Next, we define the vector-valued RKHS that will be used in the Kernelized Stein Discrepancy.

Definition 3.4 (Product RKHS (Liu & Wang, 2016))**.** Let $k : \Omega \times \Omega \to \mathbb{R}$ be a continuous, symmetric, and positive-definite kernel such that $\forall x \in \Omega, k(\cdot, x) \in \mathcal{S}(\mu) \cap \mathcal{S}(\pi)$. Using the Moore–Aronszajn theorem (Aronszajn, 1950), we consider the associated real-valued RKHS $\mathcal{H}_0$. Let $\mathcal{H}$ be the product RKHS induced by $\mathcal{H}_0$, i.e. $\forall f = (f_1, \ldots, f_d)^\top$, $f \in \mathcal{H} \iff \forall 1 \leq i \leq d, f_i \in \mathcal{H}_0$. The inner product of $\mathcal{H}$ is defined by

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

Let $L^2_\mu(\Omega)$ be the set of functions from Ω to $\mathbb{R}$ that are square-integrable w.r.t. μ. Let $L^2_\mu(\Omega, \Omega)$ be the set of functions from Ω to Ω that are component-wise in $L^2_\mu(\Omega)$, i.e.

$$\forall f \in L^2_\mu(\Omega, \Omega),\ \forall 1 \leq i \leq d, f_i \in L^2_\mu(\Omega).$$

We proved that, assuming k is square-integrable w.r.t. μ, that the integral operator

$$\begin{aligned} T_k :\ & L^2_\mu(\Omega) \to L^2_\mu(\Omega) \\ & T_k f \mapsto \int_\Omega k(\cdot, x) f(x)\mathrm{d}\mu(x) \end{aligned}$$

is a mapping from $L^2_\mu(\Omega)$ to $\mathcal{H}_0$, i.e. $T_k : L^2_\mu(\Omega) \to \mathcal{H}_0$. (See proof in Appendix A.3). This allows to define another integral operator

$$\begin{aligned} S_\mu :\ & L^2_\mu(\Omega, \Omega) \to \mathcal{H} \\ & f \mapsto T_k f, \end{aligned}$$

where T_k is applied component-wise. The proof in Appendix A.3 also shows that $\mathcal{H}$ is a subset of $L^2_\mu(\Omega, \Omega)$. Thus, we can define the inclusion map

$$\iota : \mathcal{H} \hookrightarrow L^2_\mu(\Omega, \Omega),$$

whose adjoint is $\iota^\star = S_\mu$. Then, have the following equality:

$$\begin{aligned} &\forall f \in L^2_\mu(\Omega, \Omega), \forall g \in \mathcal{H}, \\ &\langle f, \iota g\rangle_{L^2_\mu(\Omega,\Omega)} = \langle \iota^\star f, g\rangle_{\mathcal{H}} = \langle S_\mu f, g\rangle_{\mathcal{H}}. \end{aligned}$$

We can now define the KSD.

Definition 3.5 (Kernelized Stein Discrepancy (Liu et al., 2016))**.** Let $\mathcal{H}$ be a product RKHS as defined in Definition 3.4. The *Kernelized Stein Discrepancy* (KSD) is then defined as:

$$\mathfrak{K}(\mu|\pi) = \max_{f\in\mathcal{H}} \left\{\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)] \mid \|f\|_{\mathcal{H}} \leq 1\right\}.$$

The construction of $\mathcal{H}$ was motivated by the fact that the closed-form solution of the KSD is given by the following theorem.

Theorem 3.6 (Steepest trajectory (Liu et al., 2016))**.** *The function that maximizes the KSD is given by:*

$$\frac{\phi^\star_\mu}{\|\phi^\star_\mu\|_{\mathcal{H}}} = \arg\max_{f\in\mathcal{H}} \left\{\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)] \mid \|f\|_{\mathcal{H}} \leq 1\right\}.$$

where $\phi_\mu^\star = \mathbb{E}_{x\sim\mu}[\nabla \log \pi(x)k(\cdot,x) + \nabla_x k(\cdot,x)]$. *It is the steepest trajectory in* $\mathcal{H}$ *that maximizes* $\mathfrak{K}(\mu|\pi)$. *The KSD is then given by*

$$\mathfrak{K}(\mu|\pi) = \mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \phi_\mu^\star(x)].$$

The proof strategy is to remark that, for any function $f \in \mathcal{H}$, $\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)] = \langle f, \phi_\mu^\star\rangle$. Then, the result follows from the Cauchy-Schwarz inequality. (See proof in Appendix A.4). From here, we make the mild assumption that

$$\phi_\mu^\star \in \bigcap_{\mu\in\mathcal{P}_2(\Omega)} \mathcal{S}(\mu).$$

Choosing $k(\cdot,x)$ such that

$$\forall y \in \Omega, \lim_{d(\{x\},\partial\Omega)\to 0} k(x,y) = 0, \text{where}$$
$$d(A,B) = \inf\{\|a-b\|_2 \mid a\in A \wedge b \in B\}$$

is enough to ensure this assumption. This would be the case for a modified Gaussian kernel:

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2 f(x,y)}\right),$$

where f is a positive function that tends to 0 (sufficiently fast for k to be in $L^2_\mu(\Omega,\Omega)$) when the distance between $\partial\Omega$ and either x or y tends to 0. This leads to the following result of the SVGD theory.

Theorem 3.7 (KL steepest descent trajectory (Liu & Wang, 2016))**.** *Let* $\mathcal{H}$ *be a product RKHS (Definition 3.4). Let* $\phi_\mu^\star \in \mathcal{H}$ *be as defined in Theorem 3.6. Let* $\varepsilon > 0$ *and*

$$T_\varepsilon : (\Omega\to\Omega) \to \Omega$$
$$\phi \mapsto I_d + \varepsilon\phi.$$

Then,

$$\arg\min_{\phi\in\mathcal{H}} \{\nabla_\varepsilon KL(T_\varepsilon(\phi)_\#\mu||\pi)|_{\varepsilon=0} \mid \|\phi\|_\mathcal{H} \le 1\} = \frac{\phi_\mu^\star}{\|\phi_\mu^\star\|_\mathcal{H}},$$

and $\nabla_\varepsilon KL((I_d+\varepsilon\phi_\mu^\star)_\#\mu||\pi)|_{\varepsilon=0} = -\mathfrak{K}(\mu|\pi)$.

(See proof in Appendix A.5). This last result is the key of the SVGD algorithm. It means that $\phi_\mu^\star$ is the optimal direction (within $\mathcal{H}$) to update μ in order to minimize the KL-divergence between μ and π. Indeed, the slope of the function $g : \varepsilon \mapsto KL(T_\varepsilon(\phi_\mu^\star/\|\phi_\mu^\star\|_\mathcal{H})_\#\mu||\pi)$ is minimal at 0. As $0 \in \mathcal{H}$ (that nullifies the gradient), the result ensures that g is decreasing over $[0,\delta]$, for $\delta > 0$ small enough. Consequently, SVGD iteratively updates μ in the direction induced by $\phi_\mu^\star$, with a small step size ε:

$$\mu_{n+1} = (I_d + \varepsilon\phi_{\mu_n}^\star)_\#\mu_n. \tag{6}$$

Furthermore, given the above assumption on $\phi_\mu^\star$, we have the following lemma.

Lemma 3.8. *Let* $\mathcal{H}$ *be a product RKHS as defined in Definition 3.4. Then,* $\phi_\mu^\star \in \mathcal{H}$ *as defined in Theorem 3.6. Given the above assumption on* $\phi_\mu^\star$, *we have that*

$$\mathfrak{K}(\mu|\pi) = \left\|\phi_\mu^\star\right\|_\mathcal{H}^2.$$

Proof. We showed in Appendix A.4 that

$$\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)] = \langle f, \phi_\mu^\star\rangle_\mathcal{H}$$

for any $f \in \mathcal{H}$. Thus, $\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi\phi_\mu^\star(x)] = \langle\phi_\mu^\star,\phi_\mu^\star\rangle_\mathcal{H}$. ■

In particular, this lemma states that the derivative of the KL-divergence when considering the direction $\phi_\mu^\star$ is negative, meaning that the sequence $(KL(\mu_n||\pi))_{n\in\mathbb{N}}$ is decreasing.

4. SBS variants

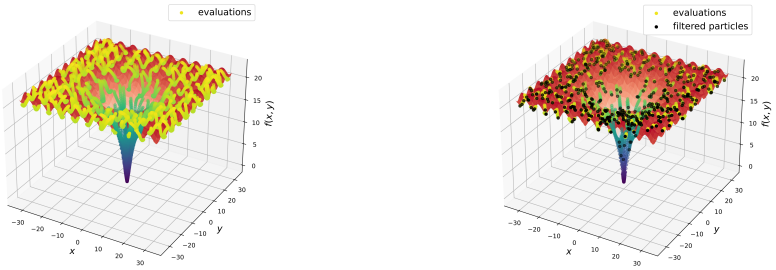
In addition to the main SBS method, we introduce two variants that can be more efficient in practice. The first one uses a particle filtering approach that removes the less promising particles (without replacing them). The second one is a hybrid method that uses SBS as a continuation for other global optimization methods, or –seen the other way around– those methods are used to initialize SBS. The particle filtering variant uses less budget than the main SBS. The hybrid variant uses some of the budget to run one of the pre-existing methods and then to initialize SBS with better starting points; the aim is to approximate the global minimum better than SBS with the same budget.

Particle filtering SBS (SBS-PF). We use a simple idea: to remove particles (i.e. candidate minimizers of f) that are less promising or stuck in bad local minima. We chose to remove a particle that does not move and have a significantly higher function value than the others. Therefore, this strategy is very likely to remove particles that are stuck in bad local minima. The difference between SBS and this variant is visualized in Figure 2. One can see that, in SBS-PF, the unpromising candidates are rapidly removed and are not replaced so that the remaining particles are more likely to converge to the global minimum. This strategy results

Algorithm 2 Initialization choice of SBS-HYBRID

Input: number of candidates n; CMA-ES budget b
Output: n candidates

Run CMA-ES for b function evaluations
Run WOA with n candidates
if CMA-ES found a better value than WOA **then**
  Sample n candidates from the last CMA-ES Gaussian
else
  Use the n candidates from WOA
end if
return the n candidates



(a) SBS **(b)** Particle filtering-based SBS-PF

Figure 2. Illustration of the plain SBS (left) and its particle filtering variant (right) on the 2d Ackley function (see Table 1). The color gradient represents the value of the function, from blue (low, preferred) to red (high). For SBS, particles are initialized uniformly over the domain. Then, they are updated in the direction induced by $\phi^{\star}_{\hat{\mu}_n}$ with a small step size. The trajectories of the particles draw the discretized flow of measures Φ_t. On the particle filtering SBS-PF variant, the particles are initialized and updated in the same way, but those being unpromising are rapidly removed and are not replaced; this is visible as there are no persisting trajectories in the area where the function has high value. This results in a significant reduction of the budget while having similar performance.

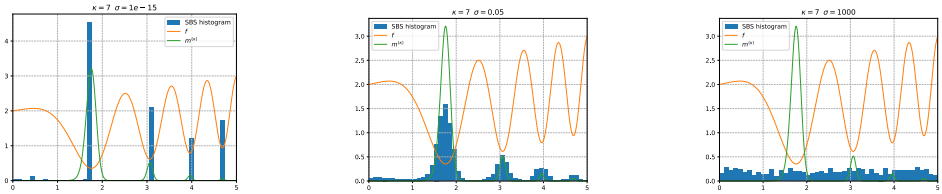


Figure 3. Illustration of the exploration/exploitation trade-off in SBS with different values of σ. In blue, the repartition of the particles, in orange, $x \mapsto \cos(x^2) + x/5 + 1$, in green $m^{(\kappa)}$.

in a significant reduction of the budget used, while having similar results as SBS.

SBS-HYBRID. Another interesting direction is to use SBS as a continuation for particles- or distribution-based methods, such as WOA or CMA-ES. Indeed, the design of SBS allows to initialize the particles with the result of such a method and then continue the optimization process. We introduce SBS-HYBRID that runs few iterations of CMA-ES and WOA to choose the most promising algorithm among them and continue the optimization with SBS (see Algorithm 2). Both WOA and CMA-ES are efficient methods, thus, running them for a small number of iterations allows to find a good starting point for SBS. Moreover, both methods are not well-fitted for a high budget for different reasons: CMA-ES uses early stopping rules (i.e. for the covariance matrix to not become ill-conditioned), and WOA takes more time to run than SBS for the same budget. SBS-HYBRID can be seen as a combination of SBS, an asymptotic consistent method, on top of very efficient non-consistent methods. The strength of SBS-HYBRIDis that it provides very good results while it is still asymptotically consistent, since the asymptotic distribution of the particles induced by WOA and CMA-ES meet the assumptions of Theorem 2.6.

5. Benchmark

In this section, we compare numerically SBS and its variants with state-of-the-art global optimization methods. We consider the following methods: CMA-ES (Hansen & Ostermeier, 1996; 2001; Hansen et al., 2003), WOA (Mirjalili & Lewis, 2016), and ADALIPO (Malherbe & Vayatis, 2017). We use classical two dimensional benchmark functions for global optimization. Some are noisy, multimodal or very smooth. We provide the implementation [1] of this experiment. For the results of Table 1, we ran each method 100 times on each function. The budget is set in order for the methods to stop in a reasonable time. For SBS and SBS-PF, we set $\sigma \triangleq 10^{-5}$. For the SBS-HYBRID, $\sigma \triangleq 10^{-10}$. As one can see, SBS is competitive with the state-of-the-art methods and score the second rank on average. SBS-PF achieves similar results on average with significantly less evaluations. Moreover, SBS-HYBRID outperforms all the other methods on average. It is a very performing method that combines the efficiency of both CMA-ES and WOA with the large budget compatibility of SBS. Even if SBS and SBS-PF are competitive, they do not clearly outperform the state-of-the-art methods. More clever particle filtering and adaptive locality of the kernel have a potential to improve further SBS (see Section 6). Note that, in order to update the particles, SBS needs to compute the gradient of the func-

[1] github.com/gaetanserre/Stochastic-Global-Optimization

Table 1. **Comparative results.** Comparison between all SBS variants with several state-of-the-art methods on two dimensional benchmark functions. For each function, we report the average best function value found (lower is better). The budget for ADALIPO is set to 2K, 200K for WOA and to 800K for the others. The average budget used by CMA-ES is 547 and 90K for SBS-PF. SBS-HYBRID runs 1K iterations of CMA-ES and WOA. As one can see, SBS and its variants are competitive with the state-of-the-art methods. The standard SBS ranks second on average. The hybrid method SBS-HYBRID outperforms all the other methods on average.

	STATE-OF-THE-ART			PROPOSED METHODS		
FUNCTIONS	ADALIPO	CMA-ES	WOA	SBS-PF	SBS	SBS-HYBRID
ACKLEY	1.53	19.29	$5.40 \cdot 10^{-7}$	0.028	0.015	$8 \cdot 10^{-3}$
BRANIN	0.4	0.39788	0.39789	0.39788	0.39788	0.39788
DROP WAVE	-0.94	-0.83	-1	-0.96	-0.96	-0.95
EGG HOLDER	-930	-395	-959	-941	-951	-946
GOLDSTEIN PRICE	3.56	6.08	3	3	3	3
HIMMELBLAU	0.007	$4.2 \cdot 10^{-16}$	$1.1 \cdot 10^{-5}$	$2.2 \cdot 10^{-7}$	$5.7 \cdot 10^{-8}$	$4.5 \cdot 10^{-15}$
HOLDER TABLE	-19.19	-7.7	-19.20848	-19.20846	-19.20845	-19.2085
MICHALEWICZ	-1.784	-1.5	-1.8012	-1.8012	-1.8013	-1.789
RASTRIGIN	0.13	4.49	$3 \cdot 10^{-13}$	0.02	0.01	0.4
ROSENBROCK	0.03	$1.05 \cdot 10^{-15}$	$1.5 \cdot 10^{-6}$	$4 \cdot 10^{-3}$	$1.7 \cdot 10^{-6}$	$1.1 \cdot 10^{-12}$
SPHERE	0.001	$7.3 \cdot 10^{-16}$	$4.5 \cdot 10^{-15}$	$8.3 \cdot 10^{-8}$	$8.6 \cdot 10^{-9}$	$2.2 \cdot 10^{-16}$
AVERAGE RANK	5.273	4.364	2.636	3.727	2.636	**2.364**
FINAL RANK	5	4	2	3	2	**1**

tion. In our implementation, we estimate it using finite differences. However, it takes the majority of the budget. More sophisticated methods, such as automatic differentiation, significantly reduce the number of evaluations, which would make SBS even more competitive.

6. Discussion

Link with Simulated Annealing. The link between SBS and Simulated Annealing (Kirkpatrick et al., 1983) is not difficult to see. Indeed, both algorithms are asymptotic methods that sample from the BD. However, the way they sample from that distribution is different. Simulated Annealing is a Markov Chain Monte-Carlo method, while SBS is a deterministic variational approach. The minimum temperature parameter of Simulated Annealing is the inverse of the κ parameter of SBS. Thus, any scheduler for the temperature used in Simulated Annealing can also be used in SBS. However, there is an extra degree of exploration/exploitation in SBS, corresponding to the kernel size used by the employed SVGD sampling.

Locality of the kernel. In classical SVGD implementations, the used RBF kernel is: $k(x, x') = \exp\left(-\frac{\|x-x'\|_2^2}{2\sigma^2}\right)$, as it is in the Stein class of any smooth density supported on $\mathbb{R}^d$. σ controls the locality of the attraction and repulsion forces applied on the particles, respectively expressed as:

$$\text{attr}(x) = \mathbb{E}_{x' \sim \hat{\mu}_n} \left[\nabla \log \pi(x') k(x, x')\right],$$
$$\text{rep}(x) = \mathbb{E}_{x' \sim \hat{\mu}_n} \left[\nabla_{x'} k(x, x')\right].$$

The first term attracts lonely particles to a close cluster of particles, and the second term repels particles that are too close to each other. They are respectively exploitation and exploration forces. Indeed, the attraction allows particles to "fall" in local minima, where a lot of particles are already stuck in. The repulsion prevents particles from getting stuck together at a narrow region of the search space, and forces them to explore the space. The value of σ controls the range of these forces. A small σ value leads to a weak repulsion and thus more exploitation. An arbitrary small σ leads to a uniform distribution over the local minima. In the contrary, a large σ leads to more exploration, as the particles will repel themselves from even from a very far distance. An arbitrary large σ leads to a uniform discretization of the space. In the case of SBS, the value of σ is not fixed and can be chosen by the user. These behaviors are illustrated in Figure 3.

7. Conclusion

In this paper, we introduced *Stein Boltzmann Sampling* (SBS), a new method for global optimization of Lipschitz functions. It is based on the *Stein Variational Gradient Descent* algorithm, which is a deterministic variational approach. We proved that SBS is consistent and showed that it is competitive with state-of-the-art methods on classical benchmark functions. We also introduced a variant of SBS that uses particle filtering to save budget while having better performances than the original version. Moreover, we introduced SBS-HYBRID, a hybrid method that combines the efficiency of CMA-ES and WOA with the large budget compatibility of SBS, outperforming all the other methods on the benchmark functions. This shows that SBS can be used as a continuation for particles or distributions based methods, particularly method that are not fitted for a large budget. For future work, we plan to study further the convergence rate of SBS and its components to make it more appealing for global optimization in real-world applications.

Acknowledgment

The authors acknowledge the support from the Industrial Data Analytics and Machine Learning Chair hosted at ENS Paris-Saclay.

References

Aronszajn, N. *Theory of Reproducing Kernels*. Transactions of the American Mathematical Society, 1950.

Billingsley, P. *Convergence of Probability Measures*. Wiley, 1999.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Proceedings of Advances in Neural Information Processing Systems*, 2022.

de Moura, L. and Ullrich, S. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*. Springer International Publishing, 2021.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. *Proceedings of Advances in Neural Information Processing Systems*, 2015.

Hansen, N. and Ostermeier, A. Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, 1996.

Hansen, N. and Ostermeier, A. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 2001.

Hansen, N., Müller, S. D., and Koumoutsakos, P. Reducing the time complexity of the derandomized evolution strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, 2003.

Hwang, C.-R. Laplace's Method Revisited: Weak Convergence of Probability Measures. *The Annals of Probability*, pp. 1177–1182, 1980.

Jordan, M. I., Kornowski, G., Lin, T., Shamir, O., and Zampetakis, M. Deterministic Nonsmooth Nonconvex Optimization, 2023.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science*, 1983.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. Kernel stein discrepancy descent. In *Proceedings of the International Conference on Machine Learning*, 2021.

Lee, J., Lee, I.-H., Joung, I., Lee, J., and Brooks, B. R. Finding multiple reaction pathways via global optimization of action. *Nature Communications*, 2017.

Liu, Q. Stein variational gradient descent as gradient flow. *Proceedings of Advances in Neural Information Processing Systems*, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Proceedings of Advances in Neural Information Processing Systems*, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, 2016.

Lu, J., Lu, Y., and Nolen, J. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 2019.

Luo, X. Minima distribution for global optimization, 2019.

Malherbe, C. and Vayatis, N. Global optimization of Lipschitz functions. In *Proceedings of the International Conference on Machine Learning*, 2017.

mathlib Community, T. The Lean mathematical library. In *Proceedings of the ACM SIGPLAN International Conference on Certified Programs and Proofs*, 2020.

Mirjalili, S. and Lewis, A. The whale optimization algorithm. *Advances in engineering software*, 2016.

Pintér, J. D. Global optimization in action. *Scientific American*, 1991.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 1972.

Villani, C. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

Villani, C. *Optimal Transport*. Springer Berlin Heidelberg, 2009.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the International Conference on Machine Learning*, 2020.

Zhou, D.-X. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 2008.

A. Proofs

In the following sections, we provide the proofs of the theorems and lemmas stated in the main text. We also provide Lean proofs of some results. The Lean proofs are available here [2]. We use the same notation that in the main text. Recall that $\mathfrak{K}(\mu|\pi)$ denotes the *Kernelized Stein Discrepancy* and $\phi_\mu^\star$ is the steepest trajectory in $\mathcal{H}$ that minimizes $\mathfrak{K}(T_{\#}\mu|\pi)$. We also introduce a new quantifier $\overline{\forall}_\mu$, such that, given a predicate P and a measure μ,

$$\left[\overline{\forall}_\mu x \in E \subseteq \Omega, P(x)\right] \triangleq \left[\exists A \subseteq E, \mu(A) = \mu(E), \forall x \in A, P(x)\right].$$

This quantifier means that the predicate P is true for almost all $x \in E$ w.r.t. the measure μ. When the considered measure is the standard Lebesgue measure, we simply write $\overline{\forall}$.

A.1. Proof of Properties 3.1

The continuous BD is a special case of the nascent minima distribution, introduced in (Luo, 2019), that has the generic form

$$m_{f,\Omega}^{(\kappa)}(x) = m^{(\kappa)}(x) = \frac{\tau^\kappa(f(x))}{\int_\Omega \tau^\kappa(f(t))\mathrm{d}t}, \tag{7}$$

where $\tau : \mathbb{R} \to \mathbb{R}_{>0}$ is monotonically decreasing. We have the following theorems for general τ.

Theorem A.1 (Nascent minima distribution properties)**.** *Let $m^{(\kappa)}$ and τ be defined in Equation (7). Then, we have the following properties:*

- *If $\lambda(X^*) = 0$, then, $\forall x \in \Omega$,*

$$\lim_{\kappa\to\infty} m^{(\kappa)}(x) = \begin{cases} \infty & \text{if } x \in X^* \\ 0 & \text{otherwise} \end{cases}.$$

- *If $0 < \lambda(X^*)$, then, $\forall x \in \Omega$,*

$$\lim_{\kappa\to\infty} m^{(\kappa)}(x) = \begin{cases} \frac{1}{\lambda(X^*)} & \text{if } x \in X^* \\ 0 & \text{otherwise} \end{cases}.$$

Proof. Let's prove the two properties together. Let $p = \tau(f(x')) > 0$, $\forall x' \notin X^*$. Then, $\exists \Omega_p$, such that $0 < \lambda(\Omega_p)$, $p < \tau(f(t))$, i.e. $f(t) < f(x')$. Thus,

$$\begin{aligned} m^{(\kappa)}(x') &= \frac{p^\kappa}{\int_{\Omega_p} \tau^\kappa(f(t))\mathrm{d}t + \int_{\Omega/\Omega_p} \tau^\kappa(f(t))\mathrm{d}t} \\ &\leq \frac{p^\kappa}{\int_{\Omega_p} \tau^\kappa(f(t))\mathrm{d}t} \\ &= \frac{1}{\int_{\Omega_p} p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t}. \end{aligned}$$

For any t in Ω_p, $p^{-1}\tau(f(t)) > 1$. Therefore $\lim_{\kappa\to,\infty} \int_{\Omega_p} p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t = \infty$. Hence,

$$\forall x' \notin X^*, \lim_{\kappa\to\infty} m^{(\kappa)}(x) = 0.$$

[2] gaetanserre.fr/assets/Lean/SBS/index.html

Now, let's consider any $x'' \in X^*$ and $p = \tau(f(x''))$. We have

$$\begin{aligned} m^{(\kappa)}(x'') &= \frac{p^\kappa}{\int_\Omega \tau^\kappa(f(t))\mathrm{d}t} \\ &= \frac{1}{\int_X^* p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t + \int_{\Omega/X^*} p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t} \\ &= \frac{1}{\int_{X^*} \mathrm{d}t + \int_{\Omega/X^*} p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t} \ (\forall t \in X^*, \tau(f(t)) = p) \\ &= \frac{1}{\lambda(X^*) + \int_{\Omega/X^*} p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t}. \end{aligned}$$

For any t in Ω/X^*, $p^{-1}\tau(f(t)) < 1$. Therefore, $\lim_{\kappa\to\infty} \int_{\Omega/X^*} p^{-\kappa}\tau^\kappa(f(t))\mathrm{d}t = 0$. Thus,

$$\forall x'' \in X^*, \lim_{\kappa\to\infty} m^{(\kappa)}(x'') = \begin{cases} \infty & \text{if } \lambda(X^*) = 0 \\ \frac{1}{\lambda(X^*)} & \text{otherwise} \end{cases}.$$

■

Theorem A.2 (Convergence of expectation)**.** *$\forall f \in C^0(\Omega, \mathbb{R})$, the following holds*

$$\lim_{\kappa\to\infty} \int_\Omega f(x)\ m^{(\kappa)}(x)\mathrm{d}x = f^*.$$

Moreover, if $X^* = x^*$, we have

$$\lim_{\kappa\to\infty} \int_\Omega x\ m^{(\kappa)}(x)\mathrm{d}x = x^*.$$

Proof. If f is constant, it is straightforward as $m^{(\kappa)}$ is a PDF. Suppose f not constant on Ω. For any $\varepsilon > 0$, let $0 < \delta \triangleq \frac{\varepsilon}{1+(\max_{x\in\Omega} f(x) - f^*)} \leq \varepsilon$. As f is continuous, $\exists \Omega_\delta = \{x \in \Omega \mid f(x) - f^* < \delta\}$, the corresponding level set. Using Theorem A.1, $\exists K \in \mathbb{N}$ such that

$$\int_{\Omega/\Omega_\delta} m^{(\kappa)}(x)\mathrm{d}x < \delta$$

holds $\forall \kappa > K$, as $m^{(\kappa)}$ tends to 0 $\forall x \notin X^*$. Thus,

$$\begin{aligned} 0 &< \int_\Omega f(x)m^{(\kappa)}(x)\mathrm{d}x - f^* \\ &= \int_\Omega f(x)m^{(\kappa)}(x)\mathrm{d}x - f^* \int_\Omega m^{(\kappa)}(x)\mathrm{d}x \\ &= \int_\Omega (f(x) - f^*)m^{(\kappa)}(x)\mathrm{d}x \\ &= \int_{\Omega_\delta} (f(x) - f^*)m^{(\kappa)}(x)\mathrm{d}x \\ &\qquad + \int_{\Omega/\Omega_\delta} (f(x) - f^*)m^{(\kappa)}(x)\mathrm{d}x \\ &< \delta \int_{\Omega_\delta} m^{(\kappa)}(x)\mathrm{d}x \\ &\qquad + (\max_{x\in\Omega} f(x) - f^*) \int_{\Omega/\Omega_\delta} m^{(\kappa)}(x)\mathrm{d}x \\ &< \delta(1-\delta) + (\max_{x\in\Omega} f(x) - f^*)\delta \\ &< (1 + (\max_{x\in\Omega} f(x) - f^*))\delta = \varepsilon. \end{aligned}$$

The proof is similar for the second statement, by setting

$$\Omega_\delta = \{x \in \Omega | \|x - x^*\| < \delta\}.$$

■

Letting $\tau = x \mapsto e^{-x}$ gives Properties 3.1.

A.2. Proof of Lemma 3.3.

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim\mu}[\mathcal{A}_\mu \phi(x)] &= \int_\Omega \nabla \log \mu(x)^\top \phi(x) + \nabla \cdot \phi(x) \mathrm{d}\mu(x) \\
&= \int_\Omega \mu(x)(\nabla \log \mu(x)^\top \phi(x))\mathrm{d}x + \int_\Omega \mu(x)(\nabla \cdot \phi(x))\mathrm{d}x \\
&= \int_\Omega \mu(x)(\nabla \log \mu(x)^\top \phi(x))\mathrm{d}x - \int_\Omega \nabla\mu(x)^\top \phi(x)\mathrm{d}x \\
&= \int_\Omega \nabla\mu(x)^\top \phi(x)\mathrm{d}x - \int_\Omega \nabla\mu(x)^\top \phi(x)\mathrm{d}x.
\end{aligned}$$

■

A.3. Proof of T_k is a map to $\mathcal{H}_0$

Proof. As k is continuous, symmetric, and positive-definite and as $\mu(\Omega) < \infty$ and as T_k is a self-adjoint operator, we can apply the Mercer's theorem to obtain a sequence of eigenfunctions $(\phi_i)_{i\in\mathbb{N}}$ and a sequence of eigenvalues $(\lambda_i)_{i\in\mathbb{N}}$ such that $(\phi_i)_{i\in I}$ is an orthornormal basis of $L^2_\mu(\Omega)$, such that $(\lambda_i)_{i\in\mathbb{N}}$ is nonnegative and converges to 0, and such that the following holds:

$$\forall s, t \in \Omega, k(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t).$$

The above series converges absolutely and uniformly on $\Omega \times \Omega$. Let define the set

$$\mathcal{H}_k = \left\{ f \in L^2_\mu(\Omega) \middle| f = \sum_{i=1}^{\infty} \lambda_i a_i \phi_i \wedge \sum_{i=1}^{\infty} \lambda_i a_i^2 < \infty \right\},$$

endowed with the inner product

$$\forall f, g \in \mathcal{H}_k, \langle f, g\rangle_{\mathcal{H}_k} = \left\langle \sum_{i=1}^{\infty} \lambda_i a_i \phi_i, \sum_{i=1}^{\infty} \lambda_i b_i \phi_i \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \lambda_i a_i b_i. \tag{8}$$

Routine works show that Equation (8) defines a inner product and that $\mathcal{H}_k$ is a Hilbert space. Let's show that $\mathcal{H}_k$ is a RKHS with kernel k, i.e., $\forall t \in \Omega$, $k(t,\cdot) \in \mathcal{H}_k$ and, $\forall f \in \mathcal{H}_k$, $f(t) = \langle f, k(t,\cdot)\rangle_{\mathcal{H}_k}$. Let $t \in \Omega$. First, Ω is compact, $\mu(\Omega) = 1 < \infty$, and $k(t,\cdot)$ is continuous on Ω, thus $k(t,\cdot) \in L^2_\mu(\Omega)$. Then, we have that

$$k(t,\cdot) = \sum_{i=1}^{\infty} \lambda_i \phi_i(t)\phi_i,$$

and

$$\sum_{i=1}^{\infty} \lambda_i \phi_i^2(t) = k(t,t) < \infty.$$

Thus, $k(t,\cdot) \in \mathcal{H}_k$. Let $f \in \mathcal{H}_k$. One can write

$$\begin{aligned}\langle f, k(t,\cdot)\rangle_{\mathcal{H}_k} &= \left\langle \sum_{i=1}^{\infty} \lambda_i a_i \phi_i, \sum_{i=1}^{\infty} \lambda_i \phi_i(t)\phi_i \right\rangle_{\mathcal{H}_k} \\ &= \sum_{i=1}^{\infty} \lambda_i a_i \phi_i(t) \\ &= f(t).\end{aligned}$$

Therefore, $\mathcal{H}_k$ is indeed a RKHS with kernel k.The Moore–Aronszajn theorem ensures that, given k, there exists an unique RKHS such that k is its kernel. Thus, $\mathcal{H}_k = \mathcal{H}_0$. That's prove that $\mathcal{H}_0 \subseteq L^2_\mu(\Omega) \implies \mathcal{H} \subseteq L^2_\mu(\Omega, \Omega)$. Let's now prove that $\forall f \in L^2_\mu(\Omega), T_k f \in \mathcal{H}_0$. Let $f \in L^2_\mu(\Omega)$. We begin by proving that $T_k f \in L^2_\mu(\Omega)$.

$$\begin{aligned}|T_k f(t)| &= \left| \int_\Omega k(t,s) f(s) \mathrm{d}\mu(s) \right| \\ &\leq \int_\Omega |k(t,s)||f(s)| \mathrm{d}\mu(s) \\ &= \langle |k(t,\cdot)|, |f| \rangle_{L^2_\mu(\Omega)} \\ &\leq \|k(t,\cdot)\|_{L^2_\mu(\Omega)} \, \|f\|_{L^2_\mu(\Omega)}.\end{aligned}$$

Then,

$$\begin{aligned}\|T_k f(t)\|^2_{L^2_\mu(\Omega)} &= \int_\Omega |T_k f(t)|^2 \mathrm{d}t \\ &\leq \int_\Omega \|k(t,\cdot)\|^2_{L^2_\mu(\Omega)} \mathrm{d}t \, \|f\|^2_{L^2_\mu(\Omega)} \\ &= \|k\|^2_{L^2_\mu} \, \|f\|^2_{L^2_\mu(\Omega)} \\ &< \infty.\end{aligned}$$

We now prove that $T_k f \in \mathcal{H}_0$.

$$\begin{aligned}T_k f &= \int_\Omega k(\cdot, s) f(s) \mathrm{d}\mu(s) \\ &= \int_\Omega \sum_{i=1}^{\infty} \lambda_i f(s) \phi_i(s) \phi_i(\cdot) \mathrm{d}\mu(s) \\ &= \sum_{i=1}^{\infty} \lambda_i \phi_i(\cdot) \int_\Omega f(s) \phi_i(s) \mathrm{d}\mu(s) \\ &= \sum_{i=1}^{\infty} \lambda_i \langle f, \phi_i \rangle_{L^2_\mu(\Omega)} \phi_i.\end{aligned}$$

As $(\phi_i)_{i\in\mathbb{N}}$ is an orthonormal basis of $L^2_\mu(\Omega)$ we have that

$$\int_\Omega \phi_i \phi_j \mathrm{d}\mu = \mathbb{1}_{\{i=j\}},$$

which implies, using Parseval's identity,

$$\sum_{i=1}^{\infty} \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} = \|f\|^2_{L^2_\mu(\Omega)} < \infty.$$

As $(\lambda_i)_{i\in\mathbb{N}}$ converges to 0, $\exists I \in \mathbb{N}$ such that $\forall i > I$, $\lambda_i < 1$. Thus,

$$\begin{aligned}
\sum_{i=1}^{\infty} \lambda_i \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} &= \sum_{i=1}^{I} \lambda_i \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} + \sum_{i=I+1}^{\infty} \lambda_i \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} \\
&\leq \sum_{i=1}^{I} \lambda_i \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} + \sum_{i=I+1}^{\infty} \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} \\
&\leq \sum_{i=1}^{I} \lambda_i \langle f, \phi_i \rangle^2_{L^2_\mu(\Omega)} + \|f\|^2_{L^2_\mu(\Omega)} \\
&< \infty.
\end{aligned}$$

Therefore, $\forall f \in L^2_\mu(\Omega)$, $T_k f \in \mathcal{H}_0$, which proves that $T_k : L^2_\mu(\Omega) \hookrightarrow \mathcal{H}_0$. ■

A.4. Proof of Theorem 3.6

Proof. First, we show that $\phi^\star_\mu \in \mathcal{H}$, i.e. $\forall 1 \leq i \leq d$, $(\phi^\star_\mu)^{(i)} \in \mathcal{H}_0$. Let define the function

$$\begin{aligned}
f^{(i)} : \Omega &\to \mathbb{R}, \\
x &\mapsto \frac{\partial \log \frac{\pi}{\mu}(x)}{\partial x_i}.
\end{aligned}$$

Then, as $\forall x \in \Omega$, $k(\cdot, x) \in \mathcal{S}(\mu)$, it is easy to show that

$$(\phi^\star_\mu)^{(i)} = T_k f^{(i)} \in \mathcal{H}_0.$$

Thus, $\phi^\star_\mu = S_\mu \nabla \log \frac{\pi}{\mu} \in \mathcal{H}$. Next, we prove that

$$\forall f \in \mathcal{H}, \mathbb{E}_{x\sim\mu}\left[\mathcal{A}_\pi f(x)\right] = \langle f, \phi^\star_\mu \rangle_{\mathcal{H}}.$$

$$\begin{aligned}
\langle f, \phi^\star_\mu \rangle_{\mathcal{H}} &= \sum_{\ell=1}^{d} \langle f^{(\ell)}, \mathbb{E}_{x\sim\mu}\left[\nabla \log \pi^{(\ell)}(x) k(x\cdot) + \nabla_x k^{(\ell)}(x, \cdot)\right] \rangle_{\mathcal{H}_0} \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{\ell=1}^{d} \langle f^{(\ell)}, \nabla \log \pi^{(\ell)}(x) k(\cdot, x) + \nabla_x k^{(\ell)}(x, \cdot) \rangle_{\mathcal{H}_0}\right] \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{\ell=1}^{d} \nabla \log \pi^{(\ell)}(x) \langle f^{(\ell)}, k(\cdot, x) \rangle_{\mathcal{H}_0} + \langle f^{(\ell)}, \nabla_x k^{(\ell)}(x, \cdot) \rangle_{\mathcal{H}_0}\right] \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{\ell=1}^{d} \nabla \log \pi^{(\ell)}(x) f^{(\ell)}(x) + \frac{\partial f^{(\ell)}(x)}{\partial x_\ell}\right] \text{ (Zhou, 2008)} \\
&= \mathbb{E}_{x\sim\mu}\left[\nabla \log \pi(x)^\top f(x) + \nabla \cdot f(x)\right].
\end{aligned}$$

Moreover, using the Cauchy-Schwarz inequality, we have that

$$\langle f, \phi^\star_\mu \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|\phi^\star_\mu\|_{\mathcal{H}}.$$

Thus,

$$\mathfrak{K}(\mu, \pi) \leq \|\phi^\star_\mu\|_{\mathcal{H}}.$$

Finally, by letting $f = \frac{\phi^\star_\mu}{\|\phi^\star_\mu\|_{\mathcal{H}}}$, we have that

$$\mathbb{E}_{x\sim\mu}\left[\mathcal{A}_\pi f\right] = \langle f, \phi^\star_\mu \rangle_{\mathcal{H}} = \|\phi^\star_\mu\|_{\mathcal{H}}.$$

For more details, see Lean proof [3]. ■

[3] gaetanserre.fr/assets/Lean/SBS/html/SteepestDirection.lean.html

A.5. Proof of Theorem 3.7

Proof. Note $T_\varepsilon = T$, $\mu_{[T]}$ the density of $T_\# \mu$ w.r.t. λ. First, when ε is sufficiently small, T is close to the identity and is guaranteed to be a one-to-one. Using change of variable, we know that $T_\#^{-1}\pi$ admits a density $\pi_{[T^{-1}]}$ w.r.t. λ and

$$\pi_{[T^{-1}]}(x) = \pi(T(x)) \cdot |\det \nabla_x T(x)|, \forall x \in \Omega.$$

Remark A.3. It is easy to see that, if T is a one-to-one map, then

$$\forall x \in \Omega, \left(\mu_{[T]} \circ T\right)(x) = \mu(x).$$

Let's show that $KL(T_\#\mu||\pi) = KL(\mu||T_\#^{-1}\pi)$.

$$\begin{aligned}
KL(T_\#\mu||\pi) &= \int_\Omega \log\left(\frac{\mu_{[T]}(x)}{\pi(x)}\right) \mathrm{d}T_\#\mu(x) \\
&= \int_{T^{-1}(\Omega)} \log\left(\frac{(\mu_{[T]} \circ T)(x)}{(\pi \circ T)(x)}\right) \mathrm{d}\mu(x) \\
&= \int_{T^{-1}(\Omega)} \log\left(\frac{(\mu_{[T]} \circ T)(x)}{(\pi_{[T^{-1}]} \circ T^{-1} \circ T)(x)}\right) \mathrm{d}\mu(x) \\
&= \int_{T^{-1}(\Omega)} \mu(x) \log\left(\frac{\mu(x)}{\pi_{[T^{-1}]}(x)}\right) \mathrm{d}x \\
&= \int_\Omega \mu(x) \log\left(\frac{\mu(x)}{\pi_{[T^{-1}]}(x)}\right) \mathrm{d}x \quad \left(T^{-1}(\Omega) = \left\{x \mid T^{-1}(x) \in \Omega\right\} = \Omega\right) \\
&= KL(\mu||T_\#^{-1}\pi).
\end{aligned}$$

For more details, see Lean proof [4]. Thus, we have

$$\begin{aligned}
\nabla_\varepsilon KL(\mu||T_\#^{-1}\pi) &= \nabla_\varepsilon \int_\Omega \mu(x) \log\left(\frac{\mu(x)}{\pi_{[T^{-1}]}(x)}\right) \mathrm{d}x \\
&= \int_\Omega \mu(x) \nabla_\varepsilon \left[\log(\mu(x)) - \log\left(\pi_{[T^{-1}]}(x)\right)\right] \mathrm{d}x \\
&= -\int_\Omega \mu(x) \nabla_\varepsilon \log\left(\pi_{[T^{-1}]}(x)\right) \mathrm{d}x \\
&= -\mathbb{E}_{x\sim\mu}\left[\nabla_\varepsilon \log\left(\pi_{[T^{-1}]}(x)\right)\right].
\end{aligned}$$

Now, let's compute $\nabla_\varepsilon \log\left(\pi_{[T^{-1}]}(x)\right)$.

$$\begin{aligned}
\nabla_\varepsilon \log\left(\pi_{[T^{-1}]}(x)\right) &= \nabla_\varepsilon \log\left(\pi(T(x)) \cdot |\det(\nabla_x T(x))|\right) \\
&= \nabla_\varepsilon \log \pi(T(x)) + \nabla_\varepsilon \log|\det(\nabla_x T(x))| \\
&= \nabla_{T(x)} \log \pi(T(x))^\top \nabla_\varepsilon T(x) + \nabla_\varepsilon \log|\det(\nabla_x T(x))| \\
&= \nabla_{T(x)} \log \pi(T(x))^\top \nabla_\varepsilon T(x) + \frac{1}{\det(\nabla_x T(x))} \nabla_\varepsilon \det(\nabla_x T(x)) \\
&= \nabla_{T(x)} \log \pi(T(x))^\top \nabla_\varepsilon T(x) + \frac{1}{\det(\nabla_x T(x))} \sum_{ij} \left(\nabla_\varepsilon \nabla_x T(x)_{ij} C_{ij}\right) \\
&= \nabla_{T(x)} \log \pi(T(x))^\top \nabla_\varepsilon T(x) + \sum_{ij} \left(\nabla_\varepsilon \nabla_x T(x)_{ij} \left(\nabla_x T(x)\right)^{-1}_{ji}\right) \\
&= \nabla_{T(x)} \log \pi(T(x))^\top \nabla_\varepsilon T(x) + \mathrm{trace}\left(\left(\nabla_x T(x)\right)^{-1} \cdot \nabla_\varepsilon \nabla_x T(x)\right),
\end{aligned}$$

where C is the cofactor matrix of $\nabla_x T(x)$. Finally, the result of the theorem is a special case of the above result. Indeed, $\forall \phi \in \mathcal{H}$, if $T = I_d + \varepsilon\phi$, then

[4] gaetanserre.fr/assets/Lean/SBS/html/KL.lean.html

- $T(x)|_{\varepsilon=0} = x$;
- $\nabla_\varepsilon T(x) = \phi(x)$;
- $\nabla_x T(x)|_{\varepsilon=0} = I_d$;
- $\nabla_\varepsilon \nabla_x T(x) = \nabla_x \phi(x)$.

This gives

$$\nabla_\varepsilon KL(T_\# \mu || \pi)|_{\varepsilon=0} = -\mathbb{E}_{x\sim\mu}\left[\nabla \log \pi(x)^\top \phi(x) + \nabla \cdot \phi(x)\right].$$

Applying Theorem 3.6 ends the proof. ■

A.6. Proof of Theorem 2.2

Proof. As $(T_t)_{0\leq t}$ is a locally Lipschitz family of diffeomorphisms representing the trajectories associated with the vector field ϕ_t, and as $\mu_t = T_{t\#}\mu$, then, a direct application of Theorem 5.34 from (Villani, 2003) gives that μ_t is the only solution of the linear transport equation

$$\begin{cases} \frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t \phi_t) & = 0, \forall t > 0, \\ \mu_0 & = \mu \end{cases},$$

where the divergence operator $(\nabla \cdot)$ is defined by duality against smooth compactly supported functions, i.e.

$$\forall \mu \in \mathcal{P}(\Omega), \forall \phi : \Omega \to \Omega, \forall \varphi \in C_c^\infty(\Omega), \langle T_{\nabla\cdot(\phi\mu)}, \varphi\rangle = -\langle T_\mu, \phi \cdot \nabla\varphi\rangle,$$

where, $\forall \mu \in \mathcal{P}(\Omega), T_\mu \in D'(\Omega)$ and $\forall \varphi \in C_c^\infty(\Omega), \langle T_\mu, \varphi\rangle = \int_\Omega \varphi \, \mathrm{d}\mu$ (see also (Villani, 2009)). Furthermore, as $\mu_{n+1} = (I_d + \varepsilon\phi^\star_{\mu_n})_\# \mu_n$ (see Equation (6)), one can write

$$\begin{aligned} \int_\Omega \varphi \mathrm{d}\mu_{n+1} &= \int_\Omega \varphi \circ (I_d + \varepsilon\phi^\star_{\mu_n}) \mathrm{d}\mu_n, \forall \varphi \in C_c^\infty(\Omega). \\ &\underset{\varepsilon\to 0}{\sim} \int_\Omega \varphi + \varepsilon(\nabla\varphi \cdot \phi^\star_{\mu_n}) \mathrm{d}\mu_n \left(\text{Taylor expansion of } \varphi(x) \text{ at } x + \varepsilon\phi^\star_{\mu_n}(x)\right) \\ &= \int_\Omega \varphi \, \mathrm{d}\mu_n + \int_\Omega \varepsilon \left(\nabla\varphi \cdot \phi^\star_{\mu_n}\right) \mathrm{d}\mu_n \\ &= \int_\Omega \varphi \, \mathrm{d}\mu_n - \int_\Omega \varepsilon\varphi \, \mathrm{d}\left(\nabla \cdot (\mu_n \phi^\star_{\mu_n})\right) \\ \iff \int_\Omega \varphi \, \mathrm{d}\mu_{n+1} - \int_\Omega \varphi \, \mathrm{d}\mu_n &= -\varepsilon \int_\Omega \varphi \, \mathrm{d}\left(\nabla \cdot (\mu_n \phi^\star_{\mu_n})\right). \end{aligned}$$

This shows that iteratively updates μ in the direction $I_d + \varepsilon\phi^\star_{\mu_n}$, given a small ε, corresponds to a finite difference approximation of the linear transport equation. ■

A.7. Proof of Theorem 2.3

Proof. Using the Leibniz integral rule, the time derivative of the KL-divergence writes

$$\begin{aligned}
\frac{\partial KL(\mu_t||\pi)}{\partial t} &= \frac{\partial}{\partial t}\int_\Omega \log\frac{\mathrm{d}\mu_t}{\mathrm{d}\pi}\mathrm{d}\mu_t \\
&= \int_\Omega \frac{\partial \mu_t(x)}{\partial t}\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x + \int_\Omega \mu_t(x)\frac{\partial \log\frac{\mu_t(x)}{\pi(x)}}{\partial t}\mathrm{d}x \\
&= \int_\Omega \frac{\partial \mu_t(x)}{\partial t}\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x + \int_\Omega \mu_t(x)\frac{\partial \log\mu_t(x)}{\partial t}\mathrm{d}x \\
&= \int_\Omega \frac{\partial \mu_t(x)}{\partial t}\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x + \int_\Omega \frac{\partial \mu_t(x)}{\partial t}\mathrm{d}x \\
&= \int_\Omega \frac{\partial \mu_t(x)}{\partial t}\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x + \frac{\partial}{\partial t}\int_\Omega \mu_t\mathrm{d}x \\
&= \int_\Omega \frac{\partial \mu_t(x)}{\partial t}\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x \left(\text{as, } \forall t\geq 0, \int_\Omega \mathrm{d}\mu_t = 1\right).
\end{aligned}$$

Furthermore, μ_t is the unique solution of the linear transport equation of Theorem 2.2, where $\phi^\star_{\mu_t} = S_{\mu_t}\nabla\log\frac{\pi}{\mu_t}$. Thus, we have

$$\begin{aligned}
\frac{\partial KL(\mu_t||\pi)}{\partial t} &= -\int_\Omega \nabla\cdot(\mu_t(x)\phi^\star_{\mu_t}(x))\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x \\
&= \int_\Omega \mu_t(x)\phi^\star_{\mu_t}(x)\cdot\nabla\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}x \quad \left(\phi^\star_{\mu_t}\in\mathcal{S}_{\mu_t}\right) \\
&= \int_\Omega \phi^\star_{\mu_t}(x)\cdot\nabla\log\frac{\mu_t(x)}{\pi(x)}\mathrm{d}\mu_t(x) \\
&= \left\langle \iota\phi^\star_{\mu_t}, \nabla\log\frac{\mu_t}{\pi}\right\rangle_{L^2_\mu(\Omega,\Omega)} \\
&= \left\langle \phi^\star_{\mu_t}, S_{\mu_t}\nabla\log\frac{\mu_t}{\pi}\right\rangle_{\mathcal{H}} \\
&= \left\langle \phi^\star_{\mu_t}, -S_{\mu_t}\nabla\log\frac{\pi}{\mu_t}\right\rangle_{\mathcal{H}} \\
&= -\left\langle \phi^\star_{\mu_t}, \phi^\star_{\mu_t}\right\rangle_{\mathcal{H}} \\
&= -\left\|\phi^\star_{\mu_t}\right\|^2_{\mathcal{H}} \\
&= -\mathfrak{K}(\mu_t|\pi).
\end{aligned}$$

■

A.8. Proof of Lemma 2.4

Proof. We recall that

$$\mathfrak{K}(\mu|\pi) = \mathbb{E}_{x\sim\mu}\left[\nabla\log\pi(x)^\top\phi^\star_\mu(x) + \nabla\cdot\phi^\star_\mu(x)\right].$$

The right implication is straightforward. Assume that $\mu = \pi$. We know that $\phi_\mu^\star$ is in $\mathcal{S}(\pi)$, thus

$$\begin{aligned}
&\mathbb{E}_{x\sim\mu}\left[\nabla \log \pi(x)^\top \phi_\mu^\star(x) + \nabla \cdot \phi_\mu^\star(x)\right] \\
&= \int_\Omega \nabla \log \pi(x)^\top \phi_\mu^\star(x) + \nabla \cdot \phi_\mu^\star(x)\, \mathrm{d}\mu(x) \\
&= \int_\Omega \nabla \log \pi(x)^\top \phi_\mu^\star(x) + \nabla \cdot \phi_\mu^\star(x)\, \mathrm{d}\pi(x) \\
&= \int_\Omega \pi(x)\left(\nabla \log \pi(x)^\top \phi_\mu^\star(x) + \nabla \cdot \phi_\mu^\star(x)\right)\, \mathrm{d}x \\
&= \int_\Omega \nabla\pi(x)^\top \phi_\mu^\star(x)\, \mathrm{d}x - \int_\Omega \nabla\pi(x)^\top \phi_\mu^\star(x)\, \mathrm{d}x \\
&= 0.
\end{aligned}$$

The left implication is more involved. Assume that $\mathfrak{K}(\mu|\pi) = 0$. Remember that

$$\mathfrak{K}(\mu|\pi) = \left\|\phi_\mu^\star\right\|_{\mathcal{H}}^2 = \left\langle \nabla \log \frac{\pi}{\mu}, \iota S_\mu \nabla \log \frac{\pi}{\mu} \right\rangle_{L^2_\mu(\Omega,\Omega)}.$$

Thus, we can rewrite the KSD as

$$\mathfrak{K}(\mu|\pi) = \int_\Omega \int_\Omega \nabla \log \frac{\pi}{\mu}(x)^\top k(x', x) \nabla \log \frac{\pi}{\mu}(x')\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(x').$$

Since k is positive definite, we have that

$$\mathfrak{K}(\mu|\pi) = 0 \iff \nabla \log \frac{\pi}{\mu}(x) = 0, \overline{\forall}_\mu x \in \Omega.$$

Finally, as $\mu(\cdot)$ and $\pi(\cdot)$ are probability densities, we have that

$$\nabla \log \frac{\pi}{\mu}(x) = 0 \iff \pi(x) = \mu(x), \overline{\forall}_\mu x \in \Omega.$$

For more details, see Lean proof [5]. ■

A.9. Proof of Lemma 2.5

Proof. We first show that π is a fixed point of $(\mu : E) \mapsto \Phi_t(\mu)$, i.e. $\Phi_t(\pi) = \pi$. To do so, recall that

$$\mathfrak{K}(\pi|\pi) = \|\phi_\pi^\star\|_{\mathcal{H}}^2.$$

Using the right implication of Lemma 2.4, we have that

$$\|\phi_\pi^\star\|_{\mathcal{H}}^2 = 0,$$

which implies that

$$\iff \phi_\pi^\star(x) = 0, \overline{\forall}_\pi x \in \Omega.$$

Thus, $\overline{\forall}_\pi x \in \Omega$,

$$T_\pi(x)|_{\varepsilon=0} = x + \varepsilon\phi_\pi^\star(x) = x,$$

implying $\Phi_t(\pi) = \pi$.

Then, suppose that $\exists \nu \in E$ such that $\nu \neq \pi$ and $\Phi_t(\nu) = \nu$ for any $t \geq 0$. We have that

$$\frac{\partial KL(\Phi_t(\nu)||\pi)}{\partial t} = 0 = -\mathfrak{K}(\nu||\pi).$$

However, using the left implication of Lemma 2.4, we obtain a contradiction.

For more details, see Lean proof [6]. ■

[5] gaetanserre.fr/assets/Lean/SBS/html/KSD.lean.html
[6] gaetanserre.fr/assets/Lean/SBS/html/KSD.lean.html

A.10. Proof of Theorem 2.6

Proof. As stated in Theorem 2.3, $t \mapsto KL(\mu_t||\pi)$ is decreasing. Moreover, as $KL(\mu||\pi)$ is finite, it exists a positive real constant c, such that, for any sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \to \infty$, $KL(\mu_{t_n}||\pi) \to c$. It implies that, for any such sequence $(t_n)_{n\in\mathbb{N}}$, it exists a subsequence $(t_k)_{k\in\mathbb{N}}$ such that $\mu_{t_k} \rightharpoonup \mu_\infty$, meaning that $\Phi_t(\mu) \rightharpoonup \mu_\infty$ (see Theorem 2.6 (Billingsley, 1999)). Therefore, μ_∞ is a fixed point of Φ_t, for any $t \geq 0$ and any $\mu \in \mathcal{P}_2(\Omega)$ such that $KL(\mu||\pi)$ is finite. Finally, using Lemma 2.5, we have that $\mu_\infty = \pi$. ■