KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the gradient flow for a relaxed approximation to the Kullback-Leibler (KL) divergence between a moving source and a fixed target distribution. This approximation, termed the KALE (KL approximate lower-bound estimator), solves a regularized version of the Fenchel dual problem defining the KL over a restricted class of functions. When using a Reproducing Kernel Hilbert Space (RKHS) to define the function class, we show that the KALE continuously interpolates between the KL and the Maximum Mean Discrepancy (MMD). Like the MMD and other Integral Probability Metrics, the KALE remains well defined for mutually singular distributions. Nonetheless, the KALE inherits from the limiting KL a greater sensitivity to mismatch in the support of the distributions, compared with the MMD. These two properties make the KALE gradient flow particularly well suited when the target distribution is supported on a low-dimensional manifold. Under an assumption of sufficient smoothness of the trajectories, we show the global convergence of the KALE flow. We propose a particle implementation of the flow given initial samples from the source and the target distribution, which we use to empirically confirm the KALE's properties.

1 Introduction

We consider the problem of transporting probability mass from a source distribution $\mathbb{P}$ to a target distribution $\mathbb{Q}$ using a Wasserstein gradient flow in probability space. When the density of the target is well defined and available, the Wasserstein gradient flow of the Kullback-Liebler (KL) divergence provides a simple way to transport mass towards the target through the Fokker-Planck equation as established in the seminal work of [25]. Its time discretization yields a practical algorithm, the Unadjusted Langevin Algorithm (ULA), which comes with strong convergence guarantees (see [18, 15]). A more recent gradient flow approach, Stein Variational Gradient Descent (SVGD) [31], also leverages the analytical expression of the density and constructs a gradient flow of the KL, albeit using a metric different from the Wasserstein metric.

The KL divergence is of particular interest due to its information theoretical interpretation [50] and its use in Baysian Inference [11]. The KL defines a strong notion of convergence between probability distributions, and as such is often widely used for learning generative models, through Maximum Likelihood Estimation [16]. Using the KL as a loss requires knowledge of the density of the target, however; moreover, this loss is well-defined only when the distributions share the same support. Consequently, we cannot use the KL in settings where the probability distributions are mutually singular, or when they are only accessible through samples. In particular, the Wasserstein gradient flow of the KL in these settings is ill-defined.

Recent works have considered the gradient flow of Integral Probability Metrics (IPM) [42] instead of the KL, in settings where the target density is unknown and only samples from $\mathbb{Q}$ are available.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

This includes the Maximum Mean Discrepancy (MMD) [3] and the Kernelized Sobolev Discrepancy (KSD) [41, 40]. One motivation for considering these particle flows is their connection with the training of Generative Adversarial Networks (GANs) [23] using IPMs such as the Wasserstein distance [6, 24, 22], the MMD [19, 29, 28, 8, 9, 4] or the Sobolev discrepancy [39]. As discussed in [41, Section 3.3], these flows define update equations that are similar to those of a generator in a GAN. Thus, studying the convergence flows can provide helpful insight into conditions for GAN convergence, and ultimately, improvements to GAN training algorithms. A second motivation lies in the connection between the training dynamics of infinitely wide 2-layer neural networks and the Wasserstein gradient flow of some particular functionals [48]. Thus, analyzing the asymptotic behavior of such flows [36, 53, 14] can ultimately provide convergence guarantees for the training dynamics of neural networks. However, establishing such results remains challenging for some classes of IPMs, such as the MMD [3].

In this paper, we construct the gradient flow of a relaxed approximation of the KL, termed the KALE (KL Approximate lower-bound estimator). Unlike the KL, the KALE is well defined given any source and target, regardless of their relative absolute continuity. The KALE is obtained by solving a regularized version of the Fenchel dual problem defining the KL, defined over a restricted functions class [44, 5], and can be estimated solely from samples from the data. The version of the KALE we consider in this work benefits from two important features that are crucial for defining and analyzing a *relaxed* gradient flow of the KL. (1) We define the function class to be a Reproducing Kernel Hilbert Space (RKHS). This ensures the optimization problem defining the KALE to be convex, thus making it possible to study its generalization properties. (2) We consider a regularized version of the problem defining the KALE, thus providing a simpler expression for the gradient flow by virtue of the envelope theorem [38]. In Section 2, we review the KALE, and show that it is a divergence that metrizes the weak convergence of probability measures, while interpolating between the KL and the MMD depending on the amount of regularization. We then construct in Section 3 the Wasserstein Gradient Flow of the KALE, and we show global convergence of the KALE flow provided that the trajectories are sufficiently regular. In Section 4, we introduce the *KALE particle descent* algorithm as well as a practical way to implement it. In Section 5, we present the results obtained by running KALE particle descent on a set of problems with different geometrical properties. We show that the sensitivity to support mismatch of the KALE inherited from the KL leads to well behaved trajectories compared to the MMD flow, making the KALE flow a desirable alternative when a KL flow cannot be defined.

Related work. [44] proposed to leverage Fenchel duality to estimate the KL between two probability distributions when their density ratios belong to the approximating class. They derive a plug-in estimator for the KL which comes with convergence guarantees. In the context of GANs, [45] used the Fenchel dual representation of f-divergences, of which the KL is a particular instance, as a GAN critic. Later, [37] used Fenchel duality to estimate the KL in the context of Variational Inference (VI) when the variational distribution is chosen to be an implicit model, thus allowing more flexible models at the expense of tractability of a KL term appearing in the expression of the ELBO. In both the GAN and VI settings, the function class defining the f-divergence was restricted to neural networks. Recently, [5] showed that controlling the smoothness of such a function class results in a divergence, the KL approximate lower-bound estimator (KALE), that metrizes the *weak convergence of distributions* [17], unlike the KL which defines a stronger topology [55]. The KALE is therefore well-suited for learning Implicit Generative Models which are only accessible through sampling, as advocated in [7]. When neural network classes are used, however, the method has no optimization guarantees, as the dual problem becomes non-convex due to the choice of the function class. This is unlike our setting [and that of 44], since our dual problem is strongly convex and comes with guarantees.

2 Interpolating between KL and MMD using KALE

In this section, we introduce the *KALE*, a relaxed approximation of the KL divergence. Although we will use the KALE to define a relaxed KL gradient flow, we show in this section that the KALE is an object of independent interest outside of the gradient flow setting: indeed, it is a valid *probability divergence* that metrizes the weak convergence of probability distributions, and interpolates between the KL and the Maximum Mean Discrepancy.

Mathematical details and notation We start by introducing some notation. We equip $\mathbb{R}^d$ with its σ-algebra, and consider the set of probability distributions defined on an open subset $\mathcal{X}$ of $\mathbb{R}^d$. Weak convergence of a sequence of probability measures $(\mathbb{P}_n)_{n\geq 0}$ towards $\mathbb{P}$ is written $\mathbb{P}_n \rightharpoonup \mathbb{P}$. A positive definite kernel on the set $\mathcal{X}$ will be denoted $k : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$, with RKHS $\mathcal{H}$. The dirac delta measure for $x \in \mathcal{X}$ will be written δ_x. We denote by $C_c^\infty(\mathcal{X}\times]0,+\infty[)$ the set of infinitely differentiable functions with compact support on $\mathcal{X}\times]0,+\infty[$, and the set of continuous bounded functions from $\mathcal{X}$ to $\mathbb{R}$ by $C_b^0(\mathcal{X})$. Sets of N points in $\mathcal{X}$ will be indexed using an superscript $\{x^{(i)}\}_{i=1}^N$, while a sequence of points in $\mathcal{X}$ will use a subscript: $(x_n)_{n\in\mathbb{N}}$. If random, elements of such sets $X^{(i)}$ or iterates of such sequences X_n will be capitalized. If not, they will be kept in lower-case.

2.1 The KL Approximate Lower-bound Estimate (KALE)

The central equation to derive KALE is the Fenchel duality formulation of the KL [2, Lemma 9.4.4]:

$$\mathrm{KL}(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{h\in C_b^0(\mathcal{X})} \int 1 + h\mathrm{d}\mathbb{P} - \int e^h \mathrm{d}\mathbb{Q}. \tag{1}$$

KALE is obtained from Eq. (1) by restricting the variational set to an RKHS $\mathcal{H}$ with reproducing kernel k, and by adding a penalty to the objective that controls the RKHS norm of the test function h. This regularization ensures that the KALE is well-defined for a broader class of probabilities compared to the KL, even when $\mathbb{P}$ and $\mathbb{Q}$ are mutually singular. Its complete definition is stated below:

Definition 1 (KALE)**.** *Let $\lambda > 0$, and $\mathcal{H}$ be a RKHS with kernel k. The Kullback-Leibler Approximate Lower bound Estimate (KALE) is given by:*

$$\textit{KALE}(\mathbb{P} \,||\, \mathbb{Q}) = (1+\lambda) \sup_{h\in\mathcal{H}} \left\{ \int 1 + hd\mathbb{P} - \int e^h d\mathbb{Q} - \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 \right\}. \tag{2}$$

The $(1+\lambda)$ scaling will prevent a degenerate decay to 0 in the large λ regime (see Proposition 1). The definition we consider here also differs from the one in [5], which first finds the optimal function $h^\star$ solving Eq. (2), and then defines KALE by evaluating the KL objective in Eq. (1), thereby discarding the regularization term when evaluating the divergence.

KALE is a probability divergence We first show that KALE is a probability divergence, and presents topological properties compatible with its use in generative models, such as GAN and Adverserial VAEs: weak continuity, and metrizing the weak convergence of probability distributions. We recall that a functional $\mathcal{D}(\cdot \,||\, \cdot)$ is a probability divergence if both $\mathcal{D}(\mathbb{P} \,||\, \mathbb{Q}) \geq 0$ and $\mathcal{D}(\mathbb{P} \,||\, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$ for any pair of probability measures $\mathbb{P}, \mathbb{Q}$.

Theorem 1 (Topological properties of KALE)**.** *Let $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$ be two probability measures defined on $\mathcal{X}$. Let $(\mathbb{P}_n)_{n\geq 0}$ be a sequence of probability measures. Then, under Assumption 1:*

(i) KALE is weakly continuous: $\mathbb{P}_n \rightharpoonup \mathbb{P} \Longrightarrow \lim_{n\to\infty} \textit{KALE}(\mathbb{P}_n \,||\, \mathbb{Q}) = \textit{KALE}(\mathbb{P} \,||\, \mathbb{Q})$

(ii) If k is universal [52], then for any $\lambda > 0$, KALE is a probability divergence. Moreover, KALE metrizes the weak topology between probability measures with finite first order moments.

(i) is novel and leverages the strong convexity of the objective, as well as the specific properties of RKHS. The proof for (ii) is inspired by [5], which in turn derives from [59] [32], and is adapted to account for the extra norm penalty term in the version of KALE in this paper.

Interpolating between the MMD and the KL using the KALE KALE includes a positive regularization parameter λ, inducing two asymptotic regimes: the regime when λ approaches 0, and the one where it grows to $+\infty$. In these regimes, the KALE asymptoticaly recovers on the one hand the KL divergence, and on the other hand the MMD. We recall that given an RKHS $\mathcal{H}$ associated to a kernel k, and two probability distributions $\mathbb{P}$ and $\mathbb{Q}$, the MMD is defined as the norm of the difference of mean embeddings of $\mathbb{P}$ and $\mathbb{Q}$,

$$\mathrm{MMD}(\mathbb{P} \,||\, \mathbb{Q}) = \|f_{\mathbb{P},\mathbb{Q}}\|_{\mathcal{H}} \quad \left(f_{\mathbb{P},\mathbb{Q}} = \int k(x,\cdot)\mathrm{d}\mathbb{P} - \int k(x,\cdot)\mathrm{d}\mathbb{Q} \triangleq \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \tag{3}$$

Proposition 1 (asymptotic properties of KALE)**.** *Let* $\mathbb{P}$ *and* $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$*. Then, under Assumption 1 the following result holds:*

$$\lim_{\lambda \to +\infty} \mathit{KALE}(\mathbb{P} \,||\, \mathbb{Q}) = \frac{1}{2} \mathit{MMD}^2(\mathbb{P} \,||\, \mathbb{Q}). \tag{4}$$

Suppose additionally that $\log \frac{d\mathbb{P}}{d\mathbb{Q}} \in \mathcal{H}$*. Then,*

$$\lim_{\lambda \to 0} \mathit{KALE}(\mathbb{P} \,||\, \mathbb{Q}) = \mathit{KL}(\mathbb{P} \,||\, \mathbb{Q}). \tag{5}$$

Proposition 1 shows that the MMD can be seen as solving a degenerate version of the KL objective. Eq. (5) is natural given the original definition of the KALE, and highlights the continuity of the KALE objective w.r.t the regularization parameter λ. Both the MMD and the KL exhibit limitations when used for defining gradient flows, however: as discussed in [3, 20, 12], the MMD induces a "flat" geometry, making its use in generative models tricky [4]. On the other hand, the KL comes with stronger convergence guarantees [2], but its use in sampling algorithms is limited to cases where the target distribution has a density, discarding cases satisfying the widely known *manifold hypothesis* [43, 12, 13], stating that typical high dimensional data used in machine learning are distributed on a lower-dimensional manifold. For this reason, we argue that the true interest of the KALE does not lie in its interpolation properties, but rather in the geometry it generates at intermediate values of λ.

The KALE's dual objective Interestingly, the KALE itself admits a dual formulation, with a strong connection to the original KL expression:

$$\begin{aligned} \mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q}) &= \min_{f>0} \int (f(\log f - 1) + 1)\, \mathrm{d}\mathbb{Q} + \frac{1}{2\lambda} \left\| \int f(x) k(x, \cdot) \mathrm{d}\mathbb{Q}(x) - \mu_{\mathbb{P}} \right\|_{\mathcal{H}}^2 \\ h^\star &= \int f^\star(x) k(x, \cdot) \mathrm{d}\mathbb{Q}(x) - \mu_{\mathbb{P}} \end{aligned} \tag{6}$$

The solution $f^\star$ of Eq. (6) can be seen as an entropically-regularized density ratio estimate on the support of $\mathbb{Q}$ (additional details on the KALE dual objective are given in the appendix). Eq. (6) also yields an elegant estimation procedure, as discussed below.

Computing KALE$(\mathbb{P} \,||\, \mathbb{Q})$ **in practice** As for other IPMs, computing $\mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q})$ for arbitrary $\mathbb{P}$ and $\mathbb{Q}$ is intractable and has to be approximated using a discretization procedure. A common procedure is to assume access to samples $\{Y^{(i)}\}_{i=1}^N$ and $\{X^{(i)}\}_{i=1}^N$ from $\mathbb{P}$ and $\mathbb{Q}$ and to solve the empirical equivalent of Eq. (6) (e.g. Eq. (6), but where $\mathbb{P}$ and $\mathbb{Q}$ are replaced by their plug-in estimators $\widehat{\mathbb{P}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y^{(i)}}$ and $\widehat{\mathbb{Q}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{X^{(i)}}$). This empirical equivalent is written

$$\min_{f>0} \quad \frac{1}{N}\sum_{i=1}^N f(X^{(i)}) \log(f(X^{(i)})) - f(X^{(i)}) - 1 + \frac{1}{2\lambda} \left\| \frac{1}{N}\sum_{i=1}^N f(X^{(i)}) k(X^{(i)}, \cdot) - \mu_{\widehat{\mathbb{P}}} \right\|_{\mathcal{H}}^2 \tag{7}$$

which is a strongly convex N-dimensional problem, and can be solved using standard euclidean optimization methods. As shown in [5], the KALE's sample complexity scales with $\frac{1}{\sqrt{N}}$, which is similar to the sample complexity of Sinkhorn divergences [21].

3 KALE Gradient Flow

Having introduced KALE as a relaxed approximation of the KL, we now construct the KALE gradient flow, and assert its well-posedness. We provide conditions for global convergence of the flow, and discuss its relationship with the MMD flow and the KL flow. All proofs are given in the appendix.

3.1 Wasserstein Gradient Flow of KALE

Wasserstein Gradient Flows of divergence functionals $\mathcal{F}(\mathbb{P} \,||\, \mathbb{Q})$ aim at transporting mass from an initial probability distribution $\mathbb{P}_0$ towards a target distribution $\mathbb{Q}$ by following a path $\mathbb{P}_t$ in probability space. The path is required to dissipate the energy, meaning that $t \mapsto \mathcal{F}(\mathbb{P}_t \,||\, \mathbb{Q})$ is a decreasing function of time. Additionally, it is constrained to satisfy a continuity equation that allows only

local movements of mass without jumping from a location to another. This equation involves a time dependent vector field V_t which serves as a force that drives the movement of mass at any time t:

$$\partial_t \mathbb{P}_t + \mathrm{div}(\mathbb{P}_t V_t) = 0. \tag{8}$$

Eq. (8) holds in the *sense of distributions*, meaning that for any test function $\varphi \in \mathcal{C}_c^\infty(\mathcal{X} \times]0+\infty[)$, we have:

$$\int \partial_t \varphi(x,t) \mathrm{d}\mathbb{P}\mathrm{d}t - \int \langle \nabla_x \varphi(x,t), V_t \rangle_{\mathbb{R}^d} \, \mathrm{d}\mathbb{P}_t \mathrm{d}t.$$

The Wasserstein gradient flow of a functional $\mathcal{F}$ is then obtained by choosing V_t as the gradient of the *first variation* of $\mathcal{F}$, defined as the Gateaux derivative of $\mathbb{P}$ along the direction χ,

$$\mathcal{D}_{\mathbb{P}} \mathcal{F}(\mathbb{P}; \chi) \triangleq \lim_{\epsilon \to 0} \epsilon^{-1} \left(\mathcal{F}(\mathbb{P} + \epsilon \chi) - \mathcal{F}(\mathbb{P}) \right),$$

where $\int d\chi = 0$, and provided that such a limit exists. This choice recovers a particle *Euclidean gradient flow* when $\mathbb{P}_0$ is a finite sum of Dirac distributions, and can thus be seen as a natural extension of gradient flows to the space of probability distributions [2, 56, 57]. In the next proposition, we show that the functional $\mathbb{P} \longmapsto \mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q})$ admits a first variation, and a well-defined gradient flow.

Proposition 2 (Gradient Flow of KALE)**.** *Let $\lambda > 0$, and $\mathbb{P}_0, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$. Under Assumptions 1 and 2, the Cauchy problem*

$$\partial_t \mathbb{P}_t - \mathit{div}(\mathbb{P}_t (1+\lambda) \nabla h_t^\star) = 0, \quad \mathbb{P}_{t=0} = \mathbb{P}_0, \tag{9}$$

where $h_t^\star$ is the unique solution of

$$h_t^\star = \arg\max_{h \in \mathcal{H}} \left\{ \int h d\mathbb{P}_t - \int e^h d\mathbb{Q} - \frac{\lambda}{2} \|h\|^2 \right\}, \tag{10}$$

admits a unique *solution $(\mathbb{P}_t)_{t \geq 0}$, which is the* Wasserstein Gradient Flow *of the KALE.*

3.2 Convergence properties of the KALE flow

Proposition 2 hints at a connection between the KALE flow and the MMD flow, which solves:

$$\partial \mathbb{P}_t - \mathrm{div}\left(\mathbb{P}_t \nabla f_{\mathbb{P}_t, \mathbb{Q}}\right) = 0, \quad \mathbb{P}_{t=0} = \mathbb{P}_0 \tag{11}$$

The MMD flow and the KALE flow thus differ in the choice of witness function characterizing their velocity field. A convergence analysis of the MMD flow was proposed for a wide range of kernels in [3] using inequalities of Lojasiewicz type; in particular, the MMD flow is guaranted to converge provided that the quantity $\mathbb{P}_t - \mathbb{Q}$ remains bounded in the *negative Sobolev norm* $\|\mathbb{P}_t - \mathbb{Q}\|_{\dot{H}^{-1}(\mathbb{P}_t)}$ [46]. In the next proposition, we extend the condition ensuring the global convergence of the MMD flow [3] to the KALE flow:

Proposition 3. *Under Assumptions 1 and 2, if $\|\mathbb{P}_t - \mathbb{Q}\|_{\dot{H}^{-1}(\mathbb{P}_t)} \leq C$ for some $C > 0$, then:*

$$\mathit{KALE}(\mathbb{P}_t \,||\, \mathbb{Q}) \leq \frac{C}{C\mathit{KALE}(\mathbb{P}_0 \,||\, \mathbb{Q}) + t}.$$

Proposition 3 ensures a convergence rate in $\mathcal{O}(1/t)$ provided that $\|\mathbb{P}_t - \mathbb{Q}\|_{\dot{H}^{-1}(\mathbb{P}_t)}$ remains bounded. This convergence rate is slower than the linear rate of the KL along its gradient flow [33], and could be an effect of the RKHS smoothing.

4 KALE Particle Descent

We are now interested in deriving a practical algorithm that computes the solution of a KALE gradient flow, given an initial source-target pair $\mathbb{P}_0$ and $\mathbb{Q}$. Because of the continuous-time dynamics, and the possibly continuous nature of $\mathbb{P}_0$ and $\mathbb{Q}$, solutions of Eq. (9) are intractable to compute and manipulate. To address this issue, we first introduce the *KALE particle descent algorithm* that returns a sequence $(\widehat{\mathbb{P}}_n^N)_{n \geq 0}$ of discrete probability measures able to approximate the forward Euler discretization of $\mathbb{P}_t$ with arbitrary precision. Additionally, we show that the KALE particle descent algorithm can be regularized using *noise injection* [3], which guarantees global convergence of the flow under a suitable noise schedule. All proofs are given in the appendix.

4.1 The KALE particle descent algorithm

Time-discretized KALE Gradient Flow As a first step towards deriving the KALE particle descent algorithm, let us first consider a time-discretized version of the KALE gradient flow (Eq. (9) and Eq. (10)), obtained by applying a forward-Euler scheme to Eq. (9) with step size γ. This time-discretized equation is given by

$$\mathbb{P}_{n+1} = (I - \gamma \nabla h_n^\star)_\# \mathbb{P}_n, \ \mathbb{P}_{n=0} = \mathbb{P}_0. \tag{12}$$

The function $h_n^\star$ is a discrete time analogue of Eq. (10) in that it is solution to the following optimization problem:

$$h_n^\star = \arg\max_{h \in \mathcal{H}_n} \left\{ \int h \mathrm{d}\mathbb{P}_n - \int e^h \mathrm{d}\mathbb{Q} - \frac{\lambda}{2} \|h\|^2 \right\}. \tag{13}$$

The solution $\mathbb{P}_n$ of Eq. (12) is a sensible approximation of $\mathbb{P}_t$: indeed, it can be shown under suitable smoothness assumptions [49], [3] that the piecewise-constant trajectory $(t \longmapsto \mathbb{P}_n$ if $t \in [n\gamma, (n+1)\gamma[)$ obtained from the time-discretized gradient flow of a functional $\mathcal{F}$ will recover the true gradient flow solution $\mathbb{P}_t$ of $\mathcal{F}$ as $\gamma \to 0$.

Approxation using finitely many samples: the KALE particle descent algorithm Despite its discrete-time nature, the sequence $(\mathbb{P}_n)_{n\geq 0}$ may still be intractable to compute: for generic $\mathbb{P}_0$ and $\mathbb{Q}$, Eq. (13) will contain intractable expectations, and have an infinite dimensional search space. To address this issue, we propose the *KALE particle descent algorithm*: this algorithm *approximates* the true solution $\mathbb{P}_n$ given N samples $\{X^{(i)}\}_{i=1}^N$ and $\{Y_0^{(i)}\}_{i=1}^N$ of $\mathbb{Q}$ and $\mathbb{P}_0$, by computing the trajectory $\widehat{\mathbb{P}}_n^N$ solving the time-discrete KALE gradient flow arising from the *empirical* source-target pair $\widehat{\mathbb{Q}}^N = \frac{1}{N}\sum_{i=1}^N X^{(i)}$ and $\widehat{\mathbb{P}}_0^N = \frac{1}{N}\sum_{i=1}^N Y_0^{(i)}$. As opposed to $\mathbb{P}_n$, it is possible to *exactly* compute $\widehat{\mathbb{P}}_n^N$: indeed, the recursion equation Eq. (12) implies that $\widehat{\mathbb{P}}_n^N$ remains discrete for all n. More precisely, we have $\widehat{\mathbb{P}}_n^N = \frac{1}{N}\sum_{i=1}^N Y_n^{(i)}$, where

$$Y_{n+1}^{(i)} = Y_n^{(i)} - \gamma \nabla \widehat{h}_n^\star(Y_n^{(i)}), \tag{14}$$

and $\widehat{h}^\star$ is defined as

$$\widehat{h}_n^\star = \arg\max_h \left\{ \int h d\widehat{\mathbb{P}}_n^N - \int h d\widehat{\mathbb{Q}}^N - \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 \right\}. \tag{15}$$

Importantly, since $\widehat{\mathbb{P}}_n^N$ and $\widehat{\mathbb{Q}}^N$ are discrete, Eq. (15) reduces to an N- dimensional convex problem, and $\widehat{h}_n^\star$ can be tractably computed. The alternate execution of Eq. (14) and Eq. (15) for a finite number of time steps $n_{\max}$ defines the *KALE particle descent algorithm*, that we lay out in Algorithm 1

Consistency of the KALE particle descent algorithm Note that source of error in the KALE particle descent algorithm lies in the use of an approximate witness function $\widehat{h}_n^\star$ instead of the true, but intractable, $h_n^\star$. Indeed, one can show, using the theory of McKean-Vlasov representative processes [35], that the n-th iterates of the sequence defined by:

$$\bar{Y}_{n+1}^{(i)} = \bar{Y}_n^{(i)} - \gamma \nabla h_n^\star(\bar{Y}_n^{(i)}), \ \bar{Y}_0^{(i)} \sim \mathbb{P}_0, \ 1 \leq i \leq N \tag{16}$$

are distributed according to the true discrete-time KALE gradient flow solution $\mathbb{P}_n$ defined in Eq. (12). As such, the discrete probability $\bar{\mathbb{P}}_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{Y_n^{(i)}}$ may be considered as an unbiased, space-discretization of Eq. (12). In the next proposition, we show the iterates $\widehat{\mathbb{P}}_n^N$ returned by the KALE particle descent algorithm can approximate the unbiased $\bar{\mathbb{P}}_n^N$ with arbitrarily low error.

Proposition 4 (Consistency of the KALE particle descent)**.** *Let $\{Y_0^{(i)}\}_{i=1}^N \sim \mathbb{P}_0$. Let $(\bar{\mathbb{P}}_n^N)_{n\geq 0}$ be the sequence of discrete probabilities arising from Eq. (16) with initial conditions $\{Y_0^{(i)}\}_{i=1}^N$, and let $(\widehat{\mathbb{P}}_n^N)_{n\geq 0}$ be the sequence arising from Eq. (14) with the* same *initial conditions $\{Y_0^{(i)}\}_{i=1}^N$. Let $n_{max} \geq 0$. Then, for all $n \leq n_{\max}$, the following bound holds:*

$$\mathbb{E}W_2(\widehat{\mathbb{P}}_n^N, \bar{\mathbb{P}}_n^N) \leq \frac{A}{\sqrt{N}}(e^{\gamma B n_{\max}} - 1)$$

with $A = \sqrt{2LdK(1 + e^{\frac{8K}{\lambda}})} \times \frac{1}{4\sqrt{K}\mu + \sqrt{2d}L}$ *and* $B = \frac{(1+\lambda)(4\sqrt{K}\mu + \sqrt{2d}L)}{\lambda}$

Proposition 4 shows that given a finite time horizon $n_{\max}$, and given suffiently many samples of $\mathbb{P}_0$ and $\mathbb{Q}$, one can approximate an exact discrete KALE flow between $n = 0$ and $n = n_{\max}$ with arbitrary precision. The proof of Proposition 4 requires showing the regularity of the KALE witness function $h^\star$ with respect to the input probability $\mathbb{P}$. Note that the dependency with the dimension d comes from a constant also present in [47].

Algorithm 1 KALE Particle Descent Algorithm

Input: $\{Y^{(i)}\}_{i=1}^N \sim \mathbb{P}_0$, $\{X^{(i)}\}_{i=1}^N \sim \mathbb{Q}$, max_iter max_iter, λ, k, γ
for $n = 1$ **to** max_iter **do**
  `h_star` $\leftarrow$ `DualSolve`$(X^{(1)}, Y^{(1)}, \ldots, X^{(N)}, Y^{(N)}, \lambda)$
  for $j = 1$ **to** N **do**
    `v` $\leftarrow$ `backprop`$(\texttt{log_ratio}, Y^{(j)})$
    $Y^{(j)} \leftarrow Y^{(j)} - \gamma \times$ `v`
  end for
end for

4.2 Regularization of KALE particle descent using Noise Injection

In practice, guaranteeing the convergence of a KALE gradient flow (and its corresponding KALE particle descent) by relying on the condition given in Proposition 3 is cumbersome for two reasons: first, this condition is hard to check, and second, it does not tell us what to do when the condition is not met. Noise injection [3, 10, 51] is a practical regularization technique originally introduced for the MMD flow, that trades off some of the "steepest descent" property of gradient flow trajectories with some additional smoothness (in negative Sobolev norm) in order to improve convergence to the target trajectory. We recall that the solution of a noise injected gradient flow with velocity field $\nabla h_n^\star$ and noise schedule β_n is defined as:

$$\mathbb{P}_{n+1} = ((x, u) \longmapsto x - \gamma \nabla h_n^\star(x + \beta_n u))_\# (\mathbb{P}_n \otimes g), \tag{17}$$

where g is a standard unit gaussian distribution. As we show in the next proposition, under a suitable noise schedule, noise injection can also be applied to ensure global convergence of the KALE flow.

Proposition 5 (Global Convergence under noise injection dynamics)**.** *Let $\mathbb{P}_n$ be defined as Eq. (17). Let $(\beta_n)_{n\geq 0}$ be a sequence of noise levels, and define $\mathcal{D}_{\beta_n} = \mathbb{E}_{y\sim\mathbb{P}_n, u\sim g} \|\nabla h^\star(x + \beta_n u)\|^2$ with g the density of a standard gaussian distribution. Then, under Assumptions 1 to 3, and for a choice of β_n such that:*

$$8\mu^2\beta_n^2 \mathit{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) \leq \mathcal{D}_{\beta_n}(\mathbb{P}_n),$$

the following inequality holds: $\mathit{KALE}(\mathbb{P}_{n+1} \mid\mid \mathbb{Q}) - \mathit{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) \leq -\frac{\gamma}{2}(1 - 3\gamma L) D_{\beta_n}(\mathbb{P}_n)$, where μ and L are defined in Assumption 2 and depend only on the choice of the kernel. Moreover, if $\sum_{i=1}^{\infty} \beta_i = +\infty$, then $\lim_{n\to\infty} \mathit{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) = 0$.

As in [3], convergence of the regularized KALE flow is guaranteed when the noise schedule satisfies an inequality for all n, which is hard to check in practice. Nonetheless, we empirically observe that in all our problems, a simple small, constant noise schedule can help the KALE flow reach a lower KALE value at convergence.

Let us stress that the noise injection scheme given in Proposition 5 is a *population* scheme that includes an intractable convolution. To use noise injection in the KALE descent algorithm, we approximate this convolution using a single sample $U_n^{(i)}$ for each particle update, which is written

$$Y_{n+1}^{(i)} = Y_n^{(i)} - \gamma \nabla \widehat{h}_n^\star(Y_n^{(i)} + \beta_n U_n^{(i)}), \;\; U_n^{(i)} \sim \mathcal{N}(0, 1). \tag{18}$$

Implementation. The particle descent algorithm can be implemented using automatic differentiation software such as the `pytorch` library in python. This allows us to easily compute the gradient of the log-density ratio estimate $\widehat{h}_n^\star$ appearing in the the particle update rule Eq. (14).

Estimating $\widehat{h}_n^\star$ can be achieved using methods such as gradient descent, coordinate descent or higher order optimization methods such as Newton's method and L-BFGS [30].

5 Experiments

In this section, we empirically study the behavior of the KALE particle descent algorithm in three settings reflecting different topological properties for the source-target pair: a pair with a target supported on a hypersurface (zero volume support), a pair with disjoint supports of positive volume, and a pair of distributions with a well defined density supported on $\mathbb{R}^d$.

KALE flow for targets defined on hypersurfaces Our first example consists in a target *supported* (and uniformly distributed) on a lower-dimensional surface that defines three non-overlapping rings. The initial source is a gaussian distribution with a mean in the vicinity of the target $\mathbb{Q}$. This setting is a perfect candidate to illustrate the failure modes of both the KL and the MMD when used in particle descent algorithms: on the one hand, the measures $\mathbb{P}_0$ and $\mathbb{Q}$ are mutually singular, and thus the the KL gradient flow from $\mathbb{P}_0$ to $\mathbb{Q}$ does not exist. By contrast, the KALE is well defined in this case, and inherits from the KL an increased sensitivity to support discrepancy. For that reason, we hypothesize that the trajectories of the KALE flow will converge towards a better limit compared to its MMD flow counterpart. We sample $N = 300$ points from the target and the initial source distribution and run an implementation of Algorithm 1 for $n = 50000$ iterations. The complete set of parameters is given in the appendix. Results are plotted in Fig. 1. We indeed notice that the KALE flow trajectory remains

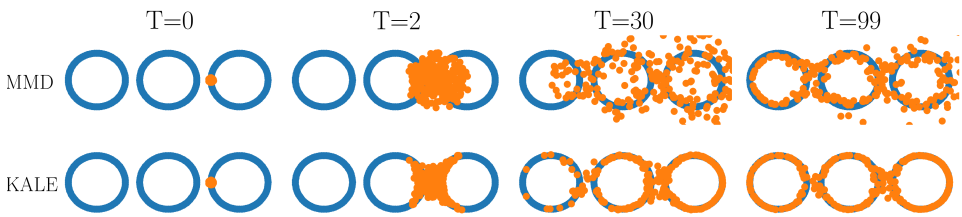


Figure 1: MMD and KALE flow trajectories for "three rings" target

close to the target support and recovers the target almost perfectly. This illustrates the ability of the KALE flow to *relax* the hard support-sharing constraints of the KL flow into soft support closeness constraints. These soft constraints are not present in the MMD flow, where particles of the source can remain scattered around the plane.

KALE flow between probabilities with disjoint support In our second example, we consider a source/target pair that are supported on disjoint subsets with a finite, positive volume each (unlike the previous example). The support of the source and the target consist respectively of a heart and a spiral, and the two distributions have a uniform density on their support. Again, because the supports of the source and the target are disjoint, the KL flow cannot be defined, nor simulated for this pair. We run a KALE particle descent algorithm, and compare it as before with an MMD flow, as well as with a "Sinkhorn descent algorithm" [20]. Results are in Fig. 2.

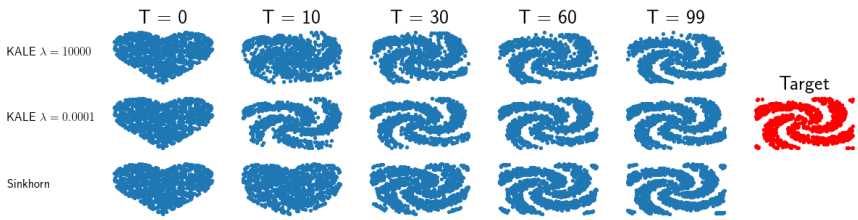
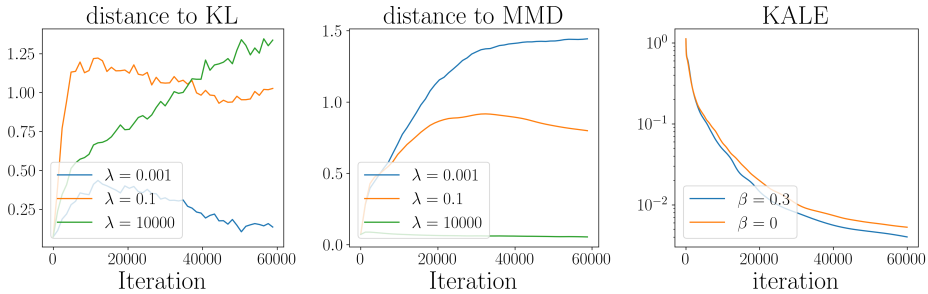


Figure 2: Shape Transfer using the KALE flow

As we can see, the soft support-sharing constraint informing the KALE flow forces the source to quickly recover the spiral shape, much before the Sinkhorn and MMD flow trajectories. However, compared to Sinkhorn, the two KALE-generated spirals have a harder time recovering outliers, disconnected from the main support of the spiral.

KALE flow for probabilities with densities We consider the setting where the target admits a density on $\mathbb{R}^d$. Hence, unlike in the two previous examples, the KL gradient flow admits a well defined density and can be simulated using the Unadjusted Langevin Algorithm (ULA). Echoing the interpolation property of KALE between the MMD and the KL shown in Proposition 1, we propose to investigate whether this property is preserved in a *gradient flow* setting. We consider a balanced mixture of 4 gaussians with means located on the 4 corners of the unit square for the target and a source distribution given by a unit gaussian in the vicinity of the unit square. We then run KL, MMD, and KALE flows with different values of λ, and compute the Wasserstein distance between reference particles at iteration n from either the MMD or KL flow and particles obtained from the KALE flow at the same iteration n. The choice of the Wasserstein distance is natural for Wasserstein Gradient Flows. As shown in Fig. 3a, for "small" λ, particles from a KALE flow remain close to the ULA particles, while for "large" λ they remain close to the MMD particles (Fig. 3b).



(a) (b) (c)

Figure 3: (a): Evolution of the Wasserstein distance between reference particles from the ULA algorithm and and the KALE descent algorithm with various values of λ. (b) Left: same as (a), but taking particles from the MMD flow as reference. (c) Evolution of the KALE along the trajectories of a KALE descent algorithm with the same mixture of gaussians as target, using $\lambda = 0.1$. Orange: without noise injection. Blue: with noise injection using a constant noise schedule.

Impact of noise injection On all three examples, using a regularized KALE flow with an appropriately tuned λ schedule always improves the proximity to the global minimum $\mathbb{P}_\infty = \mathbb{Q}$. Its effect is particularily impactful in the mixture of gaussians example, where a small, constant noise schedule β_n allows for faster mixing times for $\mathbb{P}_n$, as opposed to its unregularized counterpart, see Fig. 3c. We provide a thorough analysis of the impact of noise injection in the appendix.

6 Discussion and further work

We have constructed the KALE flow, a gradient flow between probability distributions that relaxes the KL gradient flow for probabilities with disjoint support. Using the *KALE Particle Descent Algorithm*, we have shown that in cases where a KL Gradient Flow cannot be defined, trajectories of the KALE flow exhibit better convergence properties when compared to the MMD Flow, a flow that the KALE is also able to interpolate. In cases where the KL flow can be defined, we notice empirically that the KALE flow can *approximate* the trajectories of the KL flow, but using only information from *samples* of the target. This latter property is in sharp contrast with KL Gradient Flow discretizations like the Unadjusted Langevin Algorithm: in this regard, we could use the KALE flow as a sample-based approximation of the KL flow, which is to our knowledge a novel concept. In future work, we plan to analyze theoretically whether the KALE flow is a consistent estimator of the KL flow in the large sample limit.

References

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

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Lectures in mathematics ETH Zürich. Birkhäuser, Boston, 2005.

[3] M. Arbel, A. Korba, A. SALIM, and A. Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] M. Arbel, D. J. Sutherland, M. Bińkowski, and A. Gretton. On gradient regularizers for MMD GANs. *NIPS*, 2018.

[5] M. Arbel, L. Zhou, and A. Gretton. Generalized energy based models. In *International Conference on Learning Representations*, 2021.

[6] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *ICLR*, 2017.

[7] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 06–11 Aug 2017.

[8] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer distance as a solution to biased Wasserstein gradients, 2017.

[9] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

[10] T. Birdal, M. Arbel, U. Simsekli, and L. J. Guibas. Synchronizing Probability Measures on Rotations via Optimal Transport. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1566–1576, Seattle, WA, USA, June 2020. IEEE.

[11] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[12] L. Bottou, M. Arjovsky, D. Lopez-Paz, and M. Oquab. Geometrical insights for implicit generative modeling. In *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State*, pages 229–268. Springer, 2018.

[13] L. Cayton. Algorithms for manifold learning. Technical report, 2005.

[14] L. Chizat and F. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[15] A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[16] H. Daniels. The asymptotic efficiency of a maximum likelihood estimator. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 151–163. University of California Press Berkeley, 1961.

[17] R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2 edition, 2002.

[18] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of langevin monte carlo via convex optimization. *arXiv preprint arXiv:1802.09188*, 2018.

[19] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *UAI*, 2015.

[20] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouve, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2681–2690. PMLR, 16–18 Apr 2019.

[21] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583. PMLR, 16–18 Apr 2019.

[22] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *AISTATS*, 2018.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. In *NIPS*, 2017.

[25] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[26] B. Jourdain, S. Méléard, and W. Woyczynski. Nonlinear SDE driven by Levy processes and related PDEs. *arXiv preprint arXiv:0707.2723*, 2007.

[27] S. Lang. *Fundamentals of Differential Geometry*. Graduate Texts in Mathematics. Springer-Verlag, New York, 1999.

[28] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2200–2210, Red Hook, NY, USA, 2017. Curran Associates Inc.

[29] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1718–1727. JMLR.org, 2015.

[30] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989.

[31] Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in neural information processing systems*, pages 3115–3123, 2017.

[32] S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and convergence properties of generative adversarial learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[33] Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942 – 1992, 2021.

[34] R. J. McCann. A Convexity Principle for Interacting Gases. *Advances in Mathematics*, 128(1):153–179, June 1997.

[35] H. P. McKean. A class of markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911, 1966.

[36] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[37] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2391–2400. PMLR, 06–11 Aug 2017.

[38] P. Milgrom and I. Segal. Envelope Theorems for Arbitrary Choice Sets. *Econometrica*, 70(2), 2002.

[39] Y. Mroueh, C.-L. Li, T. Sercu, A. Raj, and Y. Cheng. Sobolev gan. 11 2017.

[40] Y. Mroueh and M. Rigotti. Unbalanced sobolev descent. *Advances in Neural Information Processing Systems*, 33, 2020.

[41] Y. Mroueh, T. Sercu, and A. Raj. Sobolev descent. In *AISTATS*, 2019.

[42] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[43] H. Narayanan and S. Mitter. Sample complexity of testing the manifold hypothesis. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[44] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, Nov. 2010.

[45] S. Nowozin, B. Cseke, and R. Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. 06 2016.

[46] F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, June 2000.

[47] L. Pillaud-Vivien, F. Bach, T. Lelièvre, A. Rudi, and G. Stoltz. Statistical estimation of the poincaré constant and application to sampling multimodal distributions. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2753–2763. PMLR, 26–28 Aug 2020.

[48] G. M. Rotskoff and E. Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915*, 2018.

[49] F. Santambrogio. Optimal Transport for Applied Mathematicians – Calculus of Variations, PDEs and Modeling. page 356.

[50] C. E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[51] Z. Shen, Z. Wang, A. Ribeiro, and H. Hassani. Sinkhorn barycenter via functional gradient descent. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 986–996. Curran Associates, Inc., 2020.

[52] C.-J. Simon-Gabriel and B. Schölkopf. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *arXiv:1604.05251 [math, stat]*, Dec. 2019. arXiv: 1604.05251.

[53] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *arXiv preprint arXiv:1808.09372*, 2018.

[54] I. Tolstikhin, B. K. Sriperumbudur, K. Mu, and et. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017.

[55] T. Van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[56] C. Villani. Trend to equilibrium for dissipative equations, functional inequalities and mass transportation. *Contemporary Mathematics*, 353:95, 2004.

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

[58] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv:1707.00087 [math, stat]*, June 2017. arXiv: 1707.00087.

[59] P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the discrimination-generalization tradeoff in GANs. In *International Conference on Learning Representations*, 2018.

[60] D.-X. Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, Oct. 2008.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Yes, we discussed limitations arising from the condition that noise injection should satisfy to ensure global convergence for instance.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is theoretical.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] See appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]