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
Centre Borelli
Department of Mathematics
École Normale Supérieure Paris-Saclay
gaetan.serre@ens-paris-saclay.fr

Argyris Kalogeratos
Centre Borelli
Department of Mathematics
École Normale Supérieure Paris-Saclay
argyris.kalogeratos@ens-paris-saclay.fr

Nicolas Vayatis
Centre Borelli
Department of Mathematics
École Normale Supérieure Paris-Saclay
nicolas.vayatis@ens-paris-saclay.fr

Abstract

In this paper, we introduce a new flow-based method for global optimization of continuous Sobolev functions, called *Stein Boltzmann Sampling* (SBS). Our method samples from the Boltzmann distribution, as it 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* (SVGD) algorithm. We prove the asymptotic convergence of our method by introducing a novel framework of the SVGD theory, suitable for global optimization, that allows to address more general target distributions over a compact subset of $\mathbb{R}^d$. We present two SBS variants and provide a detailed comparison with several state-of-the-art global optimization algorithms on various benchmark functions, showing that SBS and its variants are highly competitive. Its design of our method, the theoretical results, and the experiments suggest that SBS is particularly well-suited to be used as a continuation for particles or distribution-based methods, conjointly with particles filtering strategies, to produce sharp approximations while making a good use of the budget.

1 Introduction

In this paper, we consider global optimization of an unknown continuous, Sobolev, 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 purpose, 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 and flow-based method called *Stein Boltzmann*

Preprint. Under review.

Sampling (SBS) for Sobolev functions. It uses the *Stein Variational Gradient Descent* (SVGD) (Liu & Wang, 2016) method to sample from the Boltzmann distribution, which has the characteristic of converging towards a distribution supported on all 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 Sobolev function using elements of the SVGD theory. We show that the SBS method achieves competitive performance on standard global optimization benchmarks versus five 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 one, BAYESOPT (Martinez-Cantin, 2014) is well-known and adapted for small budget. The third one uses MALA to sample from the Boltzmann distribution, in a similar way to our method (Grenander & Miller, 1994; Welling & Teh, 2011; Raginsky et al., 2017). The last 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 two 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 over a compact subset of $\mathbb{R}^d$ for a class of target distribution. The class of distribution considered is more general than the one considered usually and allows to show the convergence of SBS for any continuous Sobolev function. In Appendix, we provide detailed definitions and results of the SVGD theory in this novel framework. To ensure ensure the correctness and reproducibility of the technical proofs, for some of the results, we provide links to proofs in Lean, a proof assistant (de Moura & Ullrich, 2021; mathlib Community, 2020). Then, we introduce two SBS variants: one that uses particle filtering to reduce the budget needed, and a hybrid 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 five 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.

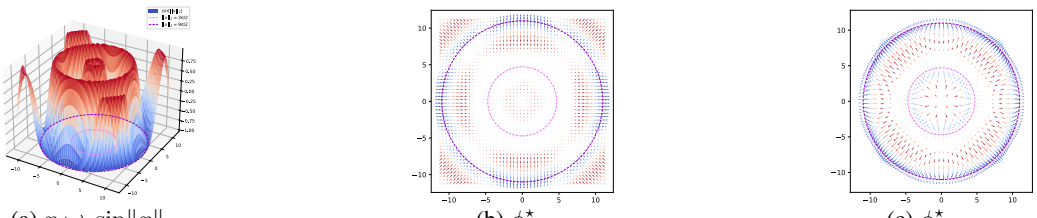
Notations. We consider the following notations: $d \in \mathbb{N}$ is the dimension of the optimization problem; $f : \Omega \to \mathbb{R}$ is the function to optimize, its domain $\Omega \subset \mathbb{R}^d$ is a smooth, connected and compact set; $x^* \in X^*$ is one of the global minima of f, i.e. $\forall x^*,\ f^* = f(x^*)$. 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 $\operatorname{supp}(f) = \{x \in \Omega \mid f(x) \neq 0\}$. 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.

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

For any natural numbers m and p, let $W^{m,p}$ be the Sobolev space of functions with m weak derivatives in $L^p_\mu(\Omega)$:

$$W^{m,p} \triangleq \left\{ f \in L^p_\mu(\Omega) \;\middle|\; \forall \alpha \in \mathbb{N}^d, |\alpha| \leq m, D^\alpha f \in L^p_\mu(\Omega) \right\},$$

where μ is clear from the context. Let the Hilbert space H^m be the Sobolev space $W^{m,2}$.



(a) $x \mapsto \sin\|x\|_2$ (b) $\phi^\star_{\mu_0}$ (c) $\phi^\star_{\hat{\mu}_{700}}$

Figure 1: Illustration of the vector field $\phi^\star$ in the discrete setting where π is the BD w.r.t. $x \mapsto \sin\|x\|_2$. One can see that, the particles are attracted to the two $d-1$ manifolds where the function is minimized, and, after some SVGD iterations, they are concentrated around these regions.

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, as it is a classical object in the global optimization theory and as it makes a link between our method and the simulated annealing method (Kirkpatrick et al., 1983) (see Section 6).

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 $m^{(\kappa)}_{f,\Omega} : \Omega \to \mathbb{R}_{\geq 0}$ defined by:

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

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 2). If $\lambda(X^*) = 0$, it tends to a distribution over X^* where the concentration of the mass depends on the local geometry of the minimizing manifolds (Hwang, 1980). More details can be found in Appendix A.1. The SBS method aims to sample from the BD with κ large enough in order for the function values at the sampled points 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) algorithm. While the SVGD theory and dynamics has been deeply studied in the literature (e.g. (Liu, 2017; Lu et al., 2019; Korba et al., 2020; Duncan et al., 2023; Sun et al., 2023)), its use for global optimization has not been considered. Thus, we introduce a new framework of the SVGD theory, suitable for global optimization, that allows to address more general target distributions over Ω and we prove classical results in this new framework (see Appendix A.2).

Given an initial measure μ, SVGD constructs iteratively a sequence of measures that get closer to the target measure (in terms of KL-divergence), noted 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 the gradient operator is understand in the sense of distributions and k is the reproducing kernel of a specific RKHS $\mathcal{H}$ (see Appendix A.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. We also provide a collection of all the key notations and their meaning used throughout the paper in Table 3. Next in this section, we prove the asymptotic convergence of SBS.

2.2 Asymptotic convergence of SBS

First, define $\mathcal{P}_n(\Omega)$ as the set of probability measures on Ω such that

$$\forall \mu \in \mathcal{P}_n(\Omega),\ \mu \ll \lambda \ \wedge\ \mu(\cdot) \in W^{1,n}(\Omega) \ \wedge\ \mathrm{supp}(\mu(\cdot)) = \Omega,$$

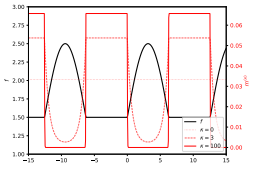


Figure 2: The Boltzmann p.d.f. becomes uniform over the set of minimizers X^* of the given function f to optimize, as κ grows, tending to infinity.

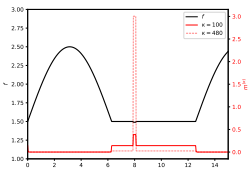


Figure 3: The volume of the global minimizers is much smaller than the volume of the local minimizers. The value of the function at the local minimizers is close to the value of the global ones. In Figure 2, $\kappa \triangleq 100$ is sufficient for the the majority of the mass to be concentrated around the global minimizers. Here, κ needs to be bigger.

where $\mu(\cdot)$ is the density of μ w.r.t. λ. To prove the asymptotic convergence of SBS, we need to prove that the sequence of measures constructed by SVGD, noted $(\mu_n)_{n\in\mathbb{N}}$ (for more details, see Appendix A.2), 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. To use theoretical results of our new SVGD framework, we need to ensure μ and π belongs to $\mathcal{P}_2(\Omega)$. For the latter, we assume that f is in $C^0(\Omega) \cap W^{1,4}(\Omega)$ so that m^κ is in $H^1(\Omega)$ (see proof in Appendix B.2). Theorem 2.2 and Theorem 2.3 are known results in the literature that we prove in our new SVGD framework. Then, we introduce two lemmas that are crucial to prove the asymptotic convergence of SBS.

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 and $\mu \in \mathcal{P}_2(\Omega)$. 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, μ_t is the unique solution of the following nonlinear transport equation:*

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

where $(\nabla \cdot)$ is the divergence operator, in the sense of distributions (see details in Appendix B.7). Moreover, the sequence $(\mu_n)_{n\in\mathbb{N}}$, constructed by SVGD, is a discretized solution of (2), considering the vector field $\phi^\star_{\mu_t}$. One can consider the resulting flow of measures

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

We provide a different proof of this theorem in Appendix B.7, 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 a discretized solution of (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 μ and π in $\mathcal{P}_2(\Omega)$ called *Kernelized Stein Discrepancy* (KSD). Both objects are defined in Appendix A.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*

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

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

(See proof in Appendix B.8). In order to show the convergence of continuous-time SVGD, we proved that the KSD is a valid discrepancy measure.

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

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

(See proof in Appendix B.9). The previous lemmas directly imply that π is the unique fixed point of the flow of measures Φ.

Lemma 2.5 (Unique fixed point)**.** *Let $\pi \in \mathcal{P}_2(\Omega)$ and Φ be the flow of measures defined in Theorem 2.2. Then, for any $t \geq 0$, π is the unique fixed point of $(\mu : \mathcal{P}_2(\Omega)) \mapsto \Phi_t(\mu)$.*

Since $\mathfrak{K}(\mu|\pi) = \left\|\phi^\star_\mu\right\|^2_{\mathcal{H}}$ (see Appendix A.2), the proof is straightforward using the previous lemmas. See complete proof in Appendix B.10. Finally, we provide a proof of the weak convergence of μ_t to π.

Theorem 2.6 (Weak convergence of SVGD)**.** *Let $\mu, \pi \in \mathcal{P}_2(\Omega)$. 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 B.11. 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: $X = \left(x^{(1)}, \dots, x^{(N)}\right) \sim \mu^{\otimes N}$, and then computes the next ones as follows:

$$\begin{aligned} X_{n+1} &= X_n + \varepsilon \phi^{\star}_{\hat{\mu}_n}\left(X_n\right), \\ \text{where } \ \hat{\mu}_n(A) &= \frac{1}{N}\sum_{i=1}^{N} \delta_{X_n^{(i)}}(A). \end{aligned} \tag{3}$$

One can see a representation of the vector field $\phi^{\star}$ in the discrete setting where π is the BD in Figure 1. 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 in $C^0(\Omega) \cap W^{1,4}(\Omega)$. Let $\kappa > 0$ and let π be the BD defined in Definition 2.1 associated with f and κ. Let $\mu_0 \in \mathcal{P}_2(\Omega)$ and let $\hat{\mu}_n$ be defined by (3). Then,*

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

The proof relies on three main components: the almost sure convergence of the empirical measure $\hat{\mu}_n$ to μ_n, the weak convergence of μ_n to π using Theorem 2.2 and Theorem 2.6 (that are applicable as $f \in C^0(\Omega) \cap W^{1,4}(\Omega)$), and the fact that the BD tends to a distribution supported over the set of minimizers X^* as κ tends to infinity. The proof is provided in Appendix B.12.

To summarize, we proved that SBS is asymptotically convergent for any functions that are continuous and belong to $W^{1,4}(\Omega)$. We adapted the theory of SVGD for target measures that are in $\mathcal{P}_2(\Omega)$ over a compact subset of $\mathbb{R}^d$ (see Appendix A.2). This is a different framework than the one usually considered in the literature, where the target density is smooth and the domain is $\mathbb{R}^d$ (e.g. (Liu & Wang, 2016; Liu et al., 2016)). Some works have been done to relax the assumptions on the target measure (e.g. (Korba et al., 2020; Sun et al., 2023)). However, thanks to the compactness of Ω, our assumptions on π are less restrictive and only consider integration constraints on its 1st order weak derivatives, making our framework more adapted to global optimization problems.

The implementation of SBS uses (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 μ_0 to be the uniform distribution on Ω as it maximizes the entropy (related to the exploration aspect of the method) and we use the RBF kernel function. These two objects are used in most literature on SVGD and meet the requirements of the theory. To better understand the previous results and objects involved, we

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

Figure 4: 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 comparable performance.

introduce an non-exhaustive list of definitions and theoretical results related to SVGD Appendix A.2.

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 μ_0 over the particles
Output: $\hat{x}$, an estimate of x^*

Sample N particles: $X_1 \leftarrow \left(x^{(1)}, \dots, x^{(N)}\right) \sim \mu_0^{\otimes N}$
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
$\hat{x} \leftarrow \arg\min_{1 \le j \le N} f(X_{n+1}^{(j)})$ the "best" particle
return $\hat{x}$

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

3 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 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 4. 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 in a significant reduction of the budget used, while having comparable 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-HYBRID is that it provides very good results while it is still asymptotically consistent, since the distribution of the particles induced by WOA and CMA-ES meet the assumptions of Theorem 2.6.

4 Choice of hyperparameters

In this section, we discuss the choice of the hyperparameters of SBS and its variants. We focus on the choice of κ and the kernel, as they carry complex information about the behavior of SBS.

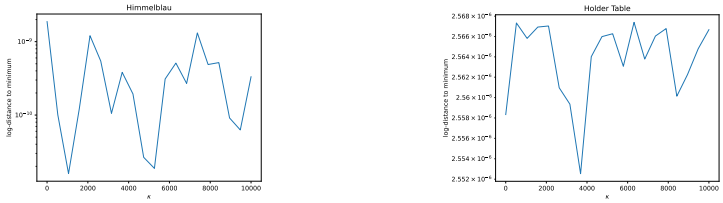


Figure 5: log-distance to the global minimum versus the value of κ for SBS on the Himmelblau function (left) and the Holder Table function (right). One can see that the choice of κ does not significantly affect the performance of SBS, on both noisy and smooth functions.

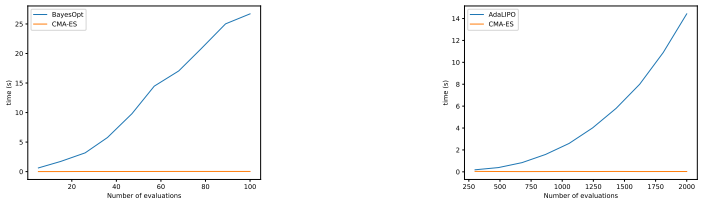


Figure 6: Execution time versus the number of evaluations for BAYESOPT, ADALIPO and CMA-ES on the Himmelblau function. One can see that the time to run is significantly higher for BAYESOPT and ADALIPO than for CMA-ES. Thus, the budget allowed for these two methods is set lower than the others.

Choice of κ. As detailed in the theoretical analysis sections, κ controls the shape of the distribution SVGD samples from. The bigger κ, the more the mass of the distribution is concentrated around the global minima of the function. Intuitively, the optimal κ to choose for a satisfying amount of the mass to be around the global minima depends on the geometry of the function around local minima (the asymptotic behavior of the BD depends on the local geometry, see (Hwang, 1980)). However, one can see in Figure 5 that, in practice, the choice of κ does not significantly affect the performance of SBS. The reason is that, if the modes of the BD that contains the most mass are the ones around the global minima, SVGD succeeds in moving some particles in these modes (given enough particles). Moreover, given κ, for those modes to not contain the most mass, it would require that the volume of one or several local minimizers that have a close value to the global minimum is much larger than the volume of the global minimizers (see Figure 3). These two conditions are interdependent. Choosing a large κ ensures that, if the latter event happens, either the volume of the local minimizers is much larger than the volume of the global minimizers (which is unlikely in practice) or the value of the local minimizers is very close to the value of the global minimizers (which is a good thing). Thus, the choice of κ can be set to a large value, such as 10^3, to have a good performance on average.

Choice of σ. In all variants of SBS, we use the RBF kernel with a bandwidth σ. The choice of σ is crucial for the performance of the methods. As detailed in Section 6, the size of σ controls the forces that occur between particles. When a lot of particles are close, they repel each other. This behavior allows to explore the domain of the function. However, it also prevents SBS from converging in narrow regions, where global minima could be located. Then, a natural choice of σ is $\frac{1}{N^2}$, where N is the number of particles. This choice ensures that, with few particles, σ is large enough to for SBS to explore the domain while, with a lot a particles, the exploration is ensured by the uniform distribution μ_0 and the small σ allows the particles to converge to the global minimum. For the SBS-PF variant, σ changes during the optimization process, as particles are being removed. For SBS-HYBRID, as the initial particles are supposed to be close to the global minimum, σ is set to a very small value, such as 10^{-10}.

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) (a particle-swarm method), ADALIPO (Malherbe & Vayatis, 2017), BAYESOPT (Martinez-Cantin, 2014), and a similar method to SBS but using MALA instead of SVGD (Grenander & Miller, 1994; Welling & Teh, 2011; Raginsky et al., 2017). We also provide a method that combines SBS-PF and SBS-HYBRID, named SBS-PF-HYBRID. We

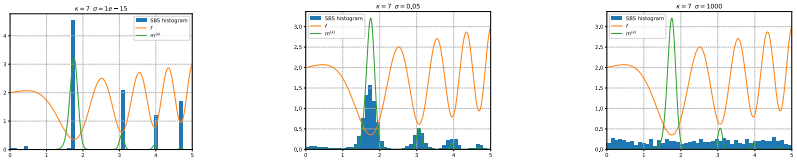


Figure 7: Illustration of the exploration/exploitation trade-off in SBS with different values of σ. In blue, the repartition of the particles, in orange, $x \mapsto \cos(x^2) + x/5 + 1$, in green $m^{(\kappa)}$.

use classical two dimensional benchmark functions for global optimization. Some are noisy and multimodal (Ackley, Drop wave, Egg Holder, Holder Table, Michalewicz, Rastrigin, Levy), some are smooth (Branin, Goldstein Price, Himmelblau, Rosenbrock, Camel, Sphere) (see more information in (Surjanovic & Bingham)). We provide the implementation [1] of this experiment. For the results of Table 1, we ran each method 10 times on each function. In the literature, the budget is defined as the number of function evaluations. However, the computational time can vary significantly between the methods and needs to be taken into account. Thus, the budget is set in order for the methods to stop in a reasonable time. As one can see in Figure 6, the time to run ADALIPO and BAYESOPT is significantly higher than the other methods. Thus, the budget allowed for these two methods is set lower than the others: 2K for ADALIPO, 100 for BAYESOPT and 800K for the others.

We introduce a metric named *Competitive ratio*:

$$\text{Competitive ratio}(m) = \frac{1}{|F|} \sum_{f \in F} \min \left(100, \frac{df_{\text{m}}}{df_{\text{best}}} \right),$$

where F is the set of benchmark functions, df_{best} is the smallest distance to the global minimum of f among all the methods and df_{m} is the distance to the global minimum of f found by the method m. This metric provides information on the average precision compared to the best method (lower is better, best is 1).

As one can see, SBS outperforms the state-of-the-art methods and score the third rank on average. SBS-PF achieves comparable results on average with significantly less evaluations ($\sim 97\%$ budget reduction). Even if the asymptotic consistency does not hold for SBS-PF, the particle filtering strategy allows to beat almost all SOTA methods (except WOA) while using only a fraction of the initial budget. Moreover, SBS-HYBRID and SBS-PF-HYBRID outperform all the other methods on average. They combine the efficiency of both CMA-ES and WOA with the large budget compatibility of SBS, while the addition of particle filtering reduce the budget by $\sim 67\%$. In parallel, SBS, SBS-PF-HYBRID and SBS-HYBRID score respectively the fourth, second and first rank on the competitive ratio metric, showing that their approximation are precise, compared to the other methods. SBS and SBS-HYBRID succeed in beating CMA-ES and WOA while being asymptotically consistent. In Appendix, we provide the results of the same experiment on 50 dimensional benchmark functions for low computational time methods (see Table 2). The budget is set to 8M. One can observe a fairly similar behavior as in the 2d case. However, the budget reduction of SBS-PF-HYBRID is less significant ($\sim 9\%$): the high dimensionality of the functions and the initial distribution makes the unpromising particles harder to distinguish.

We believe that, with more sophisticated particle filtering and adaptive locality of the kernel, SBS could be further improved (see Section 6). Note that, in order to update the particles, SBS needs to compute the gradient of the function. In our implementation, we estimate it using finite differences. However, it takes the majority of the budget. More sophisticated methods, such as automatic differentiation, could 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 (Azencott, 1989), while SBS is a deterministic variational approach.

[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). SBS-HYBRID runs 1K iterations of CMA-ES and WOA. As one can see, SBS-HYBRID and SBS respectively rank 1st and 3rd while SBS-PF-HYBRID and SBS-PF achieve competitive results with significantly less evaluations (respectively $\sim 67\%$ and $\sim 97\%$ budget reduction).

	STATE-OF-THE-ART					PROPOSED METHODS			
FUNCTIONS	LANGEVIN	BAYESOPT	ADALIPO	CMA-ES	WOA	SBS-PF	SBS	SBS-PF-HYBRID	SBS-HYBRID
ACKLEY	12.779	0.322	1.286	9.916	9×10^{-8}	0.002	8×10^{-4}	1×10^{-5}	5×10^{-6}
BRANIN	0.398	0.398	0.400	0.398	0.398	0.398	0.398	0.398	0.398
DROP WAVE	−0.052	−0.838	−0.955	−0.685	−1.000	−0.963	−0.981	−0.934	−0.981
EGG HOLDER	1049.132	−860.935	−937.983	−629.634	−959.641	−932.393	−958.142	−944.700	−946.280
GOLDSTEIN PRICE	2548.300	10.231	3.813	37.236	3.000	3.000	3.000	3.000	3.000
HIMMELBLAU	3×10^{-6}	6×10^{-4}	0.006	1×10^{-16}	3×10^{-6}	1×10^{-7}	9×10^{-11}	7×10^{-19}	9×10^{-21}
HOLDER TABLE	−9.234	−19.169	−19.184	−10.843	−19.208	−19.209	−19.209	−19.209	−19.209
MICHALEWICZ	-5×10^{-15}	−1.801	−1.790	−1.696	−1.801	−1.801	−1.801	−1.801	−1.743
RASTRIGIN	4.944	2.780	0.191	4.155	4×10^{-15}	0.100	1×10^{-9}	0.497	0.398
ROSENBROCK	0.534	0.106	0.037	1×10^{-15}	2×10^{-7}	4×10^{-5}	2×10^{-6}	5×10^{-17}	2×10^{-17}
CAMEL	161.823	−0.967	−1.016	−1.032	−1.032	−1.032	−1.032	−1.032	−1.032
LEVY	75.563	0.056	0.024	3.445	1×10^{-8}	9×10^{-8}	2×10^{-12}	1×10^{-19}	6×10^{-20}
SPHERE	1×10^{-5}	8×10^{-4}	6×10^{-4}	2×10^{-16}	3×10^{-16}	8×10^{-8}	8×10^{-12}	2×10^{-19}	1×10^{-21}
COMP. RATIO	85.732	92.385	92.385	54.309	31.511	69.558	33.567	31.462	**28.331**
AVERAGE RANK	8.46	7.23	6.54	6.00	3.54	4.92	3.38	2.77	**2.15**
FINAL RANK	9	8	7	6	4	5	3	2	**1**

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: $k(x, x') = \exp\left(-\|x - x'\|_2^2 / 2\sigma^2\right)$, as it is in the Stein class of any smooth density supported on $\mathbb{R}^d$. The bandwidth 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 7.

7 Conclusion

In this paper, we introduced SBS, a new method for global optimization. We proved that it is consistent using theory of the SVGD algorithm, that we extended to a more general class of target measure, thanks to the compactness of the domain. This new SVGD framework is particularly suitable for global optimization, as it allows to sample from the BD of any continuous function given integration constraints on its 1st order weak derivatives. We showed that SBS outperforms state-of-the-art methods in average on classical benchmark functions. We also introduced SBS-PF, a variant of SBS that uses particle filtering to save most of the budget while having comparable performance 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. Our work suggests that, in order to obtain precise approximation while having reduced budget, SBS should be use as a continuation for particles or distribution-based methods, conjointly with particles filtering strategies. For the future, we plan to study further the convergence rate of SBS and its components, and design more sophisticated particle filtering strategies to make it more appealing for global optimization in real-world applications.

Acknowledgments and Disclosure of Funding

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.

Azencott, R. Simulated annealing, 1989.

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.

Duncan, A., Nüsken, N., and Szpruch, L. On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 2023.

Evans, L. C. and Gariepy, R. F. *Measure Theory and Fine Properties of Functions, Revised Edition*. Chapman and Hall/CRC, 2015.

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

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

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., Salim, A., Arbel, M., Luise, G., and Gretton, A. A non-asymptotic analysis for stein variational gradient descent. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

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.

Martinez-Cantin, R. Bayesopt: A bayesian optimization library for nonlinear optimization, experimental design and bandits. *Journal of Machine Learning Research*, 2014.

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.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 2011.

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.

Sun, L., Karagulyan, A., and Richtarik, P. Convergence of stein variational gradient descent under a weaker smoothness condition. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Surjanovic, S. and Bingham, D. Virtual library of simulation experiments: Test functions and datasets. Retrieved April 29, 2024, from http://www.sfu.ca/~ssurjano.

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

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

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

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.

Table 2: **Comparative results.** Comparison between all SBS variants with several state-of-the-art methods on 50 dimensional benchmark functions. For each function, we report the average best function value found (lower is better). SBS-HYBRID runs 1K iterations of CMA-ES and WOA. As one can see, SBS-HYBRID and SBS respectively rank 2^{nd} and 4^{th} while SBS-PF-HYBRID and SBS-PF achieve competitive results with less evaluations (respectively $\sim 9\%$ and $\sim 97\%$ budget reduction). The high dimensionality of the functions makes the particle filtering of SBS-PF-HYBRID less efficient.

	STATE-OF-THE-ART			PROPOSED METHODS			
FUNCTIONS	WOA	LANGEVIN	CMA-ES	SBS-PF	SBS	SBS-PF-HYBRID	SBS-HYBRID
ACKLEY	19.737	21.514	19.420	18.935	19.039	19.631	19.479
MICHALEWICZ	-13.663	-0.823	-34.069	-11.905	-13.626	-32.673	-32.556
RASTRIGIN	570.841	25.207	101.817	276.598	267.643	127.023	107.124
ROSENBROCK	19021.088	0.254	3×10^{-14}	26.156	36.757	26.329	43.631
LEVY	215.674	2837.201	85.298	58.677	52.496	74.093	41.644
SPHERE	636.685	0.001	1×10^{-14}	1×10^{-4}	2×10^{-10}	9×10^{-23}	6×10^{-20}
COMP. RATIO	38.172	45.105	**1.686**	35.579	19.439	18.300	18.036
AVERAGE RANK	6.17	5.00	**2.50**	4.00	3.83	3.33	3.17
FINAL RANK	7	6	**1**	5	4	3	2

Table 3: Collection of all notations and their meanings

Notation	Definition
f	function to minimize
d	dimension of the domain of f
Ω	compact subset of $\mathbb{R}^d$, domain of f
X^*	set of global minimizers of f
$W^{p,m}$	Sobolev space of functions with p-integrable m-th order weak derivatives
H^m	$W^{2,m}$
λ	Lebesgue measure
$m^{(\kappa)}$	density of the BD with parameter κ
$\mathcal{A}_\mu$	the Stein operator associated to the measure μ
$\mathcal{S}(\mu)$	the Stein class of the measure μ
$\mathcal{P}_2(\Omega)$	the set of probability measures supported over Ω with density in H^1
π	target measure, the BD of f in SBS context
$\mathcal{H}_0$	the foundational RKHS of SVGD
k	the kernel of the RKHS $\mathcal{H}_0$
$\mathcal{H}$	the product RKHS of SVGD constructed using $\mathcal{H}_0$
T_μ	an integral operator from $L^2_\mu(\Omega)$ to $\mathcal{H}_0$
S_μ	an integral operator from $L^2_\mu(\Omega,\Omega)$ to $\mathcal{H}$ constructed using T_μ
ϕ^*_μ	the optimal transport vector field in $\mathcal{H}$ constructed by SVGD
$\mathfrak{K}(\mu\|\pi)$	the Kernelized Stein Discrepancy
$(\mu_n)_{n\in\mathbb{N}}$	sequence of measures constructed by SVGD
$\hat{\mu}_n$	empirical measure of the SVGD particles
$(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$	net extension of $(\mu_n)_{n\in\mathbb{N}}$
$\Phi : \mathbb{R}_{\geq 0} \times \mathcal{P}_2(\Omega)$	the flow of measures associated to $(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$

A Background

In this section, we introduce some background results related to the Boltzmann distribution (BD) and the SVGD theory. Please note that, concerning the BD, these results are not new. Concerning SVGD, we introduce a novel framework, suitable for global optimization, in which we prove classical results of SVGD theory. The purpose of this section is to provide a self-contained presentation of the theory behind SBS for the reader and to show the consistency of our novel SVGD framework.

A.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 B.1.

Properties A.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, $\forall x \in \Omega$,

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

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

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

A visual representation of the BD is given in Figure 2. 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 (1). We provide the proof of the properties in Appendix B.1. To sample from this 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.

A.2 Stein Variational Gradient Descent

Sampling from an intractable distribution is a common task in Bayesian inference, where the target distribution is a posterior one. 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, less sensitive to the choice of the step-size but not suitable for non-convex objectives. The remainder of this section introduces key definitions and theoretical results related to SVGD and shows that they hold when considering a compact domain Ω and a target density in $H^1(\Omega)$: a novel framework particularly suitable for global optimization that we use to prove the consistency of SBS (see Section 2.2).

A.2.1 Definitions

For any natural number n, we start by defining the set of probability measures on Ω that have a density w.r.t. the Lebesgue measure and are in $W^{1,n}(\Omega)$. Let $\mathcal{P}_n(\Omega)$ denote the set of probability measures on Ω such that

$$\forall \mu \in \mathcal{P}_n(\Omega),\ \mu \ll \lambda\ \wedge\ \mu(\cdot) \in W^{1,n}(\Omega)\ \wedge\ \mathrm{supp}(\mu(\cdot)) = \Omega,$$

where $\mu(\cdot)$ is the density of μ w.r.t. λ. In SVGD theory, μ and π must belong to $\mathcal{P}_2(\Omega)$. Thus, their densities lie in $H^1(\Omega)$. The condition on their support ensures that the KL divergence is well-defined. In the following, we denote the density w.r.t. λ of an arbitrary measure μ by the function $\mu : \Omega \to \mathbb{R}_{\geq 0}$.

A.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 (∇) and $(\nabla \cdot)$ are respectively the gradient and the divergence operators, in the sense of distributions. 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 A.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 $\phi \in H^1(\Omega)$ and

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

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 A.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 B.3). Now, one can consider:

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

If $\mu \neq \pi$, (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) chose $\mathcal{F}$ to be a universal vector-valued RKHS, which allows to find closed-form solution to (5). The Stein discrepancy restricted to that RKHS is known as *Kernelized Stein Discrepancy*.

A.2.3 Kernelized Stein Discrepancy

From now on, we consider $\mu, \pi \in \mathcal{P}_2(\Omega)$ such that π is the target measure. Next, we define the vector-valued RKHS that will be used in the Kernelized Stein Discrepancy.

Definition A.4 (Product RKHS (Liu & Wang, 2016)). Let $k : \Omega \times \Omega \to \mathbb{R}$ be a continuous, symmetric, and integrally positive-definite kernel such that $\forall x \in \Omega, k(\cdot, x) \in \mathcal{S}(\mu) \cap \mathcal{S}(\pi)$ and $\nabla_{xy} k(x, y) \in L^2_\mu(\Omega)$ (in the sense of distributions). 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. $\forall f = (f_1, \dots, f_d)^\top, f \in \mathcal{H} \iff \forall 1 \leq i \leq d, f_i \in \mathcal{H}_0$. 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.

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

As k is integrally positive-definite, $\mathcal{H}_0$ is dense in $L^2_\mu(\Omega)$ (see (Sriperumbudur et al., 2011)), which shows its expressiveness. We proved that the integral operator

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

is a mapping from $L^2_\mu(\Omega)$ to $\mathcal{H}_0$, i.e. $T_\mu : L^2_\mu(\Omega) \to \mathcal{H}_0$. (See proof in Appendix B.4). This allows to define another integral operator

$$\begin{aligned} S_\mu :\ & L^2_\mu(\Omega, \Omega) \to \mathcal{H} \\ & f \mapsto (T_\mu f^{(1)}, \dots, T_\mu f^{(d)})^\top, \end{aligned}$$

where T_μ is applied component-wise. The proof in Appendix B.4 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:

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

We can now define the KSD.

Definition A.5 (Kernelized Stein Discrepancy (Liu et al., 2016))**.** Let $\mathcal{H}$ be a product RKHS as defined in Definition A.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 A.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 $\phi^\star_\mu = \mathbb{E}_{x\sim\mu}[\nabla \log \pi(x) k(\cdot, x) + \nabla_x k(\cdot, x)]$. *As* $\operatorname{supp}(\pi) = \Omega$, $\phi^\star_\mu$ *is well-defined. It is the steepest trajectory in* $\mathcal{H}$ *that maximizes* $\mathfrak{K}(\mu|\pi)$. *The KSD is then given by*

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

The proof strategy is to remark that, for any function $f \in \mathcal{H}$, $\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi f(x)] = \langle f, \phi^\star_\mu \rangle_{\mathcal{H}}$. Then, the result follows from the Cauchy-Schwarz inequality. (See proof in Appendix B.5). This leads to the following result of the SVGD theory.

Theorem A.7 (KL steepest descent trajectory (Liu & Wang, 2016))**.** *Let* $\mathcal{H}$ *be a product RKHS (Definition A.4). Let* $\phi^\star_\mu \in \mathcal{H}$ *be as defined in Theorem A.6. Let* $\varepsilon > 0$ *and*

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

Then,

$$\arg\min_{\phi \in \mathcal{H}} \left\{ \nabla_\varepsilon \mathrm{KL}(T_\varepsilon(\phi)_\# \mu || \pi)|_{\varepsilon=0} \mid \|\phi\|_{\mathcal{H}} \leq 1 \right\} = \frac{\phi^\star_\mu}{\|\phi^\star_\mu\|_{\mathcal{H}}},$$

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

(See proof in Appendix B.6). This last result is the key of the SVGD algorithm. It means that $\phi^\star_\mu$ is the optimal direction (within $\mathcal{H}$) to update μ in order to minimize the KL-divergence between μ and π. As $\mathbf{0} \in \mathcal{H}$ (that nullifies the gradient), the result ensures that the gradient of $g : \varepsilon \mapsto \mathrm{KL}(T_\varepsilon(\phi^\star_\mu / \|\phi^\star_\mu\|_{\mathcal{H}})_\# \mu || \pi)$ is at most 0 and thus g is decreasing over $[0, \delta]$, for $\delta > 0$ small enough. Consequently, SVGD iteratively updates μ in the direction induced by $\phi^\star_\mu$, with a small step size ε:

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

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

Lemma A.8. *Let* $\mathcal{H}$ *be a product RKHS as defined in Definition A.4. Then,* $\phi^\star_\mu \in \mathcal{H}$ *as defined in Theorem A.6. Given the above assumption on* $\phi^\star_\mu$, *we have that*

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

Proof. We showed in Appendix B.5 that

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

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

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 $(\mathrm{KL}(\mu_n||\pi))_{n\in\mathbb{N}}$ is decreasing, given that the step size is small enough. In order to use theorems in Section 2.2, we need to ensure that

$$\phi_\mu^\star \in \mathcal{S}(\mu).$$

Given the assumption of the kernel, $\phi_\mu^\star$ lies in $H^1(\Omega)$. Thus, we need to choose k in order to ensure that the integral of $\phi_\mu^\star$ over $\partial\Omega$ is null. An easy way to guarantee this is to choose k such that

$$\begin{aligned}\forall x \in \Omega, \lim_{d(\{y\},\partial\Omega)\to 0} \nabla_x k(y,x) = 0, \text{and} \\ \lim_{d(\{y\},\partial\Omega)\to 0} \nabla \log \pi(x) k(y,x) = 0, \text{where} \\ d(A,B) = \inf\left\{\|a-b\|_2 \mid a \in A \wedge b \in B\right\}.\end{aligned}$$

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 and symmetric function such that f and its gradient w.r.t. x tend to 0 near the boundary of Ω sufficiently fast to ensure the previous conditions. This assumption allows to use Lemma A.3 with $\phi_\mu^\star$, for any $\mu \in \mathcal{P}_2(\Omega)$.

B 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]. Note that a collection of all key notations and their meanings is available in Table 3. 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 $\overline{\forall}$.

B.1 Proof of Properties A.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 B.1 (Nascent minima distribution properties)**.** *Let $m^{(\kappa)}$ and τ be defined in (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, $\forall x \in \Omega$,*

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

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

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,

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

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 B.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 B.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 A.1.

B.2 Proof of $f \in C^0(\Omega) \cap W^{1,4}(\Omega) \implies m^\kappa \in H^1(\Omega)$

Proof. As f and $\exp(\cdot)$ lie in $C^0(\Omega)$, $e^{-\kappa f}$ is also in $C^0(\Omega)$. As Ω is compact, $e^{-2\kappa f}$ is bounded. Thus, $e^{-\kappa f}$ lies in $L^2(\Omega)$:

$$\int_\Omega e^{-2\kappa f(x)}\,\mathrm{d}x < \lambda(\Omega) * C < \infty.$$

Moreover, $\forall \alpha \in \mathbb{N}^d$ such that $|\alpha| \leq 1$, we have

$$D^\alpha(e^{-\kappa f}) = -\kappa e^{-\kappa f} D^\alpha f.$$

As f is in $W^{1,4}(\Omega)$, $D^\alpha f$ is in $L^4(\Omega)$. Thus, $D^\alpha(e^{-\kappa f})$ is also in $L^2(\Omega)$:

$$\begin{aligned}
\int_\Omega \left(D^\alpha(e^{-\kappa f(x)})\right)^2\,\mathrm{d}x &= \int_\Omega -\kappa e^{-2\kappa f(x)} \left(D^\alpha f(x)\right)^2\,\mathrm{d}x \\
&= \langle -\kappa e^{-\kappa f}, (D^\alpha f)^2 \rangle_{L^2(\Omega)} \\
&\leq \left\| -\kappa e^{-2\kappa f} \right\|_{L^2(\Omega)} \left\| (D^\alpha f)^2 \right\|_{L^2(\Omega)} \\
&= \left\| -\kappa e^{-2\kappa f} \right\|_{L^2(\Omega)} \|D^\alpha f\|^2_{L^4(\Omega)} \\
&< \infty.
\end{aligned}$$

■

B.3 Proof of Lemma A.3.

Proof. As $\mu(\cdot)$ and ϕ are in $H^1(\Omega)$, and as Ω is smooth, one can apply the integration by parts formula in $\Omega \subset \mathbb{R}^d$ (see (Evans & Gariepy, 2015)):

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

■

B.4 Proof of T_μ is a map to $\mathcal{H}_0$

Proof. As k is continuous, symmetric, and positive-definite and as $\mu(\Omega) < \infty$ and as T_μ 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 (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 $\forall f \in L^2_\mu(\Omega), T_\mu f \in \mathcal{H}_0$. Let $f \in L^2_\mu(\Omega)$. We begin by proving that $T_\mu f \in L^2_\mu(\Omega)$.

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

Then,

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

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

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

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

B.5 Proof of Theorem A.6

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

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

As $\text{supp}(\mu) = \Omega$, $f^{(i)}$ is well-defined and, as π and μ are in $H^1(\Omega)$, $f^{(i)}$ is in $L^2(\Omega)$. Then, as $\forall x \in \Omega$, $k(\cdot, x) \in \mathcal{S}(\mu)$, it is easy to show that

$$(\phi_\mu^\star)^{(i)} = T_\mu f^{(i)} \in \mathcal{H}_0.$$

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

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

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

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

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

Thus, as $\|f\|_\mathcal{H} \leq 1$,

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

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

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

■

B.6 Proof of Theorem A.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 B.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 $\mathrm{KL}(T_{\#}\mu||\pi) = \mathrm{KL}(\mu||T_{\#}^{-1}\pi)$.

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

For more details, see Lean proof [3]. Thus, we have

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

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

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

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

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

This gives

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

Applying Theorem A.6 ends the proof. ■

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

B.7 Proof of Theorem 2.2

Proof. First, as Ω is a subset of a metric space (Euclidean space) and is compact, it is also complete for the induced metric. In addition, as it is connected, it is also path-connected. These properties combined with the fact that Ω is smooth ensure that Ω is a smooth complete manifold. Finally, 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 unique solution of the nonlinear 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.

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

where $\mathcal{M}(\Omega)$ is the set of measures on Ω, for any μ in $\mathcal{M}(\Omega)$, T_μ is the distribution associated with μ, and, for any φ 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 (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 nonlinear transport equation. ■

B.8 Proof of Theorem 2.3

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

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

Furthermore, μ_t is the unique solution of the nonlinear transport equation of Theorem 2.2, where $\phi^\star_{\mu_t} = S_{\mu_t} \nabla \log \frac{\pi}{\mu_t}$ (see Appendix B.5). Thus, we have

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

■

B.9 Proof of Lemma 2.4

Proof. We recall that, using Appendix B.5,

$$\mathfrak{K}(\mu|\pi) = \left\| \phi^\star_\mu \right\|^2_{\mathcal{H}} = \mathbb{E}_{x\sim\mu}\left[\mathcal{A}_\pi \phi^\star_\mu\right].$$

The right implication is straightforward. Assume that $\mu = \pi$. We know that $\phi^\star_\mu$ is in $\mathcal{S}(\mu) = \mathcal{S}(\pi)$, thus, using Lemma A.3, we have that

$$\mathbb{E}_{x\sim\mu}\left[\mathcal{A}_\pi \phi^\star_\mu\right] = \mathfrak{K}(\mu|\pi) = \mathbb{E}_{x\sim\pi}\left[\mathcal{A}_\pi \phi^\star_\mu\right] = 0.$$

The left implication is more involved. Assume that $\mathfrak{K}(\mu|\pi) = 0$. In Appendix B.5, we have shown that

$$\phi^\star_\mu = S_\mu \nabla \log \frac{\pi}{\mu}.$$

This implies that

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

Thus, one 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

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

Moreover, as the density of μ is supported over Ω, there is no set $E \subset \Omega$ such that $\lambda(E) > 0$ and $\mu(E) = 0$. Thus, a predicate $P(x)$ is true for almost all $x \in \Omega$, w.r.t. μ if and only if $P(x)$ is true for almost all x in Ω, w.r.t. λ.

Finally, if $\overline{\forall} x \in \Omega$, $\nabla \log \frac{\pi}{\mu}(x) = 0$, it implies that $\exists c \in \mathbb{R}_{>0}$ such that, $\mu(x) = c\mu(x)$. As $\mu(\cdot)$ and $\pi(\cdot)$ are probability densities over Ω, $c = 1$:

$$\mu(\Omega) = 1 = \int_\Omega \mu(x) \,\mathrm{d}x = c \int_\Omega \pi(x) \,\mathrm{d}x = c.$$

Thus,

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

For more details, see Lean proof [4]. ■

B.10 Proof of Lemma 2.5

Proof. We first show that π is a fixed point of $(\mu : \mathcal{P}_2(\Omega)) \mapsto \Phi_t(\mu)$, 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, $\overline{\forall}_\pi x \in \Omega$,

$$T_\pi(x) = x + \varepsilon \phi_\pi^\star(x) = x,$$

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

Then, suppose that $\exists \nu \in \mathcal{P}_2(\Omega)$ such that $\nu \neq \pi$ and $\Phi_t(\nu) = \nu$ for any $t \geq 0$. We have that

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

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

For more details, see Lean proof[4]. ■

B.11 Proof of Theorem 2.6

Proof. By construction of $\mathcal{P}_2(\Omega)$, $\mathrm{KL}(\mu||\pi)$ is finite. Moreover, as stated in Theorem 2.3, $t \mapsto \mathrm{KL}(\mu_t||\pi)$ is decreasing. Thus, it exists a positive real constant c, such that, for any sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \to \infty$, $\mathrm{KL}(\mu_{t_n}||\pi) \to c$. 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 $\mathrm{KL}(\mu||\pi)$ is finite. Finally, using Lemma 2.5, we have that $\mu_\infty = \pi$. ■

B.12 Proof of Theorem 2.7

Proof. In order to apply any SVGD theoretical results, we need to ensures that every assumptions are satisfied. First, we know by hypothesis that $\mu_0 \in H^1(\Omega)$ and the fact that $f \in C^0(\Omega) \cap W^{1,4}(\Omega)$ ensure that $\pi \in H^1(\Omega)$ as well (see Appendix B.2). We assume the kernel k to satisfy the conditions of Appendix A.2.3. We know by hypothesis and by construction of the BD that $\mathrm{supp}(\mu_0) = \mathrm{supp}(\pi) = \Omega$. Finally, $\mathrm{KL}(\mu_0||\pi)$ is finite. This allows to use Theorem 2.2 and Theorem 2.6, or any other theoretical results relative to SVGD.

Using the strong law of large numbers, one can show that, for any $n \in \mathbb{N}$, the discrete measure $\hat{\mu}_n$ converges almost surely to μ_n as $\hat{\mu}_n$ is the empirical measure arising from an i.i.d sequence of samples from μ_n. Thus, $\hat{\mu}_n \xrightarrow[N\to\infty]{a.s.} \mu_n$. Moreover, in Appendix B.7, we showed that

$$\forall \varphi \in C_c^\infty(\Omega), \langle T_{\mu_{n+1}-\mu_n}, \varphi \rangle \xrightarrow[\varepsilon\to 0]{} \langle T_{\nabla\cdot(\mu_n \phi_{\mu_n}^\star)}, \varphi \rangle.$$

Let μ be the net limit of $(\mu_n)_{n\in\mathbb{N}}$ as ε tends to 0:

$$\begin{aligned} \nu_\bullet : \mathbb{R}_{\geq 0} &\to \mathcal{P}_2(\Omega), \\ t &\mapsto \mu_{\lfloor t/\varepsilon \rfloor}, \end{aligned}$$

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

such that

$$(\mu_n)_{n\in\mathbb{N}} \xrightarrow[\varepsilon\to 0]{} (\nu_t)_{t\in\mathbb{R}_{\geq 0}},$$

for a certain notion of convergence in $\mathcal{P}_2(\Omega)$, e.g.

$$\forall t \in \mathbb{R}_{\geq 0}, \sup_{n\in\mathbb{N}} \{\|\mu_n - \nu_t\|_{\mathrm{TV}}\} \xrightarrow[\varepsilon\to 0]{} 0.$$

We have that, for any $t \in \mathbb{R}_{\geq 0}$ and any $\varphi \in C_c^\infty(\Omega)$,

$$\begin{aligned}
\lim_{\varepsilon\to 0} \langle T_{\nu_{t+\varepsilon}-\nu_t}, \varphi\rangle &= \lim_{\varepsilon\to 0} \langle T_{\mu_{\lfloor (t+\varepsilon)/\varepsilon\rfloor}} - \mu_{\lfloor t/\varepsilon\rfloor}, \varphi\rangle \\
&= \lim_{\varepsilon\to 0} \langle T_{\mu_{\lfloor t/\varepsilon\rfloor+1}} - \mu_{\lfloor t/\varepsilon\rfloor}, \varphi\rangle \\
&= -\varepsilon \langle T_{\nabla\cdot(\mu_{\lfloor t/\varepsilon\rfloor}\phi^\star_{\mu_{\lfloor t/\varepsilon\rfloor}})}, \varphi\rangle \\
&= -\varepsilon \langle T_{\nabla\cdot(\nu_t\phi^\star_{\nu_t})}, \varphi\rangle.
\end{aligned}$$

This allows to state on the derivative of ν:

$$\begin{aligned}
\lim_{\varepsilon\to 0} \int_\Omega \varphi \,\mathrm{d}\frac{\nu_{t+\varepsilon}-\nu_t}{\varepsilon} &= -\int_\Omega \varphi \nabla\cdot(\nu_t\phi^\star_{\nu_t})\,\mathrm{d}\nu_t \\
\iff \frac{\partial \nu_t}{\partial t} &= -\nabla\cdot(\nu_t\phi^\star_{\nu_t}).
\end{aligned}$$

Thus, as $\nu_0 = \mu_0$, $\nu_\bullet$ is a solution of (2). Finally, as $(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$ is the unique solution of the nonlinear transport equation, we have that

$$(\mu_n)_{n\in\mathbb{N}} \xrightarrow[\varepsilon\to 0]{} (\mu_t)_{t\in\mathbb{R}_{\geq 0}}.$$

This result is expected by construction of $(\mu_n)_{n\in\mathbb{N}}$ and (2).

Now, using Theorem 2.6, we have that $(\mu_n)_{n\in\mathbb{N}} \underset{\substack{\varepsilon\to 0\\ n\to\infty}}{\rightharpoonup} \pi$ and thus, $(\hat{\mu}_n)_{n\in\mathbb{N}} \underset{\substack{\varepsilon\to 0\\ n\to\infty\\ N\to\infty}}{\rightharpoonup} \pi$. The fact that the Boltzmann distribution ensures that π tends to a distribution supported on X^* as κ tends to ∞ (see Properties A.1) gives the desired result.

■