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 **Argyris Kalogeratos** **Nicolas Vayatis**
Centre Borelli, École Normale Supérieure Paris-Saclay

Abstract

We present a deterministic particle-based method for global optimization of continuous Sobolev functions, called *Stein Boltzmann Sampling* (SBS). SBS initializes uniformly a number of particles representing candidate solutions, then uses the *Stein Variational Gradient Descent* (SVGD) algorithm to sequentially and deterministically move those particles in order to approximate a target distribution whose mass is concentrated around promising areas of the domain of the optimized function. The target is chosen to be a properly parametrized Boltzmann distribution. For the purpose of global optimization, we adapt the generic SVGD theoretical framework for addressing more general target distributions over a compact subset of $\mathbb{R}^d$, and we prove SBS's asymptotic convergence. In addition to the main SBS algorithm, we present two variants: the SBS-PF that includes a particle filtering strategy, and the SBS-HYBRID one that uses SBS or SBS-PF as a continuation after other particle- or distribution-based optimization methods. An extensive comparison with state-of-the-art methods on benchmark functions demonstrates that SBS and its variants are highly competitive, while the combination of the two variants provides the best trade-off between accuracy and computational cost.

1 Introduction

We consider the problem of global optimization of an unknown a priori nonconvex, continuous Sobolev function, under the concern of making efficient use of the computational budget, (i.e. function evaluations at candidate minimizers). Optimizing an unknown function is a typical situation in real applications, e.g. hyperparameter calibration or complex system design emerge in several domains (e.g. [30, 19]). For this, sequential methods are usually employed, where at each iteration the algorithm uses information extracted from the previous candidate solutions to propose new ones. Such methods rely on a deterministic or stochastic process to explore the search space, and on a selection process to choose the next candidate solutions given the previous ones.

In this work, we introduce a new sequential and deterministic particle-based method, called *Stein Boltzmann Sampling* (SBS), for continuous Sobolev functions. SBS uses the *Stein Variational Gradient Descent* (SVGD) [22] method to sample from a target distribution whose mass is concentrated at areas of the domain where minimizers are possible to be found. We choose as target the Boltzmann distribution (BD), which by definition converges toward a distribution with a support spanning over all minimizers of the optimized function. The idea of sampling from the BD for approximating the minimizers of a function is not new (e.g. [2, 4]), yet utilizing SVGD for global optimization is novel, and therefore, part of our contribution concerns the adaptation of the generic SVGD theoretical framework to our objective. SVGD is a generic variational inference method that approximates a target distribution. Specifically, SVGD constructs a flow in the space of probability measures (similarly to a gradient flow evolving in $\mathbb{R}^d$) that moves toward the target distribution. In the discrete case, candidate solutions are represented by particles, and their updates that displace them are affected by attraction and repulsion forces. The SBS optimization process is illustrated in Fig. 1 (some elements will be clarified in Sec. 4), where the sequence of updates over the candidate solutions are shown as trajectories of particles aiming to reach the global minimum. The pseudocode of the proposed SBS method can be found in Alg. 1.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

The related global optimization literature is rich of methods. ADALIPO [25] is a method that is consistent over Lipschitz functions and is adapted for a very low computational budget. The well-known BAYESOPT method [26] is also adapted for low budgets. Then, there are approaches that use MALA to sample from the Boltzmann distribution, in a similar way to our method [12, 39, 31, 8]. CMA-ES [13] and WOA [28], are two inconsistent methods, but are known to be very efficient in practice. Due to either early stopping conditions or time complexity, these two methods do not scale well computationally, hence they are not suited for when the available budget is large and several function evaluations need to be performed. The recent method in [32] subsamples a finite subset of constraints from an uncountable one and uses an SDP solver to approximate the global minimum.

The rest of the contribution of the paper is as follows: we provide a new proof of the SVGD convergence over a compact subset of $\mathbb{R}^d$ for a class of target distributions, which is more general than the one usually considered in the literature, and allows to show the asymptotic convergence of SBS for any continuous Sobolev function (see Sec. 3). In the appendix, we provide detailed definitions and results of the SVGD theory, adapted to the context of global optimization. To ensure the correctness and reproducibility, for some technical results we provide links to proofs in the Lean proof assistant [6, 27]. Then, we introduce two SBS variants: one that uses particle filtering to reduce the budget needed (see Fig. 1b), 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 (see Sec. 4). We discuss the optimal values for the hyperparameters of SBS and compare our approaches with five state-of-the-art methods on several standard global optimization benchmark functions (see Sec. 5 and 6). Finally, we interpret, in the global optimization context, the internal attraction and repulsion forces between particles, which come in effect during the SVGD sampling (see Sec. 7).

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 minimizers of f, i.e. $\forall x^* \in X^*$, $f^* = f(x^*)$. Given an arbitrary function f, its support is $\text{supp}(f) = \{x \in \Omega \mid f(x) \neq 0\}$. Let λ be the standard Lebesgue measure on the Borel sets of $\mathbb{R}^d$. 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 $m, p \in \mathbb{N}$, let $W^{m,p}$ be the Sobolev space of functions with m weak derivatives in $L^p_\mu(\Omega)$:

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

where D^α is the weak derivative operator w.r.t. to the multi-index α, and μ is clear from context. Let the Hilbert space H^m be the Sobolev space $W^{m,2}$.

The full list of notations is provided in the appendix (see Tab. 3).

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: $X_1 \leftarrow \left(x^{(1)}, ..., x^{(N)}\right) \sim \mu_0^{\otimes N}$
for $i = 1$ **to** n **do**
  Compute the vector field $\phi^\star_{\hat{\mu}_i}$ - - see Sec. 2
  $X_{i+1} \leftarrow X_i + \varepsilon \phi^\star_{\hat{\mu}_i}(X_i)$ - - update the particles
  $\hat{\mu}_{i+1} \leftarrow \frac{1}{N} \sum_{j=1}^N \delta_{X_{i+1}^{(j)}}$ - - empirical measure
end for
$\hat{x} \leftarrow \arg\min_{1 \leq j \leq N} f\left(X_{n+1}^{(j)}\right)$ - - the "best" particle
return $\hat{x}$

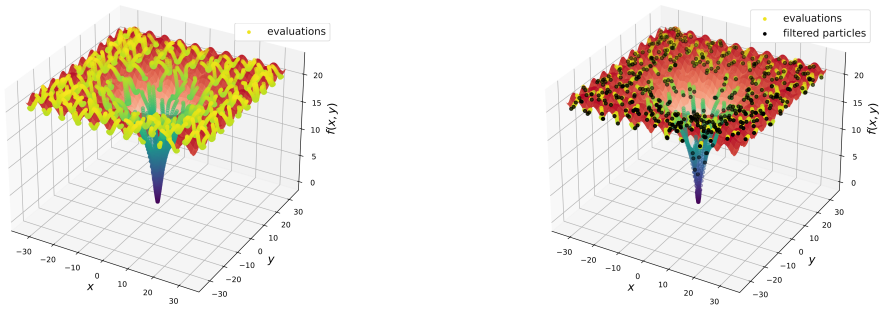
Algorithm 2 Initialization choice of SBS-HYBRID

Input: number of vectors (particles) 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 Gaussian
else
  Use the N candidates from WOA
end if
return the N candidates

2 Stein Boltzmann Sampling

Let us now introduce the proposed *Stein Boltzmann Sampling* (SBS) method. While *Stein Variational Gradient Descent* (SVGD) has been thoroughly studied in the literature [20, 23, 18, 7, 35], this work is the first to consider it in a global optimization context. Therefore, part of the contribution of this work is that adaptation of the SVGD theoretical framework so as to be suitable for global optimization, and to allow addressing more general target distributions over Ω. For consistency and completeness, we prove classical results in our adapted framework in Appendix A.2.



(a) SBS **(b)** SBS-PF

Figure 1: Illustration of the flow of measures and the trajectories of particles over the iterations. The color gradient represents the 2d Ackley function value, from blue (low) to red (high). The trajectories draw the discretized flow of measures. **a)** SBS: the particles are initialized uniformly at random over the domain, and then get updated by making a small step in the direction induced by SVGD forces. **b)** SBS-PF variant with particle filtering: the particles are initialized and updated as before, but the less promising ones get rapidly removed and are not replaced. This is visible as there are less persisting trajectories in areas where the function has high values. This strategy results in a significant reduction of the budget while having comparable performance.

Given the initial particles $(X_0^{(i)})_{1\leq i\leq N} \in \Omega^N$, SVGD constructs an update direction in order to move them toward a target distribution π. This gives the following differential equation for the particles:

$$\frac{\partial X_t^{(i)}}{\partial t} = \frac{1}{N}\sum_{j=1}^{N} \nabla \log \pi\left(X_t^{(j)}\right) k\left(X_t^{(i)}, X_t^{(j)}\right) + \nabla_{X_t^{(j)}} k\left(X_t^{(i)}, X_t^{(j)}\right), \tag{1}$$

where k is the reproducing kernel of a specific RKHS $\mathcal{H}$ (see Appendix A.2 for more details). A usual choice for k is the Gaussian kernel with bandwidth σ. An illustration of this equation is given in Fig. 3. The forces driving the particles are determined by a mixture of individual and collective information. A deep analysis of particle-based models for a large number of particles is exceedingly complex, sometimes even impossible. A popular workaround is to study the convergence of the distribution of the particles at time t, that describes their evolution [20, 18, 29, 5, 10]. For deterministic methods, passing to the distribution is simply an application of the law of large numbers, while for stochastic methods it utilizes tools from the mean-field theory. At time t, the update direction μ_t for the particles distribution is given by:

$$\phi^\star_{\mu_t} \triangleq \int_\Omega \nabla \log \pi(x) k(\cdot, x) + \nabla_x k(\cdot, x)\, \mathrm{d}\mu_t, \tag{2}$$

where the gradient operator is understood in the distributional sense. Moreover, in the classical SVGD literature, the sequence $\mu_{n+1} \triangleq (I_d + \varepsilon \phi^\star_{\mu_n})_{\#}\mu_n$ is also studied (e.g. [22, 18]).

To use SVGD as a global optimization method, we need a target distribution that concentrates its mass around the global minimizers of the optimized function, and the continuous Boltzmann distribution (BD) has this feature. Moreover, it is a classical object in the global optimization theory, and makes a link between our method Simulated Annealing [16] (see Sec. 7).

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

$$m^{(\kappa)}_{f,\Omega}(x) = m^{(\kappa)}(x) = \frac{e^{-\kappa f(x)}}{\int_\Omega e^{-\kappa f(t)}\, \mathrm{d}t}, \quad \forall \kappa \in \mathbb{R}_{\geq 0}. \tag{3}$$

A characteristic property of the BD is that, as κ tends to infinity, the BD tends to a distribution supported only over the set of minimizers X^*. If $\lambda(X^*) > 0$, the BD tends to a uniform distribution over X^* (see Fig. 2a). 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 manifold [14] (see details in Appendix A.1).

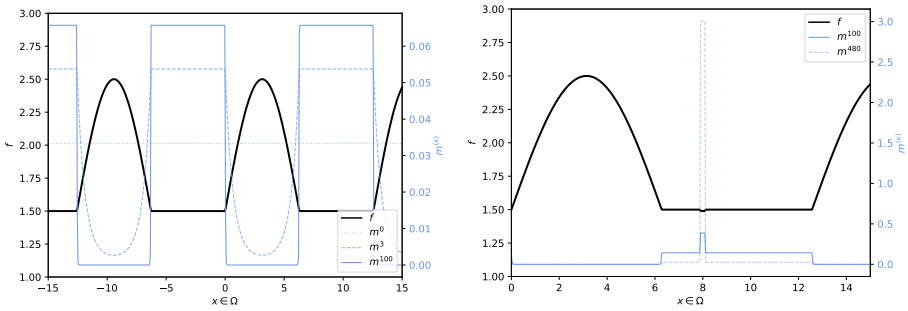
The proposed SBS method is essentially an SVGD sampler applied to the BD. Note that any distribution whose density has the characteristic of being asymptotically supported only over X^* could be used as a target distribution in the following theoretical results.

3 Theory of SBS

First, let $\mathcal{P}_n(\Omega)$ be the set of probability measures on Ω such that for each element $\mu \in \mathcal{P}_n(\Omega)$:

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

where $\mu(\cdot) : \Omega \to \mathbb{R}_{\geq 0}$ is the density of the measure μ w.r.t. λ. To prove the asymptotic convergence of SBS, we need to prove that the measures constructed



(a) **(b)**

Figure 2: a) The density of the Boltzmann distribution $m^{(\kappa)}$ (Definition 2.1) (**blue** lines) becomes uniform over the set of minimizers X^* of the given function f to optimize (**black** lines), as its parameter κ tends to infinity. **b)** In this example, the volume of the set X^* is much smaller than the volume of local minimizers in the flat region. The value of the function at the local minimizers is also closer to the value of the global ones. Setting κ to 100 does not suffice to concentrate the majority of the mass of $m^{(\kappa)}$ around the global minimizers.

using Eq. 2 converges to the distribution induced by the BD, noted as π. To do so, we need to study the net of measures induced by the update direction of SVGD, noted $(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$. To use theoretical results of our adapted 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 $m^{(\kappa)}$ is in $H^1(\Omega)$ (see proof in Appendix B.2). We prove the weak convergence of the net $(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$ to π in the following theorem.

Theorem 3.1 (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 $\phi^\star_{\mu_t}$ (see Eq. 2), such that $T_0 = I_d$. Let $\mu_t = T_{t\#}\mu$. Then, $\mu_t \underset{t}{\rightharpoonup} \pi$.*

The proof is in Appendix B.11 and is inspired by the proof of [23, Theorem 2.8]. It relies on Theorem A.10, a known result of the literature, and Lemmas 3.2 and 3.3, two original lemmas.

Lemma 3.2 (KSD valid discrepancy)**.** *Let μ, $\pi \in \mathcal{P}_2(\Omega)$, and $\mathfrak{K}$ a discrepancy measure defined in Appendix A.2. Then, $\mu = \pi \iff \mathfrak{K}(\mu|\pi) = 0$.*

This result has been stated in [22] without further details. We provide a formalized proof is in Appendix B.9. This lemma implies directly that π is the unique fixed point of the flow of measures constructed by SVGD.

Lemma 3.3 (Unique fixed point)**.** *Let $\pi \in \mathcal{P}_2(\Omega)$ and Φ be the flow of measures induced by the net $(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$ (see in Theorem A.9). 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 lemma. The complete proof is in Appendix B.10.

Asymptotic convergence

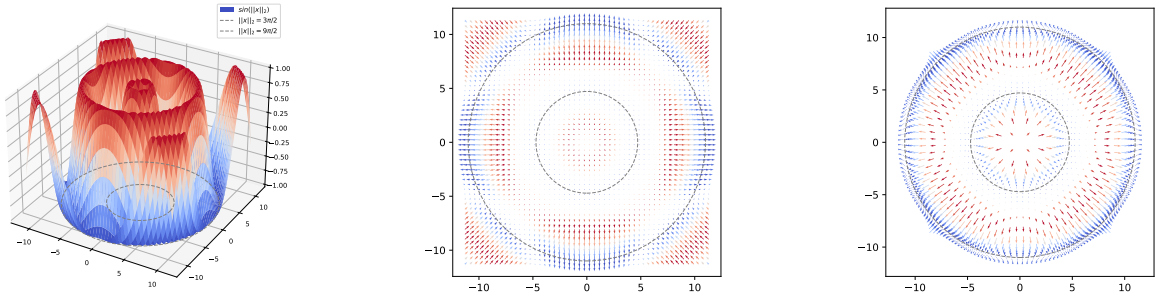
As a direct consequence of Theorem 3.1 and the fact that $\phi^\star_{\mu_t}$ results from the passage to the distribution of particles of SVGD (see Sec. 2), we have the following result.

Theorem 3.4 (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 (Definition 2.1) associated with f and κ. Let $\mu_0 \in \mathcal{P}_2(\Omega)$ and let $\hat{\mu}_i$ be the empirical measure of the particles at iteration i. Then,*

$$\left\{ f\left(X_i\right) \;\middle|\; X_i = (x^{(1)}, \ldots, x^{(N)}) \sim \hat{\mu}_i^{\otimes N} \right\} \underset{\substack{\varepsilon\to 0 \\ i\to\infty}}{\underset{N\to\infty}{\underset{\kappa\to\infty}{\rightharpoonup}}} \{f^*\}.$$

Note that the order of the limits is important. The proof is a direct consequence of the law of large numbers and Theorem 3.1 (that are applicable as $f \in C^0(\Omega) \cap W^{1,4}(\Omega)$), and finally the fact that the BD tends to a distribution supported over the set of minimizers X^* as κ tends to infinity.

To summarize, we proved that SBS is asymptotically convergent for any continuous function belonging to $W^{1,4}(\Omega)$. Note that, since Ω is compact, $C^\infty(\Omega) \subset W^{1,4}(\Omega)$, and therefore, the result holds for any smooth function on Ω. We adapted the SVGD theoretical framework for target distributions that are in $\mathcal{P}_2(\Omega)$ over a compact subset of $\mathbb{R}^d$ (see Appendix A.2). This is different to what is usually considered in the literature, where the target distribution density is smooth and its domain is $\mathbb{R}^d$ (e.g. [22, 21]). Some works have tried to relax the assumptions on the target distribution (e.g. [18, 35]). However, thanks to the compactness of Ω, our assumptions on π are less restrictive and only consider integration constraints on its 1^{st} order weak derivatives, which makes our framework more adapted for global optimization.



(a) $x \mapsto \sin\|x\|_2$ **(b)** $t = 0$ **(c)** $t = 700$

Figure 3: Illustration of the vector field induced by Eq. 1 in a discrete-time setting where π is the BD. **a)** The optimized function $x \mapsto \sin\|x\|_2$ and the two manifolds at which it is minimized (dashed gray lines). **b)** The initial particles (not shown) start getting attracted toward the two ring-shaped manifolds. **c)** After some SVGD iterations, there are stronger forces in the vector field and the particles get concentrated around those minimizing regions.

The implementation of SBS estimates the gradients using finite differences. At each iteration, it updates the set of particles by a small step in the direction induced by $\phi^\star_{\hat{\mu}_i}$, which is computed using the Adam optimizer [15] that gives better experimental results. We choose the initial distribution μ_0 to be the uniform distribution on Ω, as it maximizes the entropy (i.e. high initial exploration), and we also use the RBF kernel function. These two objects are used in most of the SVGD literature. To better understand the previous results and involved objects, we recall some definitions and theoretical results related to SVGD in Appendix A.2.

4 SBS variants

In addition to the main SBS method, we introduce two variants that can be more efficient in practice. The first one, called SBS-PF, uses a particle filtering approach that removes the less promising particles (without replacing them). The second one, called SBS-HYBRID, 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. SBS-PF uses less budget than SBS, and SBS-HYBRID 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.

SBS-PF. We use a simple particle filtering idea: to remove particles (i.e. candidate minimizers of f) that are less promising or stuck in bad local minima. We choose to remove particles that do not move and correspond to significantly higher function values than the others, hence particles that are very likely stuck in bad local minima. This is done by removing particles using their function values and the distance between their previous and actual positions. More precisely, if these two quantities are respectively higher than the q-th and lower than the p-th percentiles of the function values and "previous-to-actual" distances of the particles, then the particle is removed. The difference between SBS and this variant is visualized in Fig. 1. One can see that, in SBS-PF, the least promising candidates are rapidly removed without being 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 optimization results to SBS. Note that the strategy to prove Theorem 3.1 is not directly applicable to SBS-PF, thus, the asymptotic convergence of SBS-PF is not guaranteed. However, the empirical results show that SBS-PF is efficient in practice, and it is a good alternative to SBS when the budget is limited.

SBS-HYBRID. This variant is based on the idea of using SBS or SBS-PF 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 one such method, and then resume the optimization process with an SBS variant. More specifically, we introduce SBS-HYBRID that runs few iterations of both CMA-ES and WOA to choose the most promising result, and then continues the optimization with SBS (see Alg. 2). Both WOA and CMA-ES are efficient methods, thus, a small number of iterations allows to find a good starting region for SBS. Moreover, both methods are not well-suited for a large budget, but for different reasons: CMA-ES uses early stopping rules (e.g. the condition number of the covariance matrix), and WOA takes more time to run than SBS for the same budget. SBS-HYBRID can be seen as a combination of the asymptotically consistent SBS method, on top of very efficient but non-consistent methods. Among the strengths of SBS-HYBRID, we can mention that: i) it empirically provides high-quality results, and ii) it is still asymptotically consistent, since the initial distribution of the particles induced by WOA and CMA-ES meet the assumptions of Theorem 3.1.

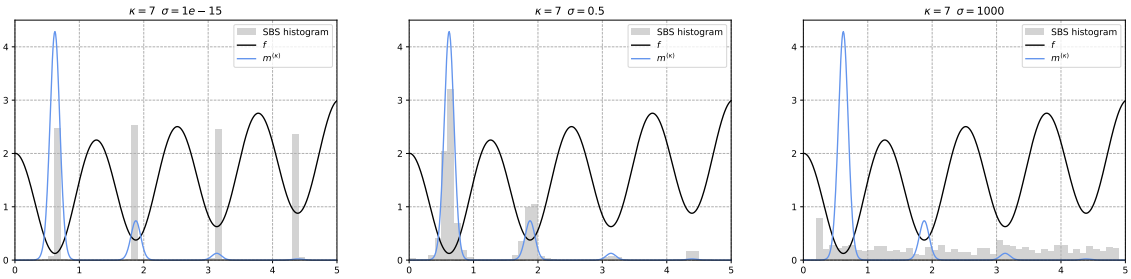


Figure 4: Illustration of the exploration/exploitation trade-off in SBS with different values of σ. In **black**, the function $x \mapsto \cos(5x) + x/5 + 1$; in **grey**, the distribution of the particles; in **blue**, the BD $m^{(\kappa)}$. When σ is too small, the particles are uniformly distributed over X^*. When σ is too large, they are uniformly distributed over the whole domain Ω.

5 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 σ, which carry complex information about the behavior of the method.

Choice of κ. As detailed earlier, κ controls the shape of the BD from which SVGD samples. The bigger κ is, the more the mass of the distribution gets concentrated around the global minimizers of the function. Intuitively, the optimal κ such that a satisfying amount of the mass is around the global minimizers depends on the geometry of the function around local minima (the asymptotic behavior of the BD depends on the local geometry, see [14]). Nevertheless, one can see in Fig. 5a 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 contain most of the density mass are the ones around the global minimizers, then SVGD would succeed in moving some particles in the areas of those modes, provided there are enough particles. The three following parameters control how the mass is repartitioned: the value of κ, the ratio between the volume of the region of local minimizers and the volume of the region of global minimizers, and the value of the function at the local minimizers. These three parameters are interdependent: the value of one can compensate for the value of the others. It is rather unlikely to encounter a function where these three parameters do not compensate each other. It would require the function to have an arbitrary small volume ratio or an arbitrary small distance between the local and global minima (see Fig. 2). Thus, a very large κ, such as 10^3, compensates almost all potential issues related to the geometry of the function and ensures a good performance on average.

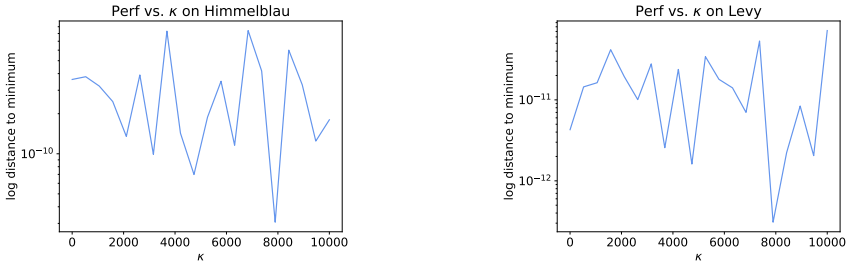
Choice of σ. All SBS variants use the RBF kernel with a bandwidth σ, the choice of which is crucial for their performance. As detailed in Sec. 7, the size of σ controls the forces developed between particles. When a lot of particles are close together, they repel each other. This behavior enforces the exploration of the function domain, but at the same time it prevents SBS from converging at narrow regions where global minimizers could be located. A natural choice is $\sigma = \frac{1}{N^2}$, where N is the number of particles, which ensures that, when the particles are few, σ gets large enough and SBS explores the domain. On the other hand, with a lot a particles the exploration is ensured by the initial uniform distribution μ_0 and the small induced σ allows the particles to converge to the global minimum, even in narrow regions. For the SBS-PF variant, σ changes during the optimization process, as particles are being filtered out. For the SBS-HYBRID variant, as the initial particles are supposed to be already well-positioned and possibly close to the global minimum, σ is set to a very small value, e.g. $\sigma = 10^{-10}$.

6 Experimental evaluation

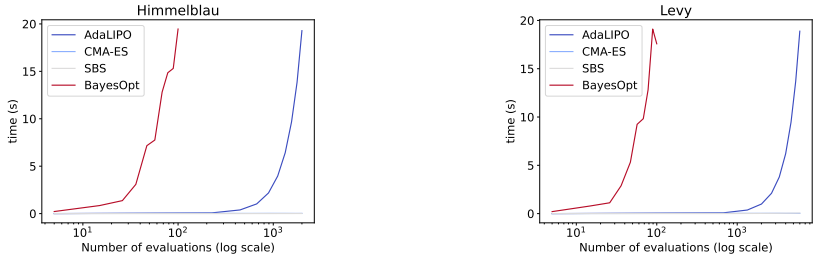
Among the SBS variants, we provide one named SBS-PF-HYBRID that employs SBS-PF as the final step of SBS-HYBRID. In our numerical comparison, we compare the SBS variants with the following state-of-the-art global optimization methods: CMA-ES [13]; WOA [28], a particle-swarm method; ADALIPO [25]; BAYESOPT [26], a similar method to SBS but using MALA instead of SVGD [12, 39, 31]; and CBO [29] that is a stochastic particle-based method similar to SBS (we implemented the algorithm as presented in [29]).

We use classical two dimensional benchmark functions for global optimization. Some are noisy and multimodal (Ackley, Drop wave, Egg Holder, Holder Table, Michalewicz, Rastrigin, Levy), others are smooth (Branin, Goldstein Price, Himmelblau, Rosenbrock, Camel, Sphere) (see more information in [36]). We provide the implementation of this experiment[1]. For the results of Tab. 1, we ran each method 10 times on each function.

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



(a) Log-distance between the computed solution and the global minimum vs. the value of κ for SBS.



(b) Execution time vs. the number of evaluations for BAYESOPT, ADALIPO, CMA-ES, and SBS

Figure 5: Insights for the compared algorithms: **a)** shows the low impact of κ on the performance of SBS. **b)** shows the time to run for BAYESOPT and ADALIPO grows exponentially and is significantly higher than for CMA-ES or SBS. In each case (a) and (b), the **left** plot is for the Himmelblau, and the **right** is for the Levy function.

In the literature, the budget is defined as the number of function evaluations, however, the actual 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 when run on a personal computer[2]. As one can see in Fig. 5b, the running time of ADALIPO and BAYESOPT is significantly higher than for the other methods. For this reason, their budget is set lower than for the other methods: 2K for ADALIPO, 100 for BAYESOPT, and 800K for the rest.

We introduce the following *empirical competitive ratio*:

$$\mathrm{ECR}(m) = \frac{1}{|F|} \sum_{f \in F} \min \left(100, \frac{df_{\mathrm{m}}}{df^*}\right),$$

where F is the set of benchmark functions, df_{m} is the distance between the global minimum of f and the approximation found by the method m, and df^* is the smallest distance among all the methods. ECR provides information about the average precision compared to the best method (lower is better, and the best is 1).

In the results of Tab. 1, one can see that SBS outperforms almost all state-of-the-art methods and scores the fourth rank on average. SBS-PF achieves comparable results on average with significantly less function evaluations ($\sim 97\%$ budget reduction). Moreover, SBS-HYBRID and SBS-PF-HYBRID outperform all the other methods on average. They combine the efficiency of either CMA-ES and WOA with the suitability of SBS for large budgets, while the addition of particle filtering reduces the budget by $\sim 67\%$. In parallel, SBS, SBS-PF-HYBRID, and SBS-HYBRID score respectively the 3^{rd}, 2^{nd}, and 1^{st} rank on the competitive ratio measure, showing that their approximations are precise on average compared to the other methods.

In the appendix, we provide the results of the same experiment on 50 dimensional benchmark functions, by restricting to only methods that can run in low computational time (see Tab. 2). There, the budget is set to 8M. One can observe that SBS and its variants outperform all competitors, and SBS-PF achieves the best results with a budget reduction of $\sim 97\%$, compared to SBS. However, the budget reduction of SBS-PF-HYBRID is less significant ($\sim 16\%$), due to the fact that the high dimensionality of the functions and the initial distribution makes the less promising particles harder to distinguish. Overall, all SBS variants seem robust to function shapes, contrary to CMA-ES for instance, which is very precise on valley-shaped functions but struggles on the multimodal functions.

7 Discussion

Link with Simulated Annealing. The link between SBS and Simulated Annealing [16] is not difficult to see. Indeed, both algorithms are asymptotic methods that sample from the BD. However, the way they sample

[2]The experiments were performed on an Apple M2 chip with 8 cores and 16GB of RAM.

Table 1: Comparison between all SBS variants with several state-of-the-art methods on two dimensional benchmark functions. For each function, we report the average distance to the global minimum and standard deviation (lower is better). The precision is truncated. SBS-HYBRID runs 1K iterations of CMA-ES and WOA. As one can see, SBS-HYBRID and SBS respectively rank 1st and 4th 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 METHOD AND VARIANTS			
FUNCTIONS	LANGEVIN	BAYESOPT	CBO	ADALIPO	CMA-ES	WOA	SBS-PF	SBS	SBS-PF-HYBRID	SBS-HYBRID
ACKLEY	6.8 ± 1.7	0.1 ± 0.1	0.0 ± 0.0	1.3 ± 0.7	19.8 ± 0.0	$\mathbf{8e^{-8} \pm 5e^{-8}}$	$0.0 \pm 9e^{-4}$	$8e^{-4} \pm 3e^{-4}$	$1e^{-5} \pm 1e^{-5}$	$7e^{-6} \pm 3e^{-6}$
BRANIN	$8e^{-5} \pm 1e^{-4}$	$2e^{-4} \pm 1e^{-4}$	0.4 ± 0.3	0.0 ± 0.0	$3e^{-7} \pm 0$	$1e^{-6} \pm 1e^{-6}$	$5e^{-7} \pm 2e^{-7}$	$3e^{-7} \pm 3e^{-11}$	$\mathbf{3e^{-7} \pm 0}$	$3e^{-7} \pm 5e^{-16}$
DROP WAVE	$0.9 \pm 1e^{-16}$	0.2 ± 0.1	0.1 ± 0.1	0.1 ± 0.0	0.3 ± 0.3	$\mathbf{1e^{-15} \pm 1e^{-15}}$	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.0	0.0 ± 0.0
EGG HOLDER	$2008.8 \pm 2e^{-13}$	90.3 ± 60.6	$833.5 \pm 1e^{-10}$	40.0 ± 18.7	$393.6 \pm 5e^{-14}$	$\mathbf{3e^{-5} \pm 5e^{-10}}$	18.0 ± 19.0	8.0 ± 10.2	7.8 ± 9.6	21.5 ± 32.0
GOLDSTEIN PRICE	391387.4 ± 301031.1	8.3 ± 6.6	$4e^{-4} \pm 2e^{-4}$	0.4 ± 0.3	8.1 ± 24.3	$1e^{-6} \pm 7e^{-7}$	$6e^{-7} \pm 5e^{-7}$	$1e^{-9} \pm 2e^{-9}$	$\mathbf{6e^{-14} \pm 6e^{-15}}$	$6e^{-13} \pm 1e^{-12}$
HIMMELBLAU	47.5 ± 31.0	$7e^{-4} \pm 9e^{-4}$	0.1 ± 0.1	0.0 ± 0.0	$9e^{-16} \pm 1e^{-15}$	$1e^{-6} \pm 1e^{-6}$	$1e^{-7} \pm 3e^{-7}$	$5e^{-11} \pm 4e^{-11}$	$4e^{-19} \pm 7e^{-19}$	$\mathbf{1e^{-20} \pm 1e^{-20}}$
HOLDER TABLE	3.4 ± 0.5	0.4 ± 0.9	$0.0 \pm 7e^{-4}$	0.0 ± 0.0	5.0 ± 5.0	$\mathbf{2e^{-6} \pm 2e^{-6}}$	$2e^{-6} \pm 1e^{-7}$	$2e^{-6} \pm 2e^{-9}$	$2e^{-6} \pm 1e^{-10}$	$2e^{-6} \pm 1e^{-10}$
MICHALEWICZ	9.6 ± 0.2	$7.9 \pm 1e^{-5}$	8.6 ± 0.0	7.9 ± 0.0	8.0 ± 0.3	$7.9 \pm 1e^{-6}$	$7.9 \pm 1e^{-6}$	$7.9 \pm 1e^{-10}$	$7.9 \pm 8e^{-14}$	$\mathbf{7.9 \pm 4e^{-15}}$
RASTRIGIN	30.5 ± 3.7	2.2 ± 1.2	$9e^{-4} \pm 7e^{-4}$	0.2 ± 0.2	5.4 ± 5.8	$\mathbf{6e^{-15} \pm 7e^{-15}}$	$5e^{-6} \pm 3e^{-6}$	$5e^{-9} \pm 1e^{-8}$	0.5 ± 0.7	0.3 ± 0.6
ROSENBROCK	6852.4 ± 4943.7	0.2 ± 0.2	0.0 ± 0.0	0.1 ± 0.0	$4e^{-16} \pm 7e^{-16}$	$2e^{-7} \pm 1e^{-7}$	$6e^{-5} \pm 9e^{-5}$	$2e^{-6} \pm 4e^{-6}$	$5e^{-17} \pm 8e^{-17}$	$\mathbf{1e^{-17} \pm 1e^{-17}}$
CAMEL	397.9 ± 9.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	$2e^{-5} \pm 1e^{-14}$	$2e^{-5} \pm 2e^{-8}$	$\mathbf{2e^{-5} \pm 1e^{-7}}$	$2e^{-5} \pm 1e^{-11}$	$2e^{-5} \pm 0$	$2e^{-5} \pm 1e^{-16}$
LEVY	83.5 ± 12.9	0.1 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	1.0 ± 2.1	$8e^{-9} \pm 7e^{-9}$	$9e^{-8} \pm 8e^{-8}$	$1e^{-12} \pm 2e^{-12}$	$3e^{-19} \pm 8e^{-19}$	$\mathbf{9e^{-20} \pm 2e^{-19}}$
SPHERE	$9e^{-5} \pm 6e^{-5}$	$5e^{-4} \pm 5e^{-4}$	0.0 ± 0.0	$0.0 \pm 9e^{-4}$	$5e^{-16} \pm 1e^{-15}$	$1e^{-16} \pm 8e^{-17}$	$5e^{-8} \pm 7e^{-8}$	$6e^{-12} \pm 7e^{-12}$	$1e^{-19} \pm 2e^{-19}$	$\mathbf{1e^{-21} \pm 3e^{-21}}$
ECR	62.2	46.8	46.7	46.7	24.0	22.4	46.7	20.8	16.2	**15.4**
AVERAGE RANK	9.38	7.85	7.46	7.15	6.69	3.00	4.15	3.38	3.15	**2.77**
FINAL RANK	10	9	8	7	6	2	5	4	3	**1**

from that distribution is different. Simulated Annealing is a Markov Chain Monte-Carlo method [2], while SBS is a deterministic variational approach. The minimum temperature parameter of Simulated Annealing is the inverse of the κ parameter of SBS. Thus, any scheduler for Simulated Annealing's temperature can also be used in SBS. However, there is an extra degree of exploration/exploitation in SBS, brought by the kernel bandwidth employed by the SVGD sampling.

Locality of the kernel. In classical SVGD implementations, the RBF kernel used 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, expressed as:

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

The first term attracts remote particles to a close cluster of particles, and the second term repels particles that are too close to each other. Hence, they are respectively exploitation and exploration forces. Indeed, the attraction allows particles to "fall" in local minima, wherein a lot of particles are already stuck. The repulsion prevents particles from getting stuck together at a narrow region of the search space, and forces them to explore the space. 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 even from a very far distance. An arbitrary large σ leads to a uniform discretization of the space. These behaviors are illustrated in Fig. 4. In the case of SBS, the value of σ is a user parameter.

Some weaknesses. Because of the gradient approximation by finite difference that occurs in SBS, our methods require a large budget. That is a common issue in gradient-based optimization algorithms. However, in the contrary of more frugal approches (e.g. BAYESOPT), SBS and its variants have a way smaller execution time. Another weakness of SBS is the difficult choice of the kernel. As explained above, the kernel controls the particles movements and the performance of one specific kernel choice highly depends on the geometry of the objective function. This choice is crucial and future users should tune this hyperparameter carefully. A way to mitigate this problem would be to find an adaptive kernel that uses only evaluations

of the objective function to choose the best way for the particles to interact. However, we believe that is a quite complex subject that is out of the scope of the current study, and should be investigated in a more general point of view.

8 Conclusion

In this paper, we introduced the Stein Boltzmann Sampling (SBS) method for global optimization, along some variants. We proved that it is asymptotically consistent using the theory of the SVGD algorithm that we extended to a more general class of target distributions, 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 1^{st} order weak derivatives. We showed in our experimental evaluation that SBS outperforms state-of-the-art methods on average on classical benchmark functions, that SBS-PF can lead to drastic reduction of the needed computational budget while having comparable performance than the original SBS version, and that SBS-HYBRID outperforms all the other methods in practice. This work suggests that, for obtaining the best trade-off between accurate approximations and low budget, SBS should be used as a continuation for others particles or distribution-based methods, conjointly with particles filtering strategies (SBS-PF-HYBRID). As future, the convergence rate of SBS and its components can be further studied, and more sophisticated particle filtering strategies can be designed to make it more appealing for global optimization in real-world applications.

Acknowledgments

The authors acknowledge the support from the Industrial Data Analytics and Machine Learning Chair hosted at ENS Paris-Saclay. Also, we sincerely thank the anonymous reviewers for their valuable comments and suggestions, which have greatly helped improve the quality of this paper.

References

[1] Nachman Aronszajn. *Theory of Reproducing Kernels.* Transactions of the American Mathematical Society, 1950.

[2] Robert Azencott. Simulated annealing, 1989.

[3] Patrick Billingsley. *Convergence of Probability Measures.* Wiley, 1999.

[4] Valentin De Bortoli and Agnès Desolneux. On quantitative laplace-type convergence results for some exponential probability measures, with two applications, 2021.

[5] José A Carrillo, Young-Pil Choi, Claudia Totzeck, and Oliver Tse. An analytical framework for consensus-based global optimization method. *Mathematical Models and Methods in Applied Sciences*, 2018.

[6] Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*. Springer International Publishing, 2021.

[7] Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *Journal of Machine Learning Research*, 2023.

[8] Murat A Erdogdu, Lester Mackey, and Ohad Shamir. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

[9] Lawrence Craig Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions, Revised Edition*. Chapman and Hall/CRC, 2015.

[10] Massimo Fornasier, Timo Klock, and Konstantin Riedl. Consensus-based optimization methods converge globally in mean-field law. *arXiv preprint arXiv:2103.15130*, 2021.

[11] Jackson Gorham and Lester Mackey. Measuring sample quality with stein's method. *Proceedings of Advances in Neural Information Processing Systems*, 2015.

[12] Ulf Grenander and Michael I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1994.

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

[14] Chii-Ruey Hwang. Laplace's Method Revisited: Weak Convergence of Probability Measures. *The Annals of Probability*, pages 1177–1182, 1980.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

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

[17] Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel stein discrepancy descent. In *Proceedings of the International Conference on Machine Learning*, 2021.

[18] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for stein variational gradient descent. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

[19] Juyong Lee, In-Ho Lee, InSuk Joung, Jooyoung Lee, and Bernard R. Brooks. Finding multiple reaction pathways via global optimization of action. *Nature Communications*, 2017.

[20] Qiang Liu. Stein variational gradient descent as gradient flow. *Proceedings of Advances in Neural Information Processing Systems*, 2017.

[21] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, 2016.

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

[23] Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 2019.

[24] Xiaopeng Luo. Minima distribution for global optimization, 2019.

[25] Cedric Malherbe and Nicolas Vayatis. Global optimization of Lipschitz functions. In *Proceedings of the International Conference on Machine Learning*, 2017.

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

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

[28] Seyedali Mirjalili and Andrew Lewis. The whale optimization algorithm. *Advances in engineering software*, 2016.

[29] René Pinnau, Claudia Totzeck, Oliver Tse, and Stephan Martin. A consensus-based model for global optimization and its mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 2017.

[30] János D Pintér. Global optimization in action. *Scientific American*, 1991.

[31] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, 2017.

[32] Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *Mathematical Programming*, pages 1–82, 2024.

[33] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 2011.

[34] Charles Stein. 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.

[35] Lukang Sun, Avetik Karagulyan, and Peter Richtarik. Convergence of stein variational gradient descent under a weaker smoothness condition. In *International Conference on Artificial Intelligence and Statistics*, 2023.

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

[37] Cédric Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

[38] Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009.

[39] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.

[40] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 2008.

Table 2: Comparison between all SBS variants with several state-of-the-art methods on 50 dimensional benchmark functions. For each function, we report the average distance to the global minimum (lower is better). The precision is truncated. SBS-HYBRID runs 1K iterations of CMA-ES and WOA. As one can see, SBS-PF ranks 1^s, as SBS, while having a significant budget reduction ($\sim 97\%$). Moreover, SBS-PF-HYBRID outperforms SBS-HYBRID while having a budget reduction of $\sim 16\%$. The high dimensionality of the functions makes the particle filtering of SBS-PF-HYBRID less efficient.

	STATE-OF-THE-ART				PROPOSED METHODS			
FUNCTIONS	LANGEVIN	CBO	WOA	CMA-ES	SBS-HYBRID	SBS-PF-HYBRID	SBS	SBS-PF
ACKLEY	21.5 ±0.0	21.1 ±0.1	19.8 ±0.1	19.6 ±0.0	19.5 ±0.1	19.6 ±0.1	**19.0** **±0.1**	19.0 ±0.1
MICHALEWICZ	8.6 ±0.6	**0.8** **±0.5**	4.4 ±0.2	25.8 ±2.4	24.2 ±1.8	24.1 ±2.5	3.8 ±0.4	2.4 ±0.8
RASTRIGIN	884.4 ±202.5	841.1 ±302.9	593.8 ±22.2	**107.4** **±20.9**	114.5 ±24.9	115.0 ±10.9	280.5 ±17.2	280.5 ±17.2
ROSENBROCK	367368.5 ±22139.8	176596.0 ±0	20145.3 ±3609.2	**0.4** **±1.2**	28.9 ±1.3	28.7 ±2.0	39.5 ±1.4	25.3 ±4.0
LEVY	2931.7 ±88.4	349.8 ±38.3	224.5 ±13.4	75.3 ±27.0	81.1 ±19.6	70.4 ±23.2	**55.9** **±2.6**	57.8 ±4.4
SPHERE	0.0 $\pm 3e^{-4}$	5000.0 ±0	646.3 ±31.5	$1e^{-14}$ $\pm 2e^{-15}$	$\mathbf{7e^{-20}}$ $\mathbf{\pm 1e^{-20}}$	$2e^{-19}$ $\pm 2e^{-19}$	$2e^{-10}$ $\pm 1e^{-11}$	$1e^{-4}$ $\pm 5e^{-6}$
ECR	43.9	36.0	35.3	**1.5**	13.4	13.2	18.1	28.2
AVERAGE RANK	7.17	6.17	5.83	3.67	3.67	3.50	**3.00**	**3.00**
FINAL RANK	7	6	5	4	3	2	**1**	**1**

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 distribution, 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
$(\mu_i)_{i\in\mathbb{N}}$	sequence of measures constructed by SVGD
$\hat{\mu}_i$	empirical measure of the SVGD particles
$(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$	net extension of $(\mu_i)_{i\in\mathbb{N}}$
$\Phi : \mathbb{R}_{\geq 0} \times \mathcal{P}_2(\Omega)$	the flow of measures associated to $(\mu_t)_{t\in\mathbb{R}_{\geq 0}}$

A Theoretical foundations

In this section, we introduce fundamental results related to the Boltzmann distribution (BD) and the SVGD theory. Please note that, concerning the BD, those results are not new. Concerning SVGD, we adapt its generic theoretical framework, allowing more general target distributions. We prove classical results of SVGD theory in this novel framework. 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 adapted 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 [24]. 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 Fig. 2a. 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 Eq. 3. 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* [22] is a method that transforms iteratively an arbitrary measure μ to a target distribution 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* [34]. The theory of SVGD has been developed in several works over the years. Note that recently, [17] 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 distribution density in $H^1(\Omega)$: a adapted framework particularly suitable for global optimization that we use to prove the consistency of SBS (see Sec. 3).

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 a 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 [20]:

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

where (∇) and $(\nabla\cdot)$ are respectively the gradient and the divergence operators, in the distributional sense. 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 [21])**.** 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 [34])**.** *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 $\mu \neq \pi$, Eq. 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* [11]. 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, [21] chose $\mathcal{F}$ to be a universal vector-valued RKHS, which allows to find closed-form solution to Eq. 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 distribution. Next, we define the vector-valued RKHS that will be used in the Kernelized Stein Discrepancy.

Definition A.4 (Product RKHS [22])**.** 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 distributional sense). Using the Moore–Aronszajn theorem [1], we consider the associated real-valued RKHS $\mathcal{H}_0$. Let $\mathcal{H}$ be the product RKHS induced by $\mathcal{H}_0$, i.e. $\forall f = (f_1, \ldots, f_d)^\top$, $f \in \mathcal{H} \iff \forall 1 \leq i \leq d, f_i \in \mathcal{H}_0$. The inner product of $\mathcal{H}$ is defined by

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

For more details, see Lean formalization [3].

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

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 [33]), 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:

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

We can now define the KSD.

Definition A.5 (Kernelized Stein Discrepancy [21])**.** 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 [21])**.** *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* $\mathrm{supp}(\pi) = \Omega$, $\phi^\star_\mu$ *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 [22])**.** *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*

$$\begin{aligned} T_\varepsilon : (\Omega \to \Omega) &\to \Omega \\ \phi &\mapsto I_d + \varepsilon\phi. \end{aligned}$$

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:

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

Furthermore, 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. 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 order to use results in Sec. 3, we need to prove that $\phi^\star_\mu \in \mathcal{S}(\mu)$. Given the assumptions of the kernel, $\phi^\star_\mu$ lies in $H^1(\Omega)$. Moreover, as $\phi^\star_\mu \in \mathcal{H}$ and $k(\cdot, x) \in \mathcal{S}(\mu)$, [21, Proposition 3.5] gives the rest of the proof. This allows to use Lemma A.3 with $\phi^\star_\mu$, for any $\mu \in \mathcal{P}_2(\Omega)$. We can now study the time-derivative of the measure net and the KL-divergence between μ and π.

Theorem A.9 (Time derivative of a measure flow [20])**.** *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:*

$$\begin{cases} \frac{\partial \mu_t}{\partial t} = -\nabla \cdot (\phi_t \mu_t), \forall t > 0 \\ \mu_0 = \mu \end{cases} \tag{7}$$

where $(\nabla\cdot)$ is the divergence operator, in the distributional sense (see details in Appendix B.7). Moreover, the sequence $(\mu_i)_{i\in\mathbb{N}}$, constructed by SVGD, is a discretized solution of Eq. 7, 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 give a new proof of this theorem in Appendix B.7, using optimal transport theory. That proof is more general in T, but less constructive. We also prove that the sequence $(\mu_i)_{i\in\mathbb{N}}$ is a discretized solution of Eq. 7 (note that Eq. 7 has also been extensively studied in [23]). This result allows to study the time-derivative of the KL-divergence between μ_t and π.

Theorem A.10 (Time-derivative of the KL-divergence [20])**.** *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*

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

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

The proof is in Appendix B.8.

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 [4]. Note that a collection of all key notations and their meanings is available in Tab. 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 $\overline{\forall}$. This quantifier can be found in Mathlib (the mathematics library of Lean), noted $\forall^m\ x\ \partial\mu, P\ x$.

B.1 Proof of Properties A.1

The continuous BD is a special case of the nascent minima distribution, introduced in [24], that has the generic form

$$m^{(\kappa)}_{f,\Omega}(x) = m^{(\kappa)}(x) = \frac{\tau^\kappa(f(x))}{\int_\Omega \tau^\kappa(f(t))\,\mathrm{d}t}, \tag{8}$$

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 Eq. 8. 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, $\forall x \in \Omega$,*

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

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,

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

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

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

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

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

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

■

Theorem 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 $L^2(\Omega)$:

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

■

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 [9]):

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

■

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 $(\lambda_i)_{i\in\mathbb{N}}$ of T_μ 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

$$\forall f, g \in \mathcal{H}_k, \langle f, g \rangle_{\mathcal{H}_k} = \left\langle \sum_{i=1}^{\infty} \lambda_i a_i \phi_i, \sum_{i=1}^{\infty} \lambda_i b_i \phi_i \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \lambda_i a_i b_i. \tag{9}$$

Routine works show that Eq. 9 defines a inner product and therefore that $\mathcal{H}_k$ is a Hilbert space (for more details, see Lean proof [5]). 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,

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

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

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 equality, that

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

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 $T_\mu : L^2_\mu(\Omega) \to \mathcal{H}_0$. ■

B.5 Proof of Theorem A.6

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

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

As $\mathrm{supp}(\mu) = \Omega$, $f^{(i)}$ 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

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

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

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

$$
\begin{aligned}
\langle f, \phi_\mu^\star \rangle_{\mathcal{H}} &= \sum_{\ell=1}^{d} \left\langle f^{(\ell)}, \mathbb{E}_{x\sim\mu}\left[\nabla \log \pi^{(\ell)}(x) k(x\cdot) + \nabla_x k^{(\ell)}(x,\cdot)\right] \right\rangle_{\mathcal{H}_0} \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{\ell=1}^{d} \langle f^{(\ell)}, \nabla \log \pi^{(\ell)}(x) k(\cdot,x) + \nabla_x k^{(\ell)}(x,\cdot)\rangle_{\mathcal{H}_0}\right] \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{\ell=1}^{d} \nabla \log \pi^{(\ell)}(x) \langle f^{(\ell)}, k(\cdot,x)\rangle_{\mathcal{H}_0} + \langle f^{(\ell)}, \nabla_x k^{(\ell)}(x,\cdot)\rangle_{\mathcal{H}_0}\right] \\
&= \mathbb{E}_{x\sim\mu}\left[\sum_{\ell=1}^{d} \nabla \log \pi^{(\ell)}(x) f^{(\ell)}(x) + \frac{\partial f^{(\ell)}(x)}{\partial x_\ell}\right] \quad [40] \\
&= \mathbb{E}_{x\sim\mu}\left[\nabla \log \pi(x)^\top f(x) + \nabla \cdot f(x)\right].
\end{aligned}
$$

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[6]. 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 $\nabla_\varepsilon \log\left(\pi_{[T^{-1}]}(x)\right)$.

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

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

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

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. ■

B.7 Proof of Theorem A.9

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 $(T_t)_{0\le t}$ 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 [37, Theorem 5.34] 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.

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

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

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 [38]). Furthermore, as $\mu_{i+1} = (I_d + \varepsilon \phi^\star_{\mu_i})_{\#}\mu_i$ (see Eq. 6), one can write

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

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

B.8 Proof of Theorem A.10

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 A.9, 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 3.2

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, $\mu(x)=c\pi(x)$. As $\mu(\cdot)$ and $\pi(\cdot)$ are probability densities over Ω, $c=1$:

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

Thus,

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

For more details, see Lean proof [7]. ■

B.10 Proof of Lemma 3.3

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 3.2, 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 3.2, we obtain a contradiction.

For more details, see Lean proof[7]. ■

B.11 Proof of Theorem 3.1

Proof. By construction of $\mathcal{P}_2(\Omega)$, $\mathrm{KL}(\mu||\pi)$ is finite. Moreover, as stated in Theorem A.10, $t \mapsto \mathrm{KL}(\mu_t||\pi)$ 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 [3, Theorem 2.6]). Therefore, by continuity of $\mathfrak{K}(\cdot|\pi)$, μ_∞ is a fixed point of Φ_t, for any $\mu \in \mathcal{P}_2(\Omega)$ such that $\mathrm{KL}(\mu||\pi)$ is finite. Finally, using Lemma 3.3, we have that $\mu_\infty = \pi$. ■

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