Kernel Stein Discrepancy Descent

Anna Korba [1] **Pierre-Cyril Aubin-Frankowski** [2] **Szymon Majewski** [3] **Pierre Ablin** [4]

Abstract

Among dissimilarities between probability distributions, the Kernel Stein Discrepancy (KSD) has received much interest recently. We investigate the properties of its Wasserstein gradient flow to approximate a target probability distribution π on $\mathbb{R}^d$, known up to a normalization constant. This leads to a straightforwardly implementable, deterministic score-based method to sample from π, named KSD Descent, which uses a set of particles to approximate π. Remarkably, owing to a tractable loss function, KSD Descent can leverage robust parameter-free optimization schemes such as L-BFGS; this contrasts with other popular particle-based schemes such as the Stein Variational Gradient Descent algorithm. We study the convergence properties of KSD Descent and demonstrate its practical relevance. However, we also highlight failure cases by showing that the algorithm can get stuck in spurious local minima.

1. Introduction

An important problem in machine learning and computational statistics is to sample from an intractable target distribution π. In Bayesian inference for instance, π corresponds to the posterior probability of the parameters, which is required to compute the posterior predictive distribution. It is known only up to an intractable normalization constant. In Generative Adversarial Networks (GANs, Goodfellow et al., 2014), the goal is to generate data which distribution is similar to the training set defined by samples of π. In the first setting, one has access to the score of π (the gradient of its log density), while in the second, one has access to samples of π. Assessing how different the target π and a given approximation μ are can be performed through a dissimilarity function $D(\mu|\pi)$. As summarized by Simon-Gabriel (2018), classical dissimilarities include f-divergences such as the KL (Kullback-Leibler) or the χ^2 (Chi-squared), the Wasserstein distances in Optimal Transport (OT), and Integral Probability Metrics (IPMs), such as the Maximum Mean Discrepancy (MMD, Gretton et al., 2012). Dissimilarity functions can hence be used to characterize π since, under mild assumptions, they only vanish at $\mu = \pi$. Setting $\mathcal{F}(\mu) = D(\mu|\pi)$, assuming also in our case that $\pi \in \mathcal{P}_2(\mathbb{R}^d)$, the set of probability measures μ with finite second moment ($\int \|x\|^2 \mathrm{d}\mu(x) < \infty$), the sampling task can then be recast as an optimization problem over $\mathcal{P}_2(\mathbb{R}^d)$

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu). \tag{1}$$

Starting from an initial distribution μ_0, one can then apply a descent scheme to (1) to converge to π. In particular, one can consider the Wasserstein gradient flow of $\mathcal{F}$ over $\mathcal{P}_2(\mathbb{R}^d)$. This operation can be interpreted as a vector field continuously displacing the particles constituting μ.

Among dissimilarities, the Kernel Stein Discrepancy (KSD, introduced independently by Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017) writes as follows

$$\mathrm{KSD}(\mu|\pi) = \sqrt{\iint k_\pi(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y)}, \tag{2}$$

where k_π is the Stein kernel, defined through the score of π, $s(x) = \nabla \log \pi(x)$, and through a positive semi-definite kernel k (see Section 2.1 for the meaning of $\boldsymbol{\nabla}\cdot_1$ or of ∇_2)

$$k_\pi(x,y) = s(x)^T s(y) k(x,y) + s(x)^T \nabla_2 k(x,y) + \nabla_1 k(x,y)^T s(y) + \boldsymbol{\nabla} \cdot_1 \nabla_2 k(x,y). \tag{3}$$

The great advantage of the KSD is that it can be readily computed when one has access to the score of π and uses a discrete measure $\hat{\mu}$, since (2) writes as a finite double sum of k_π in this case. Furthermore the definition of the KSD was inspired by Stein's method (see Anastasiou et al., 2021, for a review) and no sampling over π is required in (2). This motivated the use of the KSD in a growing number of problems. The KSD has been widely used in nonparametric statistical tests for goodness-of-fit (e.g. Xu & Matsuda, 2020; Kanagawa et al., 2020). It was also used for sampling tasks: to select a suitable set of static points to approximate π, adding a new one at each iteration (Chen et al., 2018; 2019); to compress (Riabiz et al., 2020) or reweight (Hodgkinson et al.,

[1]CREST, ENSAE, Institut Polytechnique de Paris [2]CAS, MINES ParisTech, Paris, France [3]CMAP, Ecole Polytechnique, Institut Polytechnique de Paris [4]CNRS and DMA, Ecole Normale Supérieure, Paris, France. Correspondence to: Anna Korba <anna.korba@ensae.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning,PMLR 139, 2021.

2020) Markov Chain Monte Carlo (MCMC) outputs; and to learn a static transport map from μ_0 to π (Fisher et al., 2021). In this paper, we consider $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{KSD}^2(\mu|\pi)$ and its Wasserstein gradient $\nabla_{W_2}\mathcal{F}$ to define a flow over particles to approximate π.

Related works. Minimizing a dissimilarity D is a popular approach to fit an unnormalized density model in the machine learning and computational statistics literature. For instance, Hyvärinen & Dayan (2005) proposed to minimize the Fisher divergence. Alternatively, D is often taken as the KL divergence. Indeed, since the seminal paper by Jordan et al. (1998), the Wasserstein gradient flow of the KL has been extensively studied and related to the Langevin Monte Carlo (LMC) algorithm (e.g. Wibisono, 2018; Durmus et al., 2019). However, an unbiased time-discretization of the KL flow is hard to implement (Salim et al., 2020). To tackle this point, a recent successful kernel-based approximation of the KL flow was introduced as the Stein Variational Gradient Descent (SVGD, Liu & Wang, 2016). Several variants were considered (see Chewi et al., 2020, and references therein). Another line of work considers D as the MMD (Mroueh et al., 2019; Arbel et al., 2019) with either regularized or exact Wasserstein gradient flow of the MMD, especially for GAN training. However, these two approaches require samples of π to evaluate the gradient of the MMD. As the KSD is a specific case of the MMD with the Stein kernel (Chen et al., 2018), our approach is similar to Arbel et al. (2019) but better suited for a score-based sampling task, owing to the properties of the Stein kernel.

Contributions. In this paper, in contrast with the aforementioned approaches, we choose the dissimilarity D in (1) to be the KSD. As in SVGD, our approach, KSD Descent, optimizes the positions of a finite set of particles to approximate π, but through a descent scheme of the KSD (in contrast to the KL for SVGD) in the space of probability measures. KSD Descent comes with several advantages. First, it benefits from a closed-form cost function which can be optimized with a fast and hyperparameter-free algorithm such as L-BFGS (Liu & Nocedal, 1989). Second, our analysis comes with several theoretical guarantees (namely, existence of the flow and a descent lemma in discrete time) under a Lipschitz assumption on the gradient of k_π. We also provide negative results highlighting some weaknesses of the convergence of the KSD gradient flow, such as the absence of exponential decay near equilibrium. Moreover, stationary points of the KSD flow may differ from the target π and even be local minima of the flow, which implies that some particles are stuck far from high-probability regions of π. Sometimes a simple annealing strategy mitigates this convergence issue. On practical machine learning problems, the performance of KSD Descent highly depends on the local minimas of $\log \pi$. KSD Descent achieves comparable performance to SVGD on convex (i.e., π log-concave) toy examples and Bayesian inference tasks, while it is outperformed on non-convex tasks with several saddle-points like independent component analysis.

This paper is organized as follows. Section 2 introduces the necessary background on optimal transport and on the Kernel Stein Discrepancy. Section 3 presents our approach and discusses its connections with related works. Section 4 is devoted to the theoretical analysis of KSD Descent. Our numerical results are to be found in Section 5.

2. Background

This section introduces the high-level idea of the gradient flow approach to sampling. It also summarizes the known properties of the KSD.

2.1. Notations

The space of l continuously differentiable functions on $\mathbb{R}^d$ is $C^l(\mathbb{R}^d)$. The space of smooth functions with compact support is $C_c^\infty(\mathbb{R}^d)$. If $\psi : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable, we denote by $J\psi : \mathbb{R}^d \to \mathbb{R}^{p\times d}$ its Jacobian. If $p = 1$, we denote by $\nabla\psi$ the gradient of ψ. Moreover, if $\nabla\psi$ is differentiable, the Jacobian of $\nabla\psi$ is the Hessian of ψ denoted $H\psi$. If $p = d$, $\boldsymbol{\nabla} \cdot \psi$ denotes the divergence of ψ, i.e. the trace of the Jacobian. We also denote by $\Delta\psi$ the Laplacian of ψ, where $\Delta\psi = \boldsymbol{\nabla} \cdot \nabla\psi$. For a differentiable kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\nabla_1 k$ (resp. $\nabla_2 k$) is the gradient of the kernel w.r.t. the first (resp. second) variable, while $H_1 k$ denotes its Hessian w.r.t. the first variable.

Consider the set $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures μ on $\mathbb{R}^d$ with finite second moment and $\mathcal{P}_c\left(\mathbb{R}^d\right)$ the set of probability measures with compact support. For $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $\mathrm{d}\mu/\mathrm{d}\pi$ its Radon-Nikodym density if μ is absolutely continuous w.r.t. π. For any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $L^2(\mu)$ is the space of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\int \|f\|^2 d\mu < \infty$. We denote by $\|\cdot\|_{L^2(\mu)}$ and $\langle\cdot,\cdot\rangle_{L^2(\mu)}$ respectively the norm and the inner product of the Hilbert space $L^2(\mu)$. Given a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T_{\#}\mu$ is the pushforward measure of μ by T. We consider, for $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance $W_2(\mu, \nu)$, and we refer to the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ as the Wasserstein space. In a Riemannian interpretation of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at μ is denoted $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ and is a subset of $L^2(\mu)$ (Otto, 2001). We refer to Appendix A.2 for more details on the Wasserstein distance and related flows.

2.2. Lyapunov analysis and gradient flows

To sample from a target distribution π, a now classical approach consists in identifying a continuous process which moves particles from an initial probability distribution μ_0 toward samples of π. This can be expressed as searching for

vector fields $v_t : \mathbb{R}^d \to \mathbb{R}^d$ transporting the distribution μ_t through the continuity equation (see Appendix A.1)

$$\frac{\partial \mu_t}{\partial t} + \operatorname{div}(\mu_t v_t) = 0 \tag{4}$$

where v_t should ensure the convergence of μ_t to π, for some topology over measures, in finite or infinite time. Due to v_t, eq. (4) is nonlinear over μ_t. Cauchy-Lipschitz-style assumptions for existence and uniqueness of the solution of (4) are provided in Appendix A.3. The continuity equation ensures that the mass is conserved and that it is not teleported as for a mixture $\mu_t = (1-t)\mu_0 + t\pi$. In order to adjust the position of particles only depending on the present distribution μ_t and to have an automated choice of v_t at any given time, it is favorable to have v_t as a function of μ_t, written as v_{μ_t}.

A principled way to select such a v_{μ_t} is to define it based on a Lyapunov functional $\mathcal{F}(\mu)$ over measures, decreasing along the Wasserstein gradient flow (see Appendix A.2)

$$\dot{\mathcal{F}}(\mu_t) := \frac{\mathrm{d}\mathcal{F}(\mu_t)}{\mathrm{d}t} = \langle \nabla_{W_2}\mathcal{F}(\mu_t), v_{\mu_t} \rangle_{L^2(\mu_t)} \leq 0. \tag{5}$$

Any dissimilarity $\mathcal{F}(\cdot) = D(\cdot|\pi)$ is a valid Lyapunov candidate since it is non-negative and vanishes at π. Hence, (4) can be seen as a continuous descent scheme of (1) or, conversely, (1)-(5) can be interpreted as a way to choose v_{μ_t} in (4) to steer μ_0 to π. In short, any Lyapunov-based approach rests upon three quantities $(\mathcal{F}(\mu_t), v_{\mu_t}, \dot{\mathcal{F}}(\mu_t))$, related by (5). A natural choice of v_{μ_t} satisfying (5) and realizing the steepest descent is the Wasserstein gradient itself, $v_{\mu_t} = -\nabla_{W_2}\mathcal{F}(\mu_t)$. Depending on the choice of $\mathcal{F}$, this v_{μ_t} may be hard to implement, or require specific analysis of the resulting dissipation function $\dot{\mathcal{F}}(\mu_t)$. Otherwise, to ensure that $\dot{\mathcal{F}}(\mu_t)$ only vanishes at π, one can choose v_{μ_t} so that both $\mathcal{F}(\mu_t)$ and $\dot{\mathcal{F}}(\mu_t)$ are known dissimilarities. As a matter of fact, if there exists a dissimilarity $\tilde{D}$ separating measures such that $-\dot{\mathcal{F}}(\mu) \geq \tilde{D}(\mu|\pi)$, then π is asymptotically stable for the flow and, if $\tilde{D}(\mu|\pi) \geq \mathcal{F}(\mu)$, then π is exponentially stable (by Gronwall's lemma). This relates the Lyapunov analysis to functional inequalities (Villani, 2003), expressing domination w.r.t. $\mathcal{F}$ of the Wasserstein gradient of $\mathcal{F}$ under specific assumptions on π, e.g. log-Sobolev for the KL, or Poincaré for the χ^2 (Chewi et al., 2020).

2.3. Kernel Stein Discrepancy

Consider a positive semi-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and its corresponding RKHS $\mathcal{H}_k$ of real-valued functions on $\mathbb{R}^d$. The space $\mathcal{H}_k$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ and norm $\|\cdot\|_{\mathcal{H}_k}$. Moreover, k satisfies the reproducing property: $\forall\ f \in \mathcal{H}_k,\ f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}_k}$; which for smooth kernels also holds for derivatives, e.g. $\partial_i f(x) = \langle f, (\nabla_1 k(x, \cdot))_i \rangle_{\mathcal{H}_k}$ (see Saitoh & Sawano, 2016). Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. If $\int k(x,x)\mathrm{d}\mu(x) < \infty$, then the integral operator associated to the kernel k and measure μ, denoted by $S_{\mu,k} : L^2(\mu) \to \mathcal{H}_k$ and defined as

$$S_{\mu,k} f = \int k(x, \cdot) f(x) \mathrm{d}\mu(x), \tag{6}$$

is a Hilbert-Schmidt operator and $\mathcal{H}_k \subset L^2(\mu)$. In this case, the identity embedding $\iota : \mathcal{H}_k \to L^2(\mu)$ is a bounded operator and it is the adjoint of $S_{\mu,k}$ (i.e., $\iota^* = S_{\mu,k}$ (Steinwart & Christmann, 2008, Theorems 4.26 and 4.27). Hence, for any $(f, g) \in \mathcal{H}_k \times L^2(\mu)$, $\langle \iota f, g \rangle_{L^2(\mu)} = \langle f, S_{\mu,k} g \rangle_{\mathcal{H}_k}$. We denote by $\mathcal{H}_k^d$ the Cartesian product RKHS consisting of elements $f = (f_1, \dots, f_d)$ with $f_i \in \mathcal{H}_k$, and with inner product $\langle f, g \rangle_{\mathcal{H}_k^d} = \sum_{i=1}^d \langle f_i, g_i \rangle_{\mathcal{H}_k}$. For vector-inputs, we extend $S_{\mu,k}$, applying it component-wise.

The Stein kernel k_π (3) is a reproducing kernel and satisfies a Stein identity ($\int_{\mathbb{R}^d} k_\pi(x, \cdot)\mathrm{d}\pi(x) = 0$) under mild regularity assumptions on k and π.[1] It allows for several interpretations of the KSD (2) as already discussed by Liu et al. (2016). It can be introduced as an IPM in the specific case of a Stein operator applied to $\mathcal{H}_k$ (e.g. Gorham & Mackey, 2017). It can then be identified as an asymmetric MMD in $\mathcal{H}_{k_\pi}$ (see Section 3.3). Alternatively, the squared KSD can be seen as a kernelized Fisher divergence, where the Fisher information $\|\nabla \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\|^2_{L^2(\mu)}$ is smoothed through the kernel integral operator, i.e. $\mathrm{KSD}^2(\mu|\pi) = \|S_{\mu,k} \nabla \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)\|^2_{\mathcal{H}_k^d}$. In this sense, the squared KSD has also been referred to as the Stein Fisher information (Duncan et al., 2019). Hence, minimizing the KSD can be thought as a kernelized version of score-matching (Hyvärinen & Dayan, 2005).

The metrization of weak convergence by the KSD, i.e. that $\lim_{t\to\infty} \mathrm{KSD}(\mu_t|\pi) = 0$ implies the weak convergence of μ_t to π, depends on the choice of the kernel relatively to the target. This question has been considered by Gorham & Mackey (2017), who show this is the case under assumptions akin to strong log-concavity of π at infinity (namely distant dissipativity, Eberle, 2016), and for a kernel k with a slow decay rate. This includes finite Gaussian mixtures with common covariance and kernels that are translation-invariant with heavy-tails and non-vanishing Fourier transform, such as the inverse multi-quadratic (IMQ) kernel defined by $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$ for $c > 0$ and $\beta \in (-1, 0)$, or its variants considered in Chen et al. (2018).

3. Sampling as optimization of the KSD

This section defines the KSD Descent and relates it to other gradient flows, especially the MMD gradient flow, of which the KSD Descent is a special case. In all the following, we assume that $k \in C^{3,3}(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$, and that π is such that $s = \nabla \log \pi \in C^2(\mathbb{R}^d)$.

[1] e.g., k is a Gaussian kernel and π is a smooth density fully supported on $\mathbb{R}^d$, see Liu et al. (2016, Theorem 3.7).

3.1. Continuous time dynamics

Consider the functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to [0, +\infty)$, $\mu \mapsto \frac{1}{2}\mathrm{KSD}^2(\mu|\pi)$ defined over the Wasserstein space. If $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ satisfies some mild regularity conditions (i.e., it has a C^1 density w.r.t. Lebesgue measure, and it is in the domain of $\mathcal{F}$, see Appendix A.2), the gradient of $\mathcal{F}$ at μ is well-defined and denoted by $\nabla_{W_2}\mathcal{F}(\mu) \in L^2(\mu)$. We shall consider the following assumptions on the Stein kernel:

- ($\mathbf{A}_1$) There exists a map $L(\cdot) \in C^0(\mathbb{R}^d, \mathbb{R}_+)$, which is μ-integrable for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, such that, for any $y \in \mathbb{R}^d$, the maps $x \mapsto \nabla_1 k_\pi(x, y)$ and $x \mapsto \nabla_2 k_\pi(x, y)$ are $L(y)$-Lipschitz.
- ($\mathbf{A}_2$) There exists $m > 0$ such that for any $\mu \in \mathcal{P}_c\left(\mathbb{R}^d\right)$, for all $y \in \mathbb{R}^d$, we have $\| \int \nabla_2 k_\pi(x, y)\mathrm{d}\mu(x)\| \leq m\left(1 + \|y\| + \int \|x\| \,\mathrm{d}\mu(x)\right)$.
- ($\mathbf{A}_3$) The map $(x, y) \mapsto \|H_1 k_\pi(x, y)\|_{op}$ is $\mu\otimes\nu$-integrable for every $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$.
- ($\mathbf{A}_4$) For all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\int k_\pi(x, x)\mathrm{d}\mu(x) < \infty$.

The KSD gradient flow is defined as the flow induced by the continuity equation:

$$\frac{\partial \mu_t}{\partial t} + \mathrm{div}(\mu_t v_{\mu_t}) = 0 \text{ for } v_{\mu_t} := -\nabla_{W_2}\mathcal{F}(\mu_t). \quad (7)$$

Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$) ensure that the KSD gradient flow exists and is unique, they are further discussed in Appendix A.3. Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_3$) are needed so that the Hessian of $\mathcal{F}$ is well defined (see Section 4). Assumption ($\mathbf{A}_4$) guarantees that the integral operator S_{μ,k_π} (6) is well-defined and that $\mathcal{F}(\mu) < \infty$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

Lemma 1. Assume that k, its derivatives up to order 3, and their product by $\|x - y\|$ are uniformly bounded over $\mathbb{R}^d$; and that s is Lipschitz and has a bounded Hessian over $\mathbb{R}^d$. Then Assumptions ($\mathbf{A}_1$), ($\mathbf{A}_3$) and ($\mathbf{A}_4$) hold. If, furthermore there exists $M > 0$ and M_0 such that, for all $x \in \mathbb{R}^d$, $\|s(x)\| \leq M\sqrt{\|x\|} + M_0$, then Assumption ($\mathbf{A}_2$) also holds.

See the proof in Appendix B.1. Smoothed Laplace distributions π paired with Gaussian k satisfy the assumptions of Lemma 1. For Gaussian π, s is linear, so Assumptions ($\mathbf{A}_1$), ($\mathbf{A}_3$) and ($\mathbf{A}_4$) hold for smooth kernels, but Assumption ($\mathbf{A}_2$) does not hold in general because of the $s(x)^\top s(y)$ term in k_π. Notice that most of our results are stated without Assumption ($\mathbf{A}_2$), which is only required to establish the global existence of KSD flow, in the sense that the particle trajectories are well-defined and do not explode in finite-time.

Proposition 2. Under Assumptions ($\mathbf{A}_1$) and ($\mathbf{A}_2$), the W_2 gradient of $\mathcal{F}$ evaluated at μ and its dissipation (5) along (7) are

$$\nabla_{W_2}\mathcal{F}(\mu) = \mathbb{E}_{x\sim\mu}[\nabla_2 k_\pi(x, \cdot)], \quad (8)$$

$$\dot{\mathcal{F}}(\mu_t) = -\mathbb{E}_{y\sim\mu_t}\left[\|\mathbb{E}_{x\sim\mu_t}[\nabla_2 k_\pi(x, y)]\|^2\right]. \quad (9)$$

Since the r.h.s. of (9) is negative, Proposition 2 shows that the squared KSD w.r.t. π decreases along the KSD gradient flow dynamics. In other words, $\mathcal{F}$ is indeed a Lyapunov functional for the dynamics (7) as discussed in Section 2.2.

3.2. Discrete time and discrete measures

A straightforward time-discretization of (7) is a gradient descent in the Wasserstein space applied to $\mathcal{F}(\mu) = \frac{1}{2}\mathrm{KSD}^2(\mu|\pi)$. Starting from an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, it writes as follows at iteration $n \in \mathbb{N}$,

$$\mu_{n+1} = (I - \gamma\nabla_{W_2}\mathcal{F}(\mu_n))_{\#}\,\mu_n, \quad (10)$$

for a step-size $\gamma > 0$. However for discrete measures $\hat{\mu} = \frac{1}{N}\sum_{j=1}^N \delta_{x^j}$, we can make the problem more explicit setting a loss function

$$F([x^j]_{j=1}^N) := \mathcal{F}(\hat{\mu}) = \frac{1}{2N^2}\sum_{i,j=1}^N k_\pi(x^i, x^j). \quad (11)$$

Problem (1) then corresponds to a standard non-convex optimization problem over the finite-dimensional, Euclidean, space of particle positions. The gradient of F is readily obtained as

$$\nabla_{x^i}F([x^j]_{j=1}^N) = \frac{1}{N^2}\sum_{j=1}^N \nabla_2 k_\pi(x^j, x^i).$$

since, by symmetry of k_π, $\nabla_1 k_\pi(x, y) = \nabla_2 k_\pi(y, x)$. As both F and $\nabla_{x^i}F$ can be explicitly computed, one can implement the KSD Descent either using a gradient descent (Algorithm 1) or through a quasi-Newton algorithm such as L-BFGS (Algorithm 2). As a matter of fact, L-BFGS (Liu & Nocedal, 1989) is often faster and more robust than the conventional gradient descent. It also does not require choosing critical hyper-parameters, such as a learning rate, since L-BFGS performs a line-search to find suitable step-sizes. It only requires a tolerance parameter on the norm of the gradient, which is in practice set to machine precision. A technical descent lemma for (10) (Proposition 14) showing that $\mathcal{F}$ decreases at each iteration (10) is provided in Appendix A.6. It requires the boundedness of $(\|L(\cdot)\|_{L^2(\mu_n)})_{n\geq 0}$, the L^2-norm of the Lipschitz constants of Assumption ($\mathbf{A}_1$) along the flow, as well as the convexity of $L(\cdot)$ and a compactly-supported initialization.

Algorithm 1 KSD Descent GD

Input: initial particles $(x_0^i)_{i=1}^N \sim \mu_0$, number of iterations M, step-size γ
for $n = 1$ **to** M **do**

$$[x_{n+1}^i]_{i=1}^N = [x_n^i]_{i=1}^N - \frac{\gamma}{N^2}\sum_{j=1}^N [\nabla_2 k_\pi(x_n^j, x_n^i)]_{i=1}^N, \quad (12)$$

end for
Return: $[x_M^i]_{i=1}^N$.

Algorithm 2 KSD Descent L-BFGS

Input: initial particles $(x_0^i)_{i=1}^N \sim \mu_0$, tolerance tol

Return: $[x_*^i]_{i=1}^N = \text{L-BFGS}(F, \nabla F, [x_0^i]_{i=1}^N, \text{tol})$.

Remark 1. As L-BFGS requires access to exact gradients, Algorithm 2 cannot be used in a stochastic setting. However this can be done for Algorithm 1 by subsampling over particles in the double sum in Equation (11). Moreover, in some settings like Bayesian inference, the score itself writes as a sum over observations. In this case, the loss F writes as a *double* sum over observations, and a stochastic variant of Algorithm 1 tailored for this problem could be devised, in the spirit of Clémençon et al. (2016).

3.3. Related work

Several recent works fall within the framework sketched in Section 2.2. In SVGD, Liu et al. (2016) take $\mathcal{F}$ as the KL, and set $v_{\mu_t} = -S_{\mu,k}\nabla \ln\left(\frac{d\mu_t}{d\pi}\right)$ to obtain $\dot{\mathcal{F}}$ as the (squared) KSD. Integrating this inequality w.r.t. time yields a $1/T$ convergence rate for the average KSD between μ_t and π for $t \in [0, T]$. This enabled Korba et al. (2020, Proposition 5, Corollary 6) to obtain a discrete-time descent lemma for bounded kernels, as well as rates of convergence for the averaged KSD. In contrast, since the dissipation (9) of the KSD along its W_2 gradient flow does not correspond to any dissimilarity, our descent lemma for (10) (Proposition 14) does not yield similar rates of convergence. Alternatively, in the LAWGD algorithm recently proposed by Chewi et al. (2020), $\mathcal{F}$ is the KL, and $v_{\mu_t} = -\nabla S_{\pi,k_{\mathcal{L}_\pi}}\left(\frac{d\mu_t}{d\pi}\right)$ with $k_{\mathcal{L}_\pi}$ chosen such that $\dot{\mathcal{F}}$ is the χ^2, by taking $S_{\pi,k_{\mathcal{L}_\pi}}$ as the inverse of the diffusion operator:

$$\mathcal{L}_\pi : f \mapsto -\Delta f - \langle \nabla \log \pi, \nabla f \rangle. \tag{13}$$

Their elegant approach results in a linear convergence of the KL along their flow, but implementing LAWGD in practice requires to compute the spectrum of $\mathcal{L}_\pi$. It is in general as difficult as solving a linear PDE, and Chewi et al. (2020) admit it is unlikely to scale in high dimensions.[2]

Beyond studies on the KL, Mroueh et al. (2019) considered $\mathcal{F}$ as the MMD and pick v_{μ_t} based on a kernelized Sobolev norm so that $\dot{\mathcal{F}}$ resembles the MMD, but without proving convergence of their scheme. Arbel et al. (2019) also analyzed $\mathcal{F}$ as the MMD, but for $v_{\mu_t} = -\nabla_{W_2}\frac{1}{2}\text{MMD}^2(\mu_t, \pi)$ with similar $\dot{\mathcal{F}}$ as ours and with a dedicated analysis of their MMD-GD flow. We recall that $\text{MMD}^2(\mu, \pi) = \|\int k(x,\cdot)d\mu(x) - \int k(x,\cdot)d\pi(x)\|^2_{\mathcal{H}_k}$. Since the Stein kernel satisfies the Stein's identity $\int k_\pi(x,\cdot)d\pi(x) = 0$, the KSD (2) can be identified to an MMD with the Stein kernel (Chen et al., 2018). However, the assumptions of Arbel et al. (2019) -∇k is L-Lipschitz for $L \in \mathbb{R}^+$- do not hold in general for unbounded Stein kernels. Here, we provide the right set of assumptions ($\mathbf{A}_1$)-($\mathbf{A}_2$) on k_π for the flow to exist and for a descent lemma to hold. Also, as noted on Figure 1, the sample-based MMD flow, defined through

$$\nabla_{W_2}\frac{1}{2}\text{MMD}^2(\mu, \pi) = \int \nabla_2 k(x,\cdot)\mathrm{d}(\mu - \pi)(x) \tag{14}$$

can fail dramatically while KSD flow succeeds. This suggests that the geometrical properties of the KSD flow are more favorable than the ones of the regular MMD flow. In other words, choosing an appropriate (target-dependent) kernel, as in our method or in LAWGD, appears more propitious than taking a kernel k unrelated to π.

Related to the optimization of the KSD, Stein points (Chen et al., 2018) also propose to use the KSD loss for sampling, but the loss is minimized using very different tools. While KSD descent uses a first order information by following the gradient (or L-BFGS) direction with a fixed number of particles, Stein points use a Frank-Wolfe scheme, adding particles one by one in a greedy fashion. Given N particles $x^1, \dots, x^N$, in Stein points, the next particle is set as

$$x^{N+1} \in \arg\min_x \frac{1}{2}k_\pi(x,x) + \sum_{i=1}^N k_\pi(x, x^i).$$

This problem is solved using derivative-free (a.k.a. zeroth-order) algorithms like grid-search or random sampling. The main drawback of such an approach is that it scales poorly with the dimension d when compared to first-order methods. In the same spirit, (Futami et al., 2019) have recently proposed a similar algorithm to optimize the MMD.

4. Theoretical properties of the KSD flow

In this section, we provide a theoretical study of the convergence of the KSD Wasserstein gradient flow, assessing the convexity of $\mathcal{F}$ and discussing the stationary points of its gradient flow. Remarkably, we encounter pitfalls similar to other deterministic flows derived from IPMs. This issue arises because IPMs are always mixture convex, but seldom geodesically convex. We first investigate the convexity properties of $\mathcal{F}$ along W_2 geodesics and show that exponential convergence near equilibrium cannot hold in general. Then, we examine some stationary points of its W_2 gradient flow, which explain the failure cases met in Section 5, where $\hat{\mu}_n$ converges to a degenerate measure.

4.1. Convexity properties of the KSD flow

As is well-known, decay along W_2 gradient flows can be obtained from convexity properties along geodesics. A natural object of interest is then the Hessian of the objective

[2] The update rules of SVGD, LAWGD and MMD-GD can be found in Appendix A.4.

$\mathcal{F}$. We define below this object, in a similar way as Duncan et al. (2019). We recall that $\{\nabla\psi,\ \psi \in C_c^\infty(\mathbb{R}^d)\}$ is by definition dense in $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d) \subset L^2(\mu)$ for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2008, Definition 8.4.1).

Definition 1. Consider $\psi \in C_c^\infty(\mathbb{R}^d)$ and the path ρ_t from μ to $(I + \nabla\psi)_\#\mu$ given by: $\rho_t = (I + t\nabla\psi)_\#\mu$, for all $t \in [0, 1]$. The Hessian of $\mathcal{F}$ at μ, $H\mathcal{F}_{|\mu}$, is defined as a symmetric bilinear form on $C_c^\infty(\mathbb{R}^d)$ associated with the quadratic form $\mathrm{Hess}_\mu\,\mathcal{F}(\psi,\psi) := \frac{d^2}{dt^2}\Big|_{t=0}\mathcal{F}(\rho_t)$.

Definition 1 can be straightforwardly related to the usual symmetric bilinear form defined on $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ (Otto & Villani, 2000, Section 3)[3].

Proposition 3. Under Assumptions $(\mathbf{A}_1)$ and $(\mathbf{A}_3)$, the Hessian of $\mathcal{F}$ at μ is given, for any $\psi \in C_c^\infty(\mathbb{R}^d)$, by

$$\mathrm{Hess}_\mu\,\mathcal{F}(\psi,\psi) = \mathbb{E}_{x,y\sim\mu}\left[\nabla\psi(x)^T\nabla_1\nabla_2 k_\pi(x,y)\nabla\psi(y)\right] + \mathbb{E}_{x,y\sim\mu}\left[\nabla\psi(x)^T H_1 k_\pi(x,y)\nabla\psi(x)\right]. \quad (15)$$

A proof of Proposition 3 is provided in Appendix B.3. Our computations are similar to the ones in (Arbel et al., 2019, Lemma 23) with some terms getting simpler owing to the Stein's identity satisfied by the Stein kernel. As for the squared MMD (Arbel et al., 2019, Proposition 5), the squared KSD is unlikely to be geodesically convex. Indeed, while the first term is always positive, the second term in (15) can in general take negative values, unless $H_1 k_\pi(x, y)$ is positive for all $x, y \in \mathrm{supp}(\mu)$. Nevertheless, at $\mu = \pi$, this second term vanishes, again owing to the Stein's property of k_π.

Corollary 4. Under Assumptions $(\mathbf{A}_1)$, $(\mathbf{A}_3)$ and $(\mathbf{A}_4)$, the Hessian of $\mathcal{F}$ at π is given, for any $\psi \in C_c^\infty(\mathbb{R}^d)$, by

$$\mathrm{Hess}_\pi\,\mathcal{F}(\psi,\psi) = \|S_{\pi,k_\pi}\mathcal{L}_\pi\psi\|^2_{\mathcal{H}_{k_\pi}}$$

where S_{π,k_π} and $\mathcal{L}_\pi$ are defined in Equations (6) and (13).

A proof of Corollary 4 is provided in Appendix B.4. We now study the curvature properties near equilibrium, characterized by $\mathrm{Hess}_\pi\,\mathcal{F}$. In particular, inspired by the methodology described in Villani (2003) and recently applied by Duncan et al. (2019), we expect exponential convergence of solutions initialized near π whenever the Hessian is bounded from below by a quadratic form on the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at π, included in $L^2(\pi)$.

Definition 2. We say that *exponential decay near equilibrium* holds if there exists $\lambda > 0$ such that for any $\psi \in C_c^\infty(\mathbb{R}^d)$,

$$\mathrm{Hess}_\pi\,\mathcal{F}(\psi,\psi) \geq \lambda\|\nabla\psi\|^2_{L^2(\pi)}. \quad (16)$$

[3]The W_2 Hessian of $\mathcal{F}$, denoted $H\mathcal{F}_{|\mu}$ is an operator over $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ verifying $\langle H\mathcal{F}_{|\mu}v_t, v_t\rangle_{L^2(\mu)} = \frac{d^2}{dt^2}\Big|_{t=0}\mathcal{F}(\rho_t)$ if $t \mapsto \rho_t$ is a geodesic starting at μ with vector field $t \mapsto v_t$.

According to Corollary 4, (16) can be seen as a kernelized version of the following form of the Poincaré inequality for π (Bakry et al., 2013, Chapter 5)

$$\|\mathcal{L}_\pi\psi\|^2_{L_2(\pi)} \geq \lambda_\pi\|\nabla\psi\|^2_{L_2(\pi)}. \quad (17)$$

Condition (16) is similar to the Stein-Poincaré inequality (Duncan et al., 2019, Lemma 32). We will now argue that (16) is hardly ever satisfied, thus obtaining an impossibility result reminiscent of the one for SVGD in (Duncan et al., 2019, Lemma 36), which states that exponential convergence (of the KL) for the SVGD gradient flow does not hold whenever π has exponential tails and the derivatives of $\nabla\log\pi$ and k grow at most at a polynomial rate. We start with the following characterization of exponential decay near equilibrium:

Proposition 5. Let $T_{\pi,k_\pi} = S^*_{\pi,k_\pi} \circ S_{\pi,k_\pi}$ and $L^2_0(\pi) = \{\phi \in L^2(\pi), \int \phi d\pi = 0\}$. The exponential decay near equilibrium (16) holds if and only if $\mathcal{L}_\pi^{-1} : L^2_0(\pi) \to L^2_0(\pi)$, the inverse of $\mathcal{L}_\pi|_{L^2_0(\pi)}$, is well-defined, bounded, and for all $\phi \in L^2_0(\pi)$ we have

$$\langle\phi, T_{\pi,k_\pi}\phi\rangle_{L_2(\pi)} \geq \lambda\langle\phi, \mathcal{L}_\pi^{-1}\phi\rangle_{L_2(\pi)}. \quad (18)$$

See Appendix B.5 for a proof. By the spectral theorem for compact, self-adjoint operators (Kreyszig, 1978, Section 8.3), T_{π,k_π} has a discrete spectrum $(l_n)_{n\in\mathbb{N}^*}$ which satisfies $l_n \geq 0$ and $l_n \to 0$. Under mild assumptions on π, the operator $\mathcal{L}_\pi$ also has a discrete, positive spectrum (Chewi et al., 2020, Appendix A). Proposition 5 implies the following necessary condition on the spectrum of $\mathcal{L}_\pi^{-1}$ and T_{π,k_π} for the exponential decay near equilibrium (16) to hold.

Corollary 6. If $\mathcal{L}_\pi^{-1}$ has a discrete spectrum $(\lambda_n)_{n\in\mathbb{N}^*}$ and (16) holds, then $\lambda_n = \mathcal{O}(l_n)$, i.e. the eigenvalue decay of $\mathcal{L}_\pi^{-1}$ is at least as fast as the one of T_{π,k_π}.

We also show that if $\mathcal{H}_{k_\pi}$ is infinite dimensional and exponential convergence near equilibrium holds, then $\mathcal{L}_\pi^{-1}$ has a discrete spectrum (Lemma 16). We now present our impossibility result on the exponential decay near equilibrium.

Theorem 7. Let $\pi \propto e^{-V}$. Assume that $V \in C^2(\mathbb{R}^d)$, ∇V is Lipschitz and $\mathcal{L}_\pi$ has a discrete spectrum. Then exponential decay near equilibium does not hold.

The main idea behind the proof of Theorem 7 (Appendix B.7) is that T_{π,k_π} is nuclear (Steinwart & Christmann, 2008, Theorem 4.27), which implies that its eigenvalues $(l_n)_{n\in\mathbb{N}^*}$ are summable. On the other hand the eigenvalue decay of $\mathcal{L}_\pi^{-1}$ when π is a Gaussian can be seen to be of order $O(1/n^{1/d})$ (Lemma 17 in Appendix B.7), which is not summable. The general case is obtained by comparison with a Gaussian. Remark that by Lemma 16, the assumption that $\mathcal{L}_\pi$ has a discrete spectrum in Theorem 7 can be exchanged for an assumption that $\mathcal{H}_{k_\pi}$ is infinite dimensional, which is discussed in the proof of Theorem 7. Despite the

lack of (strong) geodesic convexity near equilibrium, we still observe empirically good convergence properties of the KSD flow for discrete measures to a stationary measure. Hence, we now investigate these stationary measures.

4.2. Stationary measures of the KSD flow

The KSD gradient flow leads to a deterministic algorithm, as for SVGD and LAWGD. To study the convergence of these algorithms in continuous time, it is relevant to characterize the stationary measures, i.e. the ones which cancel the dissipation $\dot{\mathcal{F}}$ (5) of the objective functional $\mathcal{F}$ along the relative gradient flow dynamics. Unfortunately, unlike for the SVGD and LAWGD algorithms, the dissipation related to the KSD flow (9) does not yield a dissimilarity between measures. Consequently, the study of the stationary measures of the KSD is more involved. We discuss below when failure cases may happen.

Lemma 8. Assume Assumption ($\mathbf{A}_4$) holds. Then $\mathcal{H}_{k_\pi}$ does not contain non-zero constant functions.

A proof of Lemma 8 is provided in Appendix B.8. This result has the immediate consequence of considerably restricting the number of candidate fully-supported measures that are stationary for the KSD gradient flow. Consider one such measure μ_∞. At equilibrium, $\dot{\mathcal{F}}(\mu_\infty) = 0$; which implies that $\int k_\pi(x, .)d\mu_\infty(x)$ is μ_∞-a.e. equal to a constant function c. Since $x \mapsto c$ is then also an element of $\mathcal{H}_{k_\pi}$, the previous lemma implies that $c = 0$. Hence, if μ_∞ and π are full-support, $\mathcal{F}(\mu_\infty) = 0$. Provided k_π is characteristic (Sriperumbudur et al., 2011), then $\mu_\infty = \pi$.

However, as Algorithms 1 and 2 rely on discrete measures, the dissipation $\dot{\mathcal{F}}$ (9) can vanish even for $\mu \neq \pi$ because μ is not full-support. Depending on the properties of π and k, this may happen even for trivial measures such as a single Dirac mass, as stated in the following Lemma.

Lemma 9. Let x_0 such that $s(x_0) = 0$ and $Js(x_0)$ is invertible, and consider a translation-invariant kernel $k(x, y) = \phi(x - y)$, for $\phi \in C^3(\mathbb{R}^d)$. Then δ_{x_0} is a stable fixed measure of (7), i.e. it is stationary and any small push-forward of δ_{x_0} is attracted back by the flow.

Proof: For $\varepsilon > 0$ and $\psi \in C_c^\infty(\mathbb{R}^d)$, set $\mu_\varepsilon = (I+\varepsilon\psi)_\# \delta_{x_0}$. We then have $\mathcal{F}(\mu_\varepsilon) = \frac{1}{2}k_\pi(x_0 + \varepsilon\psi(x_0), x_0 + \varepsilon\psi(x_0))$. Expanding $k_\pi(x, x)$ at the first order around $x = x_0$ gives $2\mathcal{F}(\mu_\varepsilon) = \varepsilon^2\|[Js(x_0)]\psi(x_0)\|^2\phi(0) - \Delta\phi(0) + o(\varepsilon^2)$. This quantity is minimized for $\psi(x_0) = 0$, which shows that δ_{x_0} is indeed a local minimum for $\mathcal{F}$.

Importantly, this result applies whenever the score s vanishes at x_0, not only when x_0 is a local stable minimum of the potential $\log(\pi)$. This means that for a single particle, KSD descent is attracted to any stationary point of $\log(\pi)$, whereas SVGD converges only to local maxima of $\log(\pi)$ (Liu & Wang, 2016). Nonetheless, if π is log-concave, there is no spurious stationary point.

For cases more general than Lemma 9, we are interested in the sets that are kept invariant by the gradient flow. For these sets, an erroneous initialization may prevent the particles from reaching the support of π. We provide below a general result holding for any deterministic flow, beyond our specific choice (7) of v_{μ_t}, and thus holding also for SVGD.

Definition 3. Let $\mathcal{M} \subset \mathbb{R}^d$ be a closed nonempty set. We say that $\mathcal{M}$ is a flow-invariant support set for the flow $(\mu_t)_{t\geq 0}$ of (4) if for any μ_0 s.t. $\mathrm{supp}(\mu_0) \subset \mathcal{M}$, we have that the flow verifies $\mathrm{supp}(\mu_t) \subset \mathcal{M}$ for all $t \geq 0$.

Proposition 10. (Informal) Let $\mathcal{M} \subset \mathbb{R}^d$ be a smooth nonempty submanifold and $\mu_0 \in \mathcal{P}_c(\mathbb{R}^d)$ with $\mathrm{supp}(\mu_0) \subset \mathcal{M}$. Assume that, for a deterministic $(v_{\mu_t})_{t\geq 0}$ satisfying classical Caratheodory-Lipschitz assumptions (Appendix A.3), we have $v_{\mu_t}(x) \in T_{\mathcal{M}}(x)$ where $T_{\mathcal{M}}(x)$ is the tangent space to $\mathcal{M}$ at x. Then $\mathcal{M}$ is flow-invariant for (4).

The formal statement, Proposition 20, stated and proved in Appendix B.9, can be in particular applied to the ubiquitous radial kernels and to planes of symmetry of π, i.e. affine subspaces $\mathcal{M} \subset \mathbb{R}^d$ such that the density of π is symmetric w.r.t. $\mathcal{M}$. Lemma 11 is illustrated in Section 5 for a mixture of two Gaussians with the same variance (Figure 3).

Lemma 11. Let $\mathcal{M}$ be a plane of symmetry of π and consider a radial kernel $k(x, y) = \phi(\|x - y\|^2/2)$ with $\phi \in C^3(\mathbb{R})$. Then, for all $(x, y) \in \mathcal{M}^2$, $\nabla_2 k_\pi(x, y) \in T_{\mathcal{M}}(x)$ and $\mathcal{M}$ is flow-invariant for (7).

Proof idea: We show that all the terms in $\nabla_2 k_\pi(x, y)$ belong to $T_{\mathcal{M}}(x)$. This implies that the convex combination $\nabla_{W_2}\mathcal{F}(\mu)(y) = \mathbb{E}_{x\sim\mu}[\nabla_2 k_\pi(x, y)] \in T_{\mathcal{M}}(x)$. We then apply Proposition 10 to conclude.

5. Experiments

In this section, we discuss the performance of KSD Descent to sample from π in practice, on toy examples and real-world problems. The code to reproduce the experiments and a package to use KSD Descent are available at `https://github.com/pierreablin/ksddescent`. For all our experiments, we use a Gaussian kernel, as we did not notice any difference in practice w.r.t. the IMQ kernel. Its bandwith is selected by cross-validation. Implementation details and additional experiments can be found in Appendix D.

5.1. Toy examples

In the first example, we choose π to be a standard 2D Gaussian, and a Gaussian k with unit bandwidth. We initialize with 50 particles drawn from a Gaussian with mean $(1, 1)$. Figure 1 displays the trajectories of several different meth-

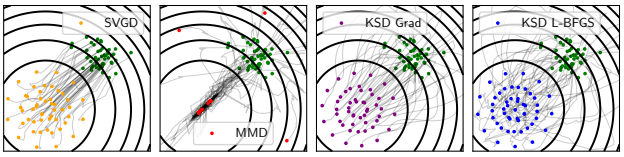

Figure 1. Toy example with 2D standard Gaussian. The green points represent the initial positions of the particles. The light grey curves correspond to their trajectories under the different v_{μ_t}.

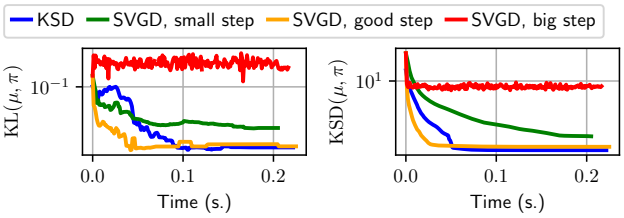

Figure 2. Convergence speed of KSD and SVGD on a Gaussian problem in 1D, with 30 particles.

ods: SVGD, KSD Descent implemented using gradient descent (Algorithm 1) and L-BFGS (Algorithm 2), and the MMD flow (Arbel et al., 2019). To assess the convergence of the algorithms, for SVGD we monitored the norm of the displacement, while for the KSD and MMD gradient flows we used the tolerance parameter of L-BFGS. KSD Descent successfully pushes the particles towards the target distribution, with a final result that is well-distributed around the mode. While KSD performs similarly to SVGD, we can notice that the trajectories of the particles behave very differently. Indeed, SVGD trajectories appear to be at first driven by the score term in the update, while the repulsive term acts later to spread the particles around the mode. In contrast, trajectories of the particles updated by KSD Descent are first influenced by the last repulsive term of the update, which seems to determine their global layout, and are then translated and contracted towards the mode under the action of the driving terms. Finally, for the MMD descent, some particles collapse around the mode, while others stay far from the target. This behavior was documented in Arbel et al. (2019), and can be partly circumvented by injecting some noise in the updates.

We then compare the convergence speed of KSD Descent and SVGD in terms of the KL or KSD objective (see Figure 2). With a fine-tuned step-size, SVGD is the fastest method. However, taking a step-size too large leads to non-convergence, while one too small leads to slow convergence. It should be stressed that it is hard to select a good step-size, or to implement a line-search strategy, since SVGD does not minimize a simple function. In contrast, the empirical KSD (11) can be minimized using L-BFGS, which does not have any critical hyper-parameter.

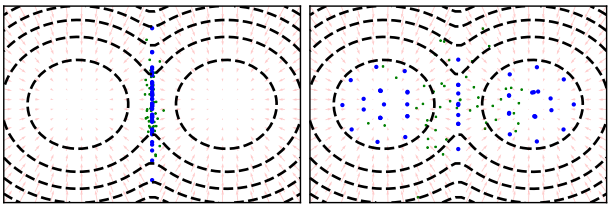

Figure 3. KSD Descent applied on a balanced mixture of Gaussian with small variance (0.1) in 2D. The centroids are at $(-1, 0)$ and $(1, 0)$. The green crosses indicate the initial particle positions, while the blue ones are the final positions. The light red arrows correspond to the score directions. **Left**: the initial positions are all chosen close to the line $x = 0$, which corresponds to an axis of symmetry of π. **Right**: even when the initialization is more spread, some particles are still caught in this spurious local minimum.

In our second example, we apply KSD Descent for π taken as a symmetric mixture of two Gaussians with the same variance. This highlights the results of Section 4.2. If initialized on the axis of symmetry, the particles are indeed stuck on it, as stated in Lemma 11. We noticed that, for a large variance of π (e.g. in $[0.2, 1]$), this axis is unstable. However, when the variance is too small (e.g. set to 0.1 as in Figure 3), the axis can even become a locally stable set. We also observed that, for a distribution initialized exclusively on one side of the axis, a single component of the mixture can be favored. This is a classical behavior of score-based methods, depending typically on the variance of π (Wenliang, 2020).

To fix this issue, we consider an annealing strategy as suggested by Wenliang (2020). It consists in adding an inverse temperature variable β to the log density of the model, i.e. $\pi^\beta(x) \propto \exp(-\beta V(x))$ for $\pi(x) \propto \exp(-V(x))$, with $\beta \in (0, 1]$. This is easily implemented with score-based methods, since it simply corresponds to multiplying $s(x)$ by β. When β is small, annealing smoothes the target distribution and the last term of the Stein kernel, repulsive at short distance, becomes dominant; on the other hand, for β close to 1, we recover the true log density. To implement this method, we start with $\beta = 0.1$, and run the KSD Descent to obtain particles at 'high temperature'. KSD Descent is then re-run starting from these particles, setting now $\beta = 1$. One can see that this strategy successfully solves the issues encountered when the KSD flow was failing to converge to the target π (Figure 4). This correction differs from the noise-injection strategy proposed in Arbel et al. (2019) for the MMD flow, which is rather related to randomized smoothing (Duchi et al., 2012). Noise-injection would prevent the use of L-BFGS in our case, as it requires exact values of the gradients of previous iterations. Annealing on the other hand is compatible with Algorithm 2.

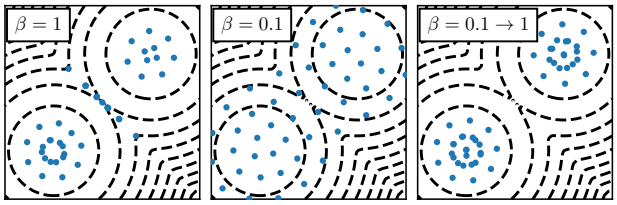

Figure 4. Effect of the annealing strategy on a mixture of Gaussians. **Left**: without annealing, some particles fall into a spurious minimum. **Middle**: with a higher temperature ($\beta = 0.1$), the particles are more spread out. **Right**: starting from the particles in the middle figure and setting $\beta = 1$ we converge to a distribution which minimizes the KSD, and has no spurious particles.

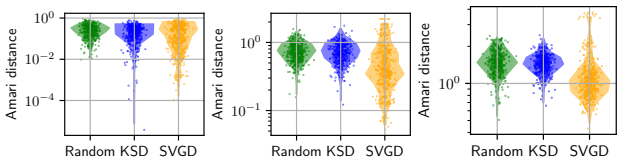

Figure 5. Bayesian ICA results. Left: $p = 2$. Middle: $p = 4$. Right: $p = 8$. Each dot corresponds to the Amari distance between an estimated matrix and the true unmixing matrix.

5.2. Bayesian Independent Component Analysis

Independent Component Analysis (ICA, Comon, 1994) is the generative model $x = W^{-1}s$, where x is an observed sample in $\mathbb{R}^p$, $W \in \mathbb{R}^{p\times p}$ is the unknown square unmixing matrix, and $s \in \mathbb{R}^p$ are the independent sources. We assume that each component has the same density $s_i \sim p_s$. The likelihood of the model is $p(x|W) = \log|W| + \sum_{i=1}^p p_s([Wx]_i)$. For our prior, we assume that W has i.i.d. entries, of law $\mathcal{N}(0,1)$. The posterior is $p(W|x) \propto p(x|W)p(W)$, and the score is given by $s(W) = W^{-\top} - \psi(Wx)x^\top - W$, where $\psi = -\frac{p_s'}{p_s}$. In practice, we choose p_s such that $\psi(\cdot) = \tanh(\cdot)$. We then use the presented algorithms to draw particles $W \sim p(W|x)$. We use $N = 10$ particles, and take 1000 samples x from the ICA model for $p \in \{2, 4, 8\}$. Each method outputs N estimated unmixing matrices, $[\tilde{W}_i]_{i=1}^N$. We compute the Amari distance (Amari et al., 1996) between each $\tilde{W}_i$ and W: the Amari distance vanishes if and only if the two matrices are the same up to scale and permutation, which are the natural indeterminacies of ICA. We repeat the experiment 50 times, resulting in 500 values for each algorithm (Figure 5). We also add the results of a random output, where the estimated matrices are obtained with i.i.d. $\mathcal{N}(0,1)$ entries. We see that for this experiment, KSD performs barely better than random, while SVGD finds matrices with lower Amari distance. One explanation is that the ICA likelihood is highly non-convex (Cardoso, 1998). This is easily seen with the invariances of the problem: permuting the rows of W does not change $p(x|W)$. As a consequence, the posterior has many saddle points, in which particles might get trapped. Unfortunately, the annealing strategy proposed above did not improve the achieved performance for this problem.

5.3. Real-world data

We compare KSD Descent and SVGD in the Bayesian logistic regression setting described in Gershman et al. (2012); Liu & Wang (2016). Given datapoints $d_1, \dots, d_q \in \mathbb{R}^p$, and labels $y_1, \dots, y_q \in \{\pm 1\}$, the labels y_i are modelled as $p(y_i = 1|d_i, w) = (1 + \exp(-w^\top d_i))^{-1}$ for some $w \in \mathbb{R}^p$. The parameters w follow the law $p(w|\alpha) = \mathcal{N}(0, \alpha^{-1}I_p)$, and $\alpha > 0$ is drawn from an exponential law $p(\alpha) = \mathrm{Exp}(0.01)$. The parameter vector is then $x = [w, \log(\alpha)] \in \mathbb{R}^{p+1}$, and we use Algorithm 2 to obtain samples from $p(x|\,(d_i, y_i)_{i=1}^q)$ for 13 datasets, with $N = 10$ particles for each.

The learning rate for SVGD and the bandwidth of the kernel for both methods are chosen through grid-search, and for each problem we select the hyper-parameters yielding the best test accuracy. For all problems, the running times of SVGD with the best step-size and of KSD Descent were similar, while KSD Descent has the advantage of having one less hyper-parameter. We present on Figure 6 the accuracy of each method on each dataset, where KSD Descent was applied without annealing since it did not change the numerical results. Our results show we match the SVGD performance without having to fine-tune the step-size, owing to Algorithm 2. We posit that KSD succeeds on this task because the posterior $p(x|\,(d_i, y_i)_{i=1}^q)$ is log-concave, and does not have saddle points.

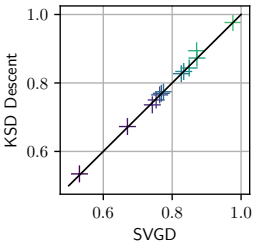

Figure 6. Accuracy of the KSD Descent and SVGD on bayesian logistic regression for 13 datasets. Both methods yield similar results. KSD is better by 2% on one dataset.

Discussion. KSD Descent benefits from a tractable loss and can be straightforwardly implemented with L-BFGS, achieving performance on par with SVGD on convex problems. However its dissipation has non-trivial stationary points, which prevents its use for non-convex problems with saddle-points such as ICA. Convergence of kernel-based sampling schemes is known to be difficult, and we provided some intuitions on the reasons for it. This leaves the door open to a more in-depth analysis of kernel-based gradient flows, especially for unbounded kernels.

Acknowledgments. A.K. thanks the GENES, and S.M. the A.N.R ABSint (Projet-ANR-18-CE40-0034) for the financial support. P.A. thanks A.N.R. ANR19-P3IA-0001.

References

Amari, S.-I., Cichocki, A., and Yang, H. H. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, pp. 757–763, 1996.

Ambrosio, L. and Crippa, G. Continuity equations and ODE flows with non-smooth velocity. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 144 (6):1191–1244, 2014.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Mackey, L., Oates, C. J., Reinert, G., and Swan, Y. Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*, 2021.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 2019.

Aubin, J.-P. and Frankowska, H. *Set-Valued Analysis*. Birkhäuser Boston, 1990.

Bakry, D., Gentil, I., and Ledoux, M. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

Bonnet, B. and Frankowska, H. Differential inclusions in Wasserstein spaces: The Cauchy-Lipschitz framework. *Journal of Differential Equations*, 271:594–637, January 2021.

Brezis, H. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

Cardoso, J.-F. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. Stein points. *International Conference on Machine Learning (ICML)*, 2018.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., Oates, C., et al. Stein point Markov Chain Monte Carlo. *International Conference on Machine Learning (ICML)*, 2019.

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

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

Clémençon, S., Colin, I., and Bellet, A. Scaling-up empirical risk minimization: optimization of incomplete u-statistics. *Journal of Machine Learning Research*, 17 (1):2682–2717, 2016.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

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

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

Eberle, A. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, 166(3): 851–886, 2016.

Fisher, M. A., Nolan, T., Graham, M. M., Prangle, D., and Oates, C. J. Measure transport with kernel Stein discrepancy. *arXiv preprint arXiv:2010.11779*, 130:1054–1062, 2021.

Futami, F., Cui, Z., Sato, I., and Sugiyama, M. Bayesian posterior approximation via greedy particle optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3606–3613, 2019.

Gershman, S., Hoffman, M., and Blei, D. Nonparametric variational inference. In *International Conference on Machine Learning (ICML)*, pp. 235–242, 2012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, volume 70, pp. 1292–1301. JMLR. org, 2017.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'ıo, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Hodgkinson, L., Salomone, R., and Roosta, F. The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*, 2020.

Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Hyvärinen, A. and Dayan, P. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Jordan, R., Kinderlehrer, D., and Otto, F. The variational formulation of the Fokker-Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. A kernel Stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*, 2020.

Klenke, A. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. A non-asymptotic analysis for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 33, 2020.

Kreyszig, E. *Introductory functional analysis with applications*, volume 1. Wiley New York, 1978.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

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

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

Mroueh, Y., Sercu, T., and Raj, A. Sobolev descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2976–2985, 2019.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

Otto, F. and Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Pankrashkin, K. Introduction to the spectral theory. *Lecture notes, Université Paris-Sud*, 2014.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Pavliotis, G. A. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

Renardy, M. and Rogers, R. C. *An introduction to partial differential equations*, volume 13. Springer Science & Business Media, 2006.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., Oates, C., et al. Optimal thinning of MCMC output. *arXiv preprint arXiv:2005.03952*, 2020.

Saitoh, S. and Sawano, Y. *Theory of Reproducing Kernels and Applications*. Springer Singapore, 2016.

Salim, A., Korba, A., and Luise, G. Wasserstein proximal gradient. In *Advances in Neural Information Processing Systems*, 2020.

Simon-Gabriel, C. J. *Distribution-Dissimilarities in Machine Learning*. PhD thesis, Eberhard Karls Universität Tübingen, Germany, 2018.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

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

Villani, C. Optimal transportation, dissipative PDE's and functional inequalities. In *Optimal transportation and applications*, pp. 53–89. Springer, 2003.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A.,

Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Wenliang, L. K. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

Wibisono, A. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. *Conference on Learning Theory*, 2018.

Xu, W. and Matsuda, T. A Stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108, pp. 320–330, 2020.