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

Pierre Glaser, Michael Arbel, Arthur Gretton

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

Pierre Glaser
Gatsby Computational Neuroscience Unit
University College London
pierreglaser@gmail.com

Michael Arbel
Université Grenoble Alpes, Inria, CNRS,
Grenoble INP, LJK,38000 Grenoble, France*
michael.n.arbel@gmail.com

Arthur Gretton
Gatsby Computational Neuroscience Unit
University College London
arthur.gretton@gmail.com

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-Leibler (KL) divergence provides a simple way to transport mass towards the target through the Fokker-Planck equation as established in the seminal work of [29]. Its time discretization yields a practical algorithm, the Unadjusted Langevin Algorithm (ULA), which comes with strong convergence guarantees [21, 17]. A more recent gradient flow approach, Stein Variational Gradient Descent (SVGD) [34], also leverages the analytic 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 [54] and its use in Bayesian Inference [12]. 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 [18]. 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.

*Work mostly completed at the Gatsby Unit.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Recent works have considered the gradient flow of Integral Probability Metrics (IPM) [44] instead of the KL, in settings where only samples (and not the density) of the target are known. This includes the Maximum Mean Discrepancy (MMD) [5] and the Kernelized Sobolev Discrepancy (KSD) [43, 41]. One motivation for considering these particle flows is their connection with the training of Generative Adversarial Networks (GANs) [26] using IPMs such as the Wasserstein distance [7, 27, 24], the MMD [22, 32, 31, 9, 10, 4] or the Sobolev discrepancy [42]. As discussed in [43, 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 particular functionals [50]. Thus, analyzing the asymptotic behavior of such flows [38, 56, 16] can ultimately provide convergence guarantees for the training dynamics of neural networks. Establishing such results remains challenging for some classes of IPMs, however, such as the MMD [5].

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 function class [46, 6], 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 makes the optimization problem defining the KALE convex and allows for practical algorithms computing it. (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 [40]. 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 the KALE particle descent algorithm on a set of problems with different geometrical properties. We show empirically 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. The Fenchel dual formulation of the KL, and more generally f-divergences, has a rich history in Machine Learning: [46] relied on this dual formulation to estimate the KL between two probability distributions when their density ratios belong to an approximating class. They derived a plug-in estimator for the KL which comes with convergence guarantees. In the context of GANs, [47] used the Fenchel dual representation of f-divergences, of which the KL is a particular instance, as a GAN critic. Later, [39] 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, [6] 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* [19], unlike the KL which defines a stronger topology [60]. The KALE is therefore well-suited for learning Implicit Generative Models which are only accessible through sampling, as advocated in [8]. 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 46], since our dual problem is strongly convex and comes with guarantees. In parallel to work related to f-divergences, [28, 14, 51, 1] have investigated the task of sampling in the case where the source and the target have disjoint supports. Again, unlike our setting, these works assume that the log-density of the target distribution is known.

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 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 denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures defined on $\mathbb{R}^d$ endowed with its Borelian σ-algebra, and by $\mathcal{P}_2(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ the set of elements of $\mathcal{P}(\mathbb{R}^d)$ with finite second moment. 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 $\mathbb{R}^d$ will be denoted $k : \mathbb{R}^d \times \mathbb{R}^d \longmapsto \mathbb{R}$, with RKHS $\mathcal{H}$. The Dirac delta measure for $x \in \mathbb{R}^d$ will be written δ_x. We denote by $C_c^\infty(\mathbb{R}^d \times (0, +\infty))$ the set of infinitely differentiable functions with compact support on $\mathbb{R}^d \times (0, +\infty)$, and by $C_b^0(\mathbb{R}^d)$ the set of continuous bounded functions from $\mathbb{R}^d$ to $\mathbb{R}$. Sets of N points in $\mathbb{R}^d$ will be indexed using a superscript $\{x^{(i)}\}_{i=1}^N$, while a sequence of points in $\mathbb{R}^d$ 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. For the sake of notational lightness, the choice of the norm used for a specific object (vectors, functions, operators) will be specified with a subscript (e.g. $\|h\|_{\mathcal{H}}$ for the RKHS norm) only if the said choice is not obvious from the context. This remark also holds when referring to the null element of a vector space ($0_{\mathcal{H}}$, $0_{\mathbb{R}^d}$, ...).

2.1 The KL Approximate Lower bound Estimator (KALE)

The central equation to derive the KALE is the (Fenchel) dual formulation of the KL [3, Lemma 9.4.4]:

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

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 an RKHS with kernel k. The Kullback-Leibler Approximate Lower bound Estimator (KALE) is given by:*

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

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 [6], 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.

Mathematical Assumptions To prove the theoretical results stated in this work, we will make the following basic assumptions on the kernel k:

Assumption 1 (Boundedness)**.** *There exists $K > 0$ such that $k(x, x) \leq K$, for all $x \in \mathbb{R}^d$.*

Assumption 2 (Smoothness)**.** *The kernel is 2-times differentiable in the sense of [58, Definition 4.35]: for all $i, j \in \{1, \ldots, d\}$ $\partial_i\partial_{i+d}k$ and $\partial_i\partial_j\partial_{i+d}\partial_{j+d}k$ exist. Moreover, we have: $\|\nabla_1 k_x\|^2 \triangleq \sum_{i=1}^d \|\partial_i k_x\|^2 \leq K_{1d}$ and $\|\boldsymbol{H}_1 k_x\|^2 = \sum_{i,j=1}^d \|\partial_i\partial_j k_x\|^2 \leq K_{2d}$, where d indicates an expected scaling with dimension.*

Assumption 1 guarantees the integrability of the objects intervening in KALE, and implies boundedness of the RKHS functions. Assumption 2 guarantees first and second order smoothness of the RKHS functions, a property invoked to control the KALE flow trajectories. Indeed, both the differential and the hessian of any $f \in \mathcal{H}$ can now be bounded in operator norm: using the Cauchy-Schwarz inequality and the kernel reproducing derivative property [58, Corollary 4.36], we have: $|\partial_i f(x)| \leq \|\partial_i k_x\| \|f\|$ and $|\partial_i\partial_j f(x)| \leq \|\partial_i\partial_j k_x\| \|f\|$, implying $\|\nabla f(x)\| \leq \sqrt{K_{1d}} \|f\|$, and $\|\boldsymbol{H}(f(x))\|_{\mathrm{Op}} \leq \|\boldsymbol{H}(f(x))\|_{\mathrm{F}} \leq \sqrt{K_{2d}} \|f\|$.

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 GANs and Adversarial VAEs: weak continuity, and metrizing the weak convergence of probability distributions. We recall that a functional $\mathcal{D}(\cdot \mid\mid \cdot)$ is a probability divergence if both $\mathcal{D}(\mathbb{P} \mid\mid \mathbb{Q}) \geq 0$ and $\mathcal{D}(\mathbb{P} \mid\mid \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$, for any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$.

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

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

(ii) If k is universal [55], then for any $\lambda > 0$, KALE is a probability divergence. Moreover, KALE metrizes the weak topology between probability measures with finite first order moments.

Central to the proof of all points in this theorem is a link between KALE and the MMD witness function $f_{\mathbb{P},\mathbb{Q}}$, which we report in the next lemma. 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 RKHS norm of the difference of mean embeddings of $\mathbb{P}$ and $\mathbb{Q}$:

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

Lemma 1. *Let $\mathbb{P}$, $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$, and $\mathcal{K} : \mathcal{H} \longmapsto \mathbb{R}$ be the objective maximized by KALE, e.g. $\mathcal{K}(h) = 1 + \int h d\mathbb{P} - \int e^h d\mathbb{Q} - \frac{\lambda}{2}\|h\|^2$. Then, under Assumption 1, $\mathcal{K}$ is Fréchet differentiable. Moreover, the following relationship holds:*

$$\nabla \mathcal{K}(0) = f_{\mathbb{P},\mathbb{Q}}$$

Intuitively, noting that $\mathcal{K}(0) = 0$, Lemma 1 ensures that KALE presents "equivalent" regularity and discriminative properties to those of MMD (a divergence which is itself, under the assumptions of this theorem, weakly continuous and that metrizes the weak convergence of probability distributions). The proof of the second point of Theorem 1 is inspired by [6], which in turn derives from [64, 35], and is adapted to account for the extra norm penalty term in the version of the KALE in this paper.

Interpolating between the MMD and the KL using the KALE The KALE includes a positive regularization parameter λ, inducing two asymptotic regimes: $\lambda \to 0$ and $\lambda \to \infty$. In these regimes, the KALE asymptotically recovers on the one hand the KL divergence, and on the other hand the MMD.

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

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

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

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

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 [5, 23, 13], 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 [3], 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* [45, 13, 15], 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:

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

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 $\mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q})$ in practice As for other IPMs, computing $\mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q})$ for arbitrary $\mathbb{P}$ and $\mathbb{Q}$ is intractable, and is therefore 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. By adapting arguments of [6], it can be shown that the discrepancy between the KALE's empirical and population value, $|\mathrm{KALE}(\widehat{\mathbb{P}}^N \,||\, \widehat{\mathbb{Q}}^N) - \mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q})|$ (often called "sample complexity"), is at most $O(\frac{1}{\sqrt{N}})$. This rate is identical that of Sinkhorn divergences [25], another family of entropically-regularized divergences.

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 the 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 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 $\varphi \in C_c^\infty(\mathbb{R}^d \times (0, +\infty))$, we have:

$$
\int \partial_t \varphi(x,t) \mathrm{d}\mathbb{P}_t \mathrm{d}t + \int \langle \nabla_x \varphi(x,t), V_t \rangle_{\mathbb{R}^d} \mathrm{d}\mathbb{P}_t \mathrm{d}t = 0.
$$

The Wasserstein gradient flow of a functional $\mathcal{F}$ is then obtained by choosing V_t as the gradient of *first variation* of $\mathcal{F}$, defined as the Gâteaux 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 [3, 61, 62]. In the next proposition, we show that the functional $\mathbb{P} \longmapsto \mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q})$ admits a well-defined gradient flow.

Proposition 2 (KALE Gradient Flow). *Let $\lambda > 0$, and $\mathbb{P}_0, \mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumptions 1 and 2, the Cauchy problem*

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

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

$$h_t^\star = \arg\max_{h\in\mathcal{H}} \left\{ 1 + \int h d\mathbb{P}_t - \int e^h d\mathbb{Q} - \frac{\lambda}{2}\|h\|^2 \right\}, \tag{10}$$

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:

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

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 [5] using inequalities of Lojasiewicz type; in particular, the MMD flow is guaranteed to converge provided that the quantity $\mathbb{P}_t - \mathbb{Q}$ remains bounded in the *negative Sobolev distance* $\|\mathbb{P}_t - \mathbb{Q}\|_{\dot{H}^{-1}(\mathbb{P}_t)}$ [48]. We recall that the negative *weighted negative Sobolev distance* [5] between μ and ν is defined as:

$$\|\mu - \nu\|_{\dot{H}^{-1}(\mathbb{P})} = \sup_{\|f\|_{\dot{H}(\mathbb{P})}\leq 1} \left| \int f \mathrm{d}(\mu - \nu) \right|,$$

which is obtained by duality with the weighted Sobolev semi-norm $\|f\|_{\dot{H}(\mathbb{P})} = (\int \|\nabla f\|^2 \mathrm{d}\mathbb{P})^{\frac{1}{2}}$. Note the important role of the latter quantity in the energy dissipation formula of the KALE gradient flow:

$$\frac{\mathrm{d}\text{KALE}(\mathbb{P}_t \mid\mid \mathbb{Q})}{\mathrm{d}t} = -\int (1+\lambda)^2 \|\nabla h^\star\|^2 \mathrm{d}\mathbb{P} = -(1+\lambda)^2 \|h^\star\|^2_{\dot{H}(\mathbb{P})}. \tag{12}$$

In the next proposition, we extend the condition ensuring the global convergence of the MMD flow [5] 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:*

$$KALE(\mathbb{P}_t \mid\mid \mathbb{Q}) \leq \frac{C}{CKALE(\mathbb{P}_0 \mid\mid \mathbb{Q}) + t}.$$

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 [36] and could be an effect of RKHS smoothing.

4 KALE Particle Descent

We now derive 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* [5], 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

$$\mathbb{P}_{n+1} = (I - \gamma(1+\lambda)\nabla h_n^\star)_\# \mathbb{P}_n, \ \mathbb{P}_{n=0} = \mathbb{P}_0. \tag{13}$$

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

$$h_n^\star = \arg\max_{h\in\mathcal{H}} \left\{ 1 + \int h \mathrm{d}\mathbb{P}_n - \int e^h \mathrm{d}\mathbb{Q} - \frac{\lambda}{2} \|h\|^2 \right\}. \tag{14}$$

The solution $\mathbb{P}_n$ of Eq. (13) is a sensible approximation of $\mathbb{P}_t$: indeed, it can be shown under suitable smoothness assumptions [52, 5] 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 $\gamma \to 0$.

Approximation 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. (14) 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 time-discrete iterates $\mathbb{P}_n$ 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 probabilities $\widehat{\mathbb{P}}_n^N$ 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. (13) 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

$$Y_{n+1}^{(i)} = Y_n^{(i)} - \gamma(1+\lambda)\nabla \widehat{h}_n^\star(Y_n^{(i)}), \tag{15}$$

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

$$\widehat{h}_n^\star = \arg\max_{h\in\mathcal{H}} \left\{ \int h d\widehat{\mathbb{P}}_n^N - \int h d\widehat{\mathbb{Q}}^N - \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 \right\}. \tag{16}$$

As in the sample-based setting of Eq. (7), $\widehat{\mathbb{P}}_n^N$ and $\widehat{\mathbb{Q}}^N$ are discrete, meaning that Eq. (16) reduces to an N- dimensional convex problem, and $\widehat{h}_n^\star$ can be tractably computed. The alternate execution of Eq. (15) and Eq. (16) for a finite number of time steps defines the *KALE Particle Descent Algorithm*, that we lay out in Algorithm 1.

Consistency of the KALE Particle Descent Algorithm Note that the 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 [37], that the n-th iterates of the sequence defined by:

$$\bar{Y}_{n+1}^{(i)} = \bar{Y}_n^{(i)} - \gamma(1+\lambda)\nabla h_n^\star(\bar{Y}_n^{(i)}), \ \bar{Y}_0^{(i)} \sim \mathbb{P}_0, \ 1 \leq i \leq N \tag{17}$$

are distributed according to the n^{th} iterate $\mathbb{P}_n$ of the true discrete-time KALE gradient flow solution defined in Eq. (13). As such, the discrete probability $\bar{\mathbb{P}}_n^N = \frac{1}{N}\sum_{i=1}^N \delta_{\bar{Y}_n^{(i)}}$ may be considered as an unbiased space-discretization of Eq. (13). In the next proposition, we show that 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. (17) 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. (15) with the* same *initial conditions $\{Y_0^{(i)}\}_{i=1}^N$. Let $n_{max} \geq 0$. Then, under Assumptions 1 and 2, for all $n \leq n_{\max}$, the following bound holds:*

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

with $A = \sqrt{2KK_{1d}(1+e^{\frac{8K}{\lambda}})} \times \frac{1}{4\sqrt{KK_{1d}}+K_{2d}}$ $B = \frac{(1+\lambda)(4\sqrt{KK_{1d}}+\sqrt{K_{2d}})}{\lambda}$, and K, K_{1d}, K_{2d} are the constants defined in Assumptions 1 and 2.

Proposition 4 shows that given a finite time horizon $n_{\max}$, and given sufficiently 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 (given in Appendix F) relies on the regularity of the KALE witness function $x \longmapsto \widehat{h}_n^\star(x)$, but also on the regularity of the mapping $\widehat{\mathbb{P}}_n^N \longmapsto \widehat{h}_n^\star$ (using the 2-Wasserstein distance as the metric on $\mathcal{P}_2(\mathbb{R}^d)$.

Algorithm 1 KALE Particle Descent Algorithm

Input: $\{Y_0^{(i)}\}_{i=1}^N \sim \mathbb{P}_0$, $\{X^{(i)}\}_{i=1}^N \sim \mathbb{Q}$, `max_iter`, λ, k, γ
Output $\{Y_{\texttt{max\_iter}}^{(i)}\}_{i=1}^N$
for $n = 0$ **to** `max_iter`-1 **do**
  `f_star` $\leftarrow$ `dual_solve`$(X^{(1)}, Y_i^{(1)}, \dots, X^{(N)}, Y_i^{(N)}, k, \lambda)$ `# See Eq.6`
  `h_star` $\leftarrow$ `compute_log_ratio`(`f_star`, $X^{(1)}, Y_i^{(1)}, \dots, X^{(N)}, Y_i^{(N)}, k, \lambda)$ `# Ditto`
  for $j = 1$ **to** N **do**
    `v` $\leftarrow (1+\lambda)$`grad`(`h_star`$(Y_i^{(j)})$)
    $Y_{i+1}^{(j)} \leftarrow Y_i^{(j)} - \gamma \times$ `v`
  end for
end for

4.2 Regularization of KALE particle descent using Noise Injection

In practice, guaranteeing the convergence of the 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 [5, 11] 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 (discrete time) noise injected gradient flow with velocity field $(1+\lambda)\nabla h_n^\star$ and noise schedule β_n is defined as the sequence $(\mathbb{P}_n)_{n\geq 0}$ whose iterates verify:

$$\mathbb{P}_{n+1} = \left((x, u) \longmapsto x - \gamma(1+\lambda)\nabla h_n^\star(x + \beta_n u)\right)_{\#} (\mathbb{P}_n \otimes g), \tag{18}$$

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. (18). Let $(\beta_n)_{n\geq 0}$ be a sequence of noise levels, and define $\mathcal{D}_{\beta_n, \mathbb{P}_n} = \mathbb{E}_{y\sim\mathbb{P}_n, u\sim g} \|\nabla h_n^\star(x + \beta_n u)\|^2$ with g the density of a standard Gaussian distribution. Then, under Assumptions 1 and 2, and for a choice of β_n such that:*

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

the following holds: $\mathit{KALE}(\mathbb{P}_{n+1} \mid\mid \mathbb{Q}) - \mathit{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) \leq -\frac{\gamma}{2}(1 - 3\gamma\sqrt{KK_{2d}})D_{\beta_n, \mathbb{P}_n}(\mathbb{P}_n)$. 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 [5], 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 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 particle descent algorithm, we approximate this convolution using a single sample $U_n^{(i)}$ for each particle update. Eq. (15) becomes:

$$Y_{n+1}^{(i)} = Y_n^{(i)} - \gamma(1+\lambda)\nabla\widehat{h}_n^\star(Y_n^{(i)} + \beta_n U_n^{(i)}), \;\; U_n^{(i)} \sim \mathcal{N}(0, 1). \tag{19}$$

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 particle update rule Eq. (15).

Computing $\widehat{h}_n^\star$ can be achieved using methods such as gradient descent, coordinate descent or higher order optimization methods such as Newton's method and L-BFGS [33].

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 positive 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 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

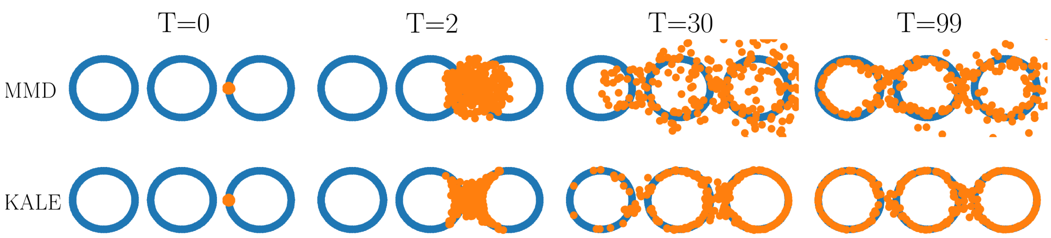


Figure 1: MMD and KALE flow trajectories for "three rings" target

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 each with a finite, positive volume (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" [23]. Results are in Fig. 2.

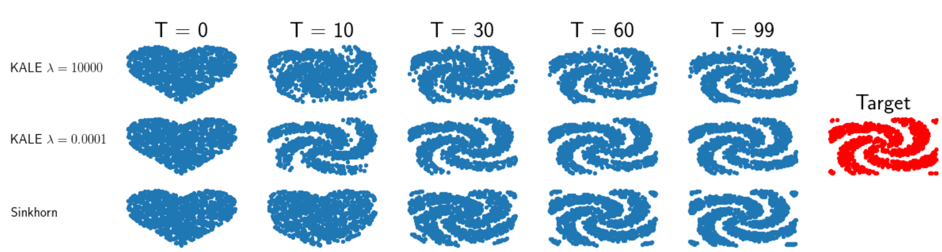
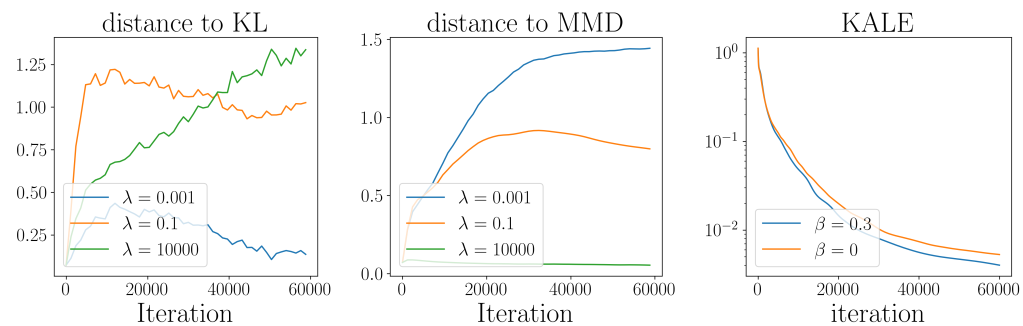


Figure 2: Shape Transfer using the KALE flow

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 positive density on $\mathbb{R}^d$. Hence, unlike in the two previous examples, the KL gradient flow is well-defined, and can be simulated using the Unadjusted Langevin Algorithm (ULA). Echoing the interpolation property of the 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" values of λ, particles from a KALE flow remain close to the ULA particles, while for "large" ones 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 the KALE particle 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 β_n schedule always improves the proximity to the global minimum $\mathbb{P}_\infty = \mathbb{Q}$. Its effect is particularly 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 further details on 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 on several examples that in cases where a KL gradient flow cannot be defined, trajectories of the KALE flow empirically 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. Future work would analyze when the KALE flow is a consistent estimator of the KL flow in the large sample limit.

References

[1] K. Ahn and S. Chewi. Efficient Constrained Sampling via the Mirror-Langevin Algorithm. *arXiv preprint arXiv:2010.16212*, 2020.

[2] S. M. Ali and S. D. Silvey. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1966.

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[4] M. Arbel, D. Sutherland, M. Bińkowski, and A. Gretton. On Gradient Regularizers for MMD GANs. In *Advances in Neural Information Processing Systems*, 2018.

[5] M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum Mean Discrepancy Gradient Flow. In *Advances in Neural Information Processing Systems*, 2019.

[6] M. Arbel, L. Zhou, and A. Gretton. Generalized Energy Based Models. In *International Conference on Learning Representations*, 2021.

[7] M. Arjovsky and L. Bottou. Towards Principled Methods for Training Generative Adversarial Networks. In *International Conference on Learning Representations*, 2017.

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, 2017.

[9] 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. *CoRR*, abs/1705.10743, 2017.

[10] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[11] T. Birdal, M. Arbel, U. Simsekli, and L. J. Guibas. Synchronizing Probability Measures on Rotations via Optimal Transport. In *Conference on Computer Vision and Pattern Recognition*, 2020.

[12] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 2017.

[13] 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*. 2018.

[14] S. Bubeck, R. Eldan, and J. Lehec. Sampling from a Log-concave Distribution with Projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 2018.

[15] L. Cayton. Algorithms for Manifold Learning. *Univ. of California at San Diego Tech. Rep*, 2005.

[16] L. Chizat and F. Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems*, 2018.

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

[18] H. Daniels. The asymptotic efficiency of a maximum likelihood estimator. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1961.

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

[20] A. Durmus and E. Moulines. Nonasymptotic Convergence Analysis for the Unadjusted Langevin Algorithm. *The Annals of Applied Probability*, 2017.

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

[22] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training Generative Neural Networks via Maximum Mean Discrepancy Optimization. In *Uncertainty in Artificial Intelligence*, 2015.

[23] 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 *International Conference on Artificial Intelligence and Statistics*, 2019.

[24] A. Genevay, G. Peyré, and M. Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[25] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample Complexity of Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2014.

[27] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, 2017.

[28] Y. Hsieh, A. Kavis, P. Rolland, and V. Cevher. Mirrored Langevin Dynamics. In *Advances in Neural Information Processing Systems*, 2018.

[29] R. Jordan, D. Kinderlehrer, and F. Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM journal on mathematical analysis*, 1998.

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

[31] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos. MMD GAN: Towards Deeper Understanding of Moment Matching Network. In *Advances in Neural Information Processing Systems*, 2017.

[32] Y. Li, K. Swersky, and R. Zemel. Generative Moment Matching Networks. In *International Conference on Machine Learning*, 2015.

[33] D. C. Liu and J. Nocedal. On the Limited Memory BFGS Method for Large Scale Optimization. *Mathematical programming*, 1989.

[34] Q. Liu. Stein Variational Gradient Descent as Gradient Flow. In *Advances in Neural Information Processing Systems*, 2017.

[35] S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and Convergence Properties of Generative Adversarial Learning. In *Advances in Neural Information Processing Systems*, 2017.

[36] 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*, 2021.

[37] H. P. McKean. A Class of Markov Processes Associated with Nonlinear Parabolic Equations. *Proceedings of the National Academy of Sciences*, 1966.

[38] 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*, 2018.

[39] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks. In *International Conference on Machine Learning*, 2017.

[40] P. Milgrom and I. Segal. Envelope Theorems for Arbitrary Choice Sets. *Econometrica*, 2002.

[41] Y. Mroueh and M. Rigotti. Unbalanced Sobolev Descent. *Advances in Neural Information Processing Systems*, 2020.

[42] Y. Mroueh, C. Li, T. Sercu, A. Raj, and Y. Cheng. Sobolev GAN. In *International Conference on Learning Representations*, 2018.

[43] Y. Mroueh, T. Sercu, and A. Raj. Sobolev Descent. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[44] A. Müller. Integral Probability Metrics and their Generating Classes of Functions. *Advances in Applied Probability*, 1997.

[45] H. Narayanan and S. Mitter. Sample Complexity of Testing the Manifold Hypothesis. In *Advances in Neural Information Processing Systems*, 2010.

[46] 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*, 2010.

[47] S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems*, 2016.

[48] F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 2000.

[49] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. ISBN 978-0-691-01586-6.

[50] 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. *CoRR*, abs/1805.00915, 2018.

[51] A. Salim and P. Richtárik. Primal Dual Interpretation of the Proximal Stochastic Gradient Langevin Algorithm. In *Advances in Neural Information Processing Systems*, 2020.

[52] F. Santambrogio. Optimal Transport for Applied Mathematicians – Calculus of Variations, PDEs and Modeling. 2015.

[53] F. Santambrogio. {Euclidean, metric, and Wasserstein} Gradient Flows: An Overview. *Bulletin of Mathematical Sciences*, 2017.

[54] C. E. Shannon. A Mathematical Theory of Communication. *The Bell system technical journal*, 1948.

[55] C.-J. Simon-Gabriel and B. Schölkopf. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *The Journal of Machine Learning Research*, 2018.

[56] J. Sirignano and K. Spiliopoulos. Mean Field Analysis of Neural Networks: A Central Limit Theorem. *Stochastic Processes and their Applications*, 2020.

[57] B. Sriperumbudur. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 2016.

[58] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[59] I. Tolstikhin, B. K. Sriperumbudur, K. Mu, and et. Minimax Estimation of Kernel Mean Embeddings. *Journal of Machine Learning Research*, 2017.

[60] T. Van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 2014.

[61] C. Villani. Trend to Equilibrium for Dissipative Equations, Functional Inequalities and Mass Transportation. *Contemporary Mathematics*, 2004.

[62] C. Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2008.

[63] J. Weed and F. Bach. Sharp Asymptotic and Finite-sample Rates of Convergence of Empirical Measures in Wasserstein distance. *Bernoulli*, 2019.

[64] 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.

[65] D.-X. Zhou. Derivative Reproducing Properties for Kernel Methods in Learning Theory. *Journal of Computational and Applied Mathematics*, 2008.

Appendix for *KALE flow: A relaxed KL Gradient Flow for Probabilities with Disjoint Support*

The appendix is structured as follows: in Appendix A, we give additional details on the variational formulation of the KL divergence as well as Wasserstein gradient flows. In Appendix B, C, we give the proofs for all statements made about the static properties of the KALE, while in Appendix D to F we give proofs for all statements made about the KALE flow and descent algorithm. Appendix G contains some additional technical lemmas that are used throughout the appendix. Finally, in Appendix H, we provide details on the experiments discussed in the main body, and the impact of noise injection on KALE particle descent trajectories.

A Mathematical Background

In this section, we lay out in more depth the theoretical framework behind the tools used in this paper. We first review the variational formulation of the KL, and more generally f-divergences. We discuss how this variational formulation can be used beyond the context of statistical estimation of the KL, which is the original context it was considered for [46]. We then provide additional details about Wasserstein gradient flows, and the theoretical tools used to study them.

A.1 The use of the variational formulation of f-divergences

f-divergences, first described in [2], form a family of divergences between probability measures parametrized by a convex, lower semi-continuous function f. The divergence D_f between two probabilities measures $\mathbb{P}$ and $\mathbb{Q}$ is defined as:

$$D_f(\mathbb{P} \,||\, \mathbb{Q}) = \begin{cases} \int f(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}})\mathrm{d}\mathbb{Q} & \text{if } \mathbb{P} \ll \mathbb{Q} \\ +\infty & \text{otherwise} \end{cases}$$

Apart from the KL, which we will discuss later, other well known instances of f-divergences include the χ^2 divergence, the Hellinger divergence and the Total Variation. Requiring the function f to be convex allows to use the theory of Fenchel duality to frame D_f as the solution of an optimization problem:

Proposition 6 ([3, Lemma 9.4.4])**.** *For any $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$, we have:*

$$D_f(\mathbb{P} \,||\, \mathbb{Q}) = \sup_{h \in C_b^0(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} h(x) d\mathbb{P} - \int f^\star(h(x)) d\mathbb{Q} \right\} \tag{20}$$

Where $f^\star$ is the Fenchel convex conjugate [49] of the convex function f, defined as:

$$f^\star(u) = \sup_{x \in \mathbb{R}^d} \langle u, x \rangle - f(x)$$

The KL divergence is a particular instance of f-divergence using the pair $(f, f^\star)$:

$$f(x) = \begin{cases} x(\log x - 1) + 1 & \text{if } x > 0 \\ 1 & \text{if } x = 0\,, \\ +\infty & \text{if } x < 0 \end{cases} \qquad f^\star(u) = e^u - 1$$

M-estimation procedures for KL$(\mathbb{P} \,||\, \mathbb{Q})$ The dual formulation in Eq. (20) is an optimization problem with an objective depending on $\mathbb{P}$ and $\mathbb{Q}$ only through *expectations*. By relying on the theory of M-estimation, [46] showed that it was possible to *consistently* approximate the population solution of Eq. (20) using only samples $\{Y^{(i)}\}_{i=1}^N$ and $\{X^{(i)}\}_{i=1}^N$ of $\mathbb{P}$ and $\mathbb{Q}$. In particular, they showed that the solution of the sample-based, regularized problem:

$$\sup_{h \in \mathcal{H}} \left\{ 1 + \int h \mathrm{d}\widehat{\mathbb{P}}^N - \int e^h \mathrm{d}\widehat{\mathbb{Q}}^N + 1 - \frac{\lambda_N}{2} I(h) \right\} \tag{21}$$

(where $I(h)$ is a convex complexity penalty) will converge in probability to the solution of Eq. (20), provided that λ_N decays to 0 as $\frac{1}{\sqrt{N}}$ and that the complexity of the function class $\mathcal{H}$ is small enough. However, their setting is general and does not exploit the specificity of an RKHS $\mathcal{H}$ with a penalty $I(h) = \|h\|_{\mathcal{H}}^2$. Consistency for the latter case (a case which is tightly linked to the definition of the KALE), was proved by [6] using tools from RKHS theory.

Why KALE differs from simple KL estimation The addition of the regularization term $\frac{\lambda_N}{2} I(h)$ (where the KALE objective is retrieved using $I(h) = \|h\|^2$) to Eq. (21) makes the solution of Eq. (20) non-infinite for the mutually singular empirical distributions $\widehat{\mathbb{P}}^N$ and $\widehat{\mathbb{Q}}^N$. However, the KL population objective Eq. (20) is unregularized, reflecting the fact that the KL is infinite for mutually singular population $\mathbb{P}$ and $\mathbb{Q}$. It is the goal of Section 2 is to show that extending the regularization technique introduced in an estimation setting to the KL population objective results in a relaxed solution to the KL problem that is a valid divergence measure between $\mathbb{P}$ and $\mathbb{Q}$. The KALE thus leverages the *biases* of the KL estimates to remain well-defined for mutually singular distributions: in the present context, the primary interest of KALE is not to estimate the KL, but to provide a *KL alternative* for mutually singular distributions. This justifies the definition of the KALE with a positive λ given in Definition 1. Note that a sample-based approximation of KALE($\mathbb{P}$ || $\mathbb{Q}$) is now:

$$\max_{h \in \mathcal{H}} \left\{ 1 + \int h \mathrm{d}\widehat{\mathbb{P}}^N - \int e^h \mathrm{d}\widehat{\mathbb{Q}}^N - \frac{\lambda}{2}\|h\|^2 \right\} \tag{22}$$

We emphasize that unlike in Eq. (21), λ is now kept fixed.

A.2 Wasserstein Gradient Flows

The Wasserstein Geometry The theory of Wasserstein-2 gradient flows considers the set of probability measures on $\mathcal{P}_2(\mathcal{X})$ (where $\mathcal{X}$ is a separable Hilbert Space set to $\mathbb{R}^d$ in our case) with finite 2nd order moments, endowed with the Wasserstein-2 metric, defined, given $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}_2(\mathbb{R}^d)$, as:

$$W_2(\mathbb{P}_0, \mathbb{P}_1) = \left(\inf_{\gamma \in \Gamma(\mathbb{P}_0, \mathbb{P}_1)} \int \|x - y\|^2 \, \mathrm{d}\gamma(x, y) \right)^{\frac{1}{2}} \tag{23}$$

$\Gamma(\mathbb{P}_0, \mathbb{P}_1)$ denotes the sets of *admissible transport plans* between $\mathbb{P}_0$ and $\mathbb{P}_1$:

$$\Gamma(\mathbb{P}_0, \mathbb{P}_1) = \left\{ \gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d); \quad (\pi^1)_{\#}\gamma = \mathbb{P}_0, \ (\pi^2)_{\#}\gamma = \mathbb{P}_1 \right\}$$

where $\pi^1 : (x, y) \longmapsto x$ and $\pi^2(x, y) \longmapsto y$ are the canonical projections on $\mathbb{R}^d \times \mathbb{R}^d$. In the proofs, we will often consider *constant speed geodesics* between two probabilities $\mathbb{P}_0$ and $\mathbb{P}_1$, defined as paths $(\mathbb{P}_t)_{0 \leq t \leq 1}$ of the form:

$$\mathbb{P}_t = \left((1 - t)\pi^1 + t\pi^2 \right)_{\#} \gamma$$

where $\gamma \in \Gamma_o(\mathbb{P}_0, \mathbb{P}_1)$ is an *optimal coupling*, in the sense that it minimizes the objective defining the $W_2(\mathbb{P}_0, \mathbb{P}_1)$ distance in Eq. (23). Convexity along geodesics, or *geodesic convexity* is a property of functionals in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$:

Definition 2 (Geodesic convexity, [3, Definition 9.1.1])**.** *We say that a functional $\mathcal{F}$ is $-M$-geodesically semiconvex for some $M > 0$ if for any $\mathbb{P}_0, \mathbb{P}_1$ and constant speed geodesic $\mathbb{P}_t, \ t \in [0, 1]$ between $\mathbb{P}_0$ and $\mathbb{P}_1$, the following holds:*

$$\mathcal{F}(\mathbb{P}_t) \leq (1 - t)\mathcal{F}(\mathbb{P}_0) + t\mathcal{F}(\mathbb{P}_1) + Mt\,(1 - t)\,W_2(\mathbb{P}_0, \mathbb{P}_1)^2.$$

Wasserstein Gradient Flows The set $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space and not a Hilbert space. Because of that, the notion of gradient (flow) of a functional $\mathcal{F}$ cannot easily be defined through duality with the differential of $\mathcal{F}$, and porting the notion of "gradient flow" to the space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ thus requires characterizing gradient flows trajectories in a Hilbertian-free way. Examples of such characterizations include curves of maximal slope [3, Section 11.1.1], or identification with limit curves of *minimizing moment schemes*. We refer to [53] for an introduction of gradient flows Wasserstein spaces. The formal definition of Wasserstein-2 gradient flows as given in [3] is as follows:

Definition 3 (Gradient Flows [3, Definition 11.1.1])**.** *We say that an absolutely continuous map $\left(t \longmapsto \mathbb{P}_t \in \mathcal{P}_2(\mathbb{R}^d)\right)$ is a solution of the* Wasserstein-2 gradient flow *equation:*

$$\partial_t \mathbb{P}_t + \mathit{div}\,(\mathbb{P}_t v_t) = 0, \tag{24}$$

if $(I \times (-v_t)) \in \boldsymbol{\partial}\mathcal{F}(\mathbb{P}_t)$, where $\boldsymbol{\partial}\mathcal{F}(\mathbb{P}_t)$ is the extended Fréchet subdifferential of $\mathcal{F}$ evaluated at $\mathbb{P}_t$.

For common functionals such as the sum of (sufficiently smooth) potential, interaction and internal energy terms,

$$\mathcal{F}(\mathbb{P}) = \int V(x)d\mathbb{P}(x) + \int W(x-y)d\mathbb{P}(x)d\mathbb{P}(y) + \int f(p(x))d\mathbb{P}(x)$$

(where $\mathbb{P}$ is assumed to be regular, of the form $\mathrm{d}\mathbb{P}(x) = p(x)\mathrm{d}x$), [3] have identified solutions of the very general Eq. (24) with solutions of the more familiar

$$\partial\mathbb{P}_t - \operatorname{div}\left(\mathbb{P}_t\nabla\frac{\delta\mathcal{F}}{\delta\mathbb{P}}(\mathbb{P}_t)\right) = 0, \tag{25}$$

where $\frac{\delta\mathcal{F}}{\delta\mathbb{P}}$ is the first variation of $\mathcal{F}$, defined (when it exists) as the function v verifying:

$$\lim_{\epsilon\to 0}\frac{\mathcal{F}(\mathbb{P}+\epsilon d\chi) - \mathcal{F}(\mathbb{P})}{\epsilon} = \int v(x)\mathrm{d}\chi, \quad \chi = \mathbb{P} - \mathbb{Q}$$

For any $\mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^d)$. Note that this case includes the MMD (given regularity assumptions on the kernel), as discussed in [5], but does not include the KALE, which is not a functional studied in [3], and to our knowledge, a novel object of study in the Wasserstein gradient flow literature. In Appendix D, we show that the identification between Eq. (24) and Eq. (25) still holds for the case of KALE, by identifying elements of its (strong) *extended Fréchet subdifferential*. For completeness, we recall the definition of a strong extended Fréchet subdifferential:

Definition 4 ((Strong) Extended Fréchet subdifferential, [3, Definition 10.3.1]). *Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \longmapsto (-\infty, +\infty]$ be a proper, geodesically convex functional that is lower semicontinuous w.r.t W_2. We say that $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ belongs to the strong extended Fréchet subdifferential $\boldsymbol{\partial}F(\mathbb{P}_0)$ if $\left(\pi^1\right)_\# \gamma = \mathbb{P}_0$, and for every $\mathbb{P}_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\boldsymbol{\mu} \in \Gamma(\gamma, \mathbb{P}_1)$:*

$$\mathcal{F}(\mathbb{P}_1) - \mathcal{F}(\mathbb{P}_0) \geq \int_{X^3}\langle x_2, x_3 - x_1\rangle\, d\boldsymbol{\mu} + o(W_{2,\boldsymbol{\mu}}(\mathbb{P}_0, \mathbb{P}_1))$$

where $W^2_{2,\boldsymbol{\mu}}(\mathbb{P}_0, \mathbb{P}_1) = \int \|x_1 - x_3\|^2\, d\boldsymbol{\mu}(x_1, x_2, x_3)$.

B Proof of Theorem 1

Throughout this proof, we will consider the function $\mathcal{K} : \mathcal{H} \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ given by:

$$\mathcal{K}(h, \mathbb{P}) = \max\left\{1 + \int h\mathrm{d}\mathbb{P} - \int e^h\mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h\|^2\right\} \tag{26}$$

$\mathcal{K}$ has the same expression as the one of Lemma 1, with a supercharged signature to include the dependency in $\mathbb{P}$, which we will use in this proof.

Proof of Lemma 1 The proof follows directly from [6, Lemma 8]. By Assumption 1, all integrability requirements are satisfied (the two *Bochner* integrals in the next equation are well-defined because of Assumption 1). Following this, the gradient of $\mathcal{K}$ is given by:

$$\nabla_h\mathcal{K}(h, \mathbb{P}) = \int k(x,\cdot)\mathrm{d}\mathbb{P} - \int k(x,\cdot)e^h\mathrm{d}\mathbb{Q} - \lambda h.$$

And its evaluation at 0 given in Lemma 1 follows. $\square$

Proof that the KALE is weakly continuous Let $(\mathbb{P}_n)_{n\in\mathbb{N}}$ such that $\mathbb{P}_n$ weakly converges to $\mathbb{P}$. Let $h^\star = \arg\max_h \mathcal{K}(h, \mathbb{P})$ and $h^\star_n = \arg\max_h \mathcal{K}(h, \mathbb{P}_n)$. We will show that both:

$$\limsup_{n\to\infty}\mathcal{K}(h^\star_n, \mathbb{P}_n) = \mathcal{K}(h^\star, \mathbb{P}) \quad \text{and} \quad \liminf_{n\to\infty}\mathcal{K}(h^\star_n, \mathbb{P}_n) = \mathcal{K}(h^\star, \mathbb{P}).$$

The result on KALE follows since when λ is kept fixed, $\mathcal{K}$ and KALE differ only by a multiplicative factor. We focus on proving the first ($\limsup$) equality, the arguments for $\liminf$ being identical.

First, by optimality of $h^\star_n$ w.r.t $\mathbb{P}_n$, we have: $\mathcal{K}(h^\star_n, \mathbb{P}_n) \geq \mathcal{K}(h^\star, \mathbb{P}_n)$, implying

$$\limsup_{n\to\infty}\mathcal{K}(h^\star_n, \mathbb{P}_n) \geq \limsup_{n\to\infty}\mathcal{K}(h^\star, \mathbb{P}_n).$$

Since $\mathbb{P}_n \rightharpoonup \mathbb{P}$, the r.h.s verifies $\limsup_{n\to\infty} \mathcal{K}(h^\star, \mathbb{P}_n) = \lim_{n\to\infty} \mathcal{K}(h^\star, \mathbb{P}_n) = \mathcal{K}(h^\star, \mathbb{P})$, from which we conclude $\limsup_{n\to\infty} \mathrm{KALE}(\mathbb{P}_n, \mathbb{Q}) \geq \mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q})$. To prove the converse, assume that $\limsup_{n\to\infty} \mathrm{KALE}(\mathbb{P}_n, \mathbb{Q}) > \mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q})$. Then there exists $\epsilon > 0$ and a subsequence $n_k \to +\infty$ with $k \to +\infty$ such that $\mathcal{K}(h^\star_{n_k}, \mathbb{P}_{n_k}) \geq \mathcal{K}(h^\star, \mathbb{P}) + \frac{\epsilon}{2}$. Let us now compare $\mathcal{K}(h^\star_{n_k}, \mathbb{P})$ with $\mathcal{K}(h^\star, \mathbb{P})$:

$$\mathcal{K}(h^\star_{n_k}, \mathbb{P}) = \mathcal{K}(h^\star_{n_k}, \mathbb{P}_{n_k}) + \int h^\star_{n_k} \mathrm{d}(\mathbb{P} - \mathbb{P}_{n_k}) \geq \mathcal{K}(h^\star, \mathbb{P}) + \frac{\epsilon}{2} - \frac{4\sqrt{K}\mathrm{MMD}(\mathbb{P} \mid\mid \mathbb{P}_{n_k})}{\lambda}$$

where for the last step, we used the Cauchy-Schwarz inequality and Lemma 5. Since the MMD is weakly continuous for bounded kernels with Lipschitz embeddings [57, Theorem 3.2], we have $\lim_{k\to\infty} \mathrm{MMD}^2(\mathbb{P}_{n_k} \mid\mid \mathbb{P}) = 0$: there exists a k_0 such that, for $k > k_0$, $\mathcal{K}(h^\star_{n_k}, \mathbb{P}) > \mathcal{K}(h^\star, \mathbb{P}) + \frac{\epsilon}{4}$, which contradicts the optimality condition defining $h^\star$. Hence, we must have

$$\limsup_{n\to\infty} \mathrm{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) = \mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}).$$

The two steps of this proof can be repeated for any convergent subsequence of $\mathcal{K}(h^\star_n, \mathbb{P}_n)$, and as a consequence, we also have: $\liminf_{n\to\infty} \mathrm{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) = \mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q})$, which proves the weak continuity of KALE. $\square$

Proof that KALE is a probability divergence that metrizes the weak convergence of probability distributions We first prove positivity and definiteness of KALE, making it a probability divergence. Positivity of KALE comes from the fact that $\mathcal{K}(h^\star, \mathbb{P}) \geq \mathcal{K}(0, \mathbb{P}) = 0$. To prove definiteness of KALE, assume $\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) = 0$. Recall that $\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) = 0 \iff h^\star = 0$, since $\mathcal{K}(0, \mathbb{P}) = 0$ and the objective is strongly convex. The optimality criterion for $0_{\mathcal{H}}$ can be characterized by differentiating $\mathcal{K}(h, \mathbb{P})$. Using Lemma 1, and the optimality of 0, we have:

$$0 = \nabla_h \mathcal{K}(0, \mathbb{P}) \triangleq \int k(x, \cdot)\mathrm{d}\mathbb{P} - \int k(x, \cdot)\mathrm{d}\mathbb{Q} = f_{\mathbb{P},\mathbb{Q}},$$

where $f_{\mathbb{P},\mathbb{Q}}$ denotes the MMD *witness function* between $\mathbb{P}$ and $\mathbb{Q}$, i.e. $\mathrm{MMD}(\mathbb{P} \mid\mid \mathbb{Q})^2 = \|f_{\mathbb{P},\mathbb{Q}}\|^2$. When k is universal, $f_{\mathbb{P},\mathbb{Q}} = 0$ is only possible when $\mathbb{P} = \mathbb{Q}$, which proves the first implication of the equivalence. The reverse implication is proven by noticing that

$$\mathbb{P} = \mathbb{Q} \Longrightarrow \nabla_h \mathcal{K}(0, \mathbb{P}) = 0.$$

Metrizing weak convergence From the weak continuity of KALE associated with the definiteness of KALE proven above, we have $\mathbb{P}_n \rightharpoonup \mathbb{Q} \Longrightarrow \mathrm{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q}) \to 0$. For the converse, assume that $\mathrm{MMD}(\mathbb{P}_n \mid\mid \mathbb{Q})$ doesn't converge to 0. Therefore, there exists a subsequence n_k with $n_k \to +\infty$ when $k \to +\infty$ and such that $\mathrm{MMD}(\mathbb{P}_{n_k} \mid\mid \mathbb{Q}) > c > 0$ for some $c > 0$. Fix $\epsilon > 0$. We have that:

$$\begin{aligned}\mathrm{KALE}(\mathbb{P}_{n_k} \mid\mid \mathbb{Q}) \geq \mathcal{K}(\epsilon \times f_{\mathbb{P}_{n_n},\mathbb{Q}}) &= \left\langle \nabla_h \mathcal{K}(0, \mathbb{P}_{n_k}), \epsilon f_{\mathbb{P}_{n_k},\mathbb{Q}} \right\rangle + \mathcal{O}(\epsilon^2 \|f_{\mathbb{P}_{n_k},\mathbb{Q}}\|) \\ &= \epsilon \|f_{\mathbb{P}_{n_k},\mathbb{Q}}\| + \mathcal{O}(\epsilon^2 \|f_{\mathbb{P}_{n_k},\mathbb{Q}}\|^2).\end{aligned}$$

Now, recall that $\|f_{\mathbb{P}_{n_k},\mathbb{Q}}\| = \mathrm{MMD}(\mathbb{P}_{n_k} \mid\mid \mathbb{Q}) \geq c > 0$, implying that for sufficiently low ϵ, we will have: $\mathrm{KALE}(\mathbb{P}_{n_k} \mid\mid \mathbb{Q}) > \frac{c\epsilon}{2}, \ \forall k \geq n_k$. Thus, $\mathrm{KALE}(\mathbb{P}_n \mid\mid \mathbb{Q})$ does not tend to 0. Hence, by contradiction $\mathrm{MMD}(\mathbb{P}_n \mid\mid \mathbb{Q})$ converges to 0 which implies that $\mathbb{P}_n$ converges weakly to $\mathbb{Q}$ since the MMD metrizes weak convergence. This concludes the proof of Theorem 1. $\square$

C Proof of Proposition 1

C.1 Proof of (i)

To prove that KALE converges to the MMD as λ increases, we will show the following inequalities:

$$\frac{1}{2}\mathrm{MMD}^2(\mathbb{P} \mid\mid \mathbb{Q}) - \mathcal{O}\left(\frac{1}{\lambda}\right) \leq \mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) \leq \frac{1}{2}\mathrm{MMD}^2(\mathbb{P} \mid\mid \mathbb{Q}) + \mathcal{O}\left(\frac{1}{\lambda}\right)$$

To prove the right inequality, we recall that $\mathcal{K}(h, \mathbb{P}) \leq \int h\mathrm{d}\mathbb{P} - \int h\mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h\|^2$, which holds by convexity of the exponential. The right-hand side is maximized for $h^\star = \frac{f_{\mathbb{P},\mathbb{Q}}}{\lambda}$ and equals $\frac{\mathrm{MMD}^2(\mathbb{P}\mid\mid\mathbb{Q})}{2\lambda}$. Consequently, we have: $\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) \leq \frac{1+\lambda}{2\lambda}\mathrm{MMD}^2(\mathbb{P} \mid\mid \mathbb{Q})$.

To prove the left inequality, we use Lemma 5 which allows to control the discrepancy between the KALE and the MMD. Indeed, we have: $h(x) = \langle h, k(x, \cdot) \rangle \leq \sqrt{K} \|h\| = \frac{4K}{\lambda}$. The following Taylor-Lagrange inequality holds, uniformly for all x:

$$e^{h(x)} \leq 1 + h(x) + \frac{e^{\frac{4K}{\lambda}} 16K^2}{2\lambda^2},$$

which gives a lower bound of $\mathcal{K}(h^\star, \mathbb{P})$:

$$(1+\lambda)\,\mathcal{K}(h, \mathbb{P}) \geq (1+\lambda)\left(\int h \mathrm{d}\mathbb{P} - \int h \mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h\|^2 - \frac{8K^2 e^{\frac{4K}{\lambda}}}{\lambda^2}\right).$$

Remark that the r.h.s is maximized for $h_1 = f_{\mathbb{P},\mathbb{Q}}/\lambda$. Because $h^\star$ maximizes the l.h.s, we have:

$$(1+\lambda)\,\mathcal{K}(h^\star, \mathbb{P}) \geq (1+\lambda)\mathcal{K}(h_1, \mathbb{P}) \geq \frac{1+\lambda}{2\lambda}\mathrm{MMD}^2(\mathbb{P} \mid\mid \mathbb{Q}) - \frac{8K^2 e^{\frac{4K}{\lambda}}(1+\lambda)}{\lambda^2}.$$

The two initial inequalities are verified, and taking them to the limit $\lambda \to \infty$ concludes the proof. □

C.2 Proof of (ii)

(ii) was proved in [6] as part of (Theorem 7). For completeness, we recall the elements of the proof. Let us highlight the dependency of $h^\star = \arg\max_h \mathcal{K}(h, \mathbb{P})$ in λ (see Eq. (26)) by noting it $h^\star_\lambda (=h^\star)$, for $\lambda \geq 0$. Because we assume that $\log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} \in \mathcal{H}$, we have:

$$h^\star_0 = \log \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}, \quad \mathrm{KL}(\mathbb{P} \mid\mid \mathbb{Q}) = 1 + \int h^\star_0 \mathrm{d}\mathbb{P} - \int e^{h^\star_0} \mathrm{d}\mathbb{Q}.$$

Thus, we have

$$\begin{aligned}
\Big| \frac{\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q})}{(1+\lambda)} - \mathrm{KL}(\mathbb{P} \mid\mid \mathbb{Q}) \Big| &= \Big| 1 + \int h^\star_\lambda \mathrm{d}\mathbb{P} - \int e^{h^\star_\lambda} \mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h^\star_\lambda\|^2_{\mathcal{H}} - \mathrm{KL}(\mathbb{P} \mid\mid \mathbb{Q}) \Big| \\
&= \Big| \int (h^\star_\lambda - h^\star_0)\, \mathrm{d}\mathbb{P} - \int e^{h_0}(1 - e^{(h^\star_\lambda - h^\star_0)}) \mathrm{d}\mathbb{Q} + \frac{\lambda}{2}\|h^\star_\lambda\|^2 \Big| \\
&\leq \Big| \int (h^\star_\lambda - h^\star_0)\, \mathrm{d}\mathbb{P} \Big| + \Big| \int e^{h_0}(1 - e^{(h^\star_\lambda - h^\star_0)}) \mathrm{d}\mathbb{Q} \Big| + \Big| \frac{\lambda}{2}\|h^\star_\lambda\|^2 \Big|
\end{aligned}$$

To bound the last term, we note that

$$\|h^\star_\lambda\| \leq \|h^\star_0\|. \tag{27}$$

Otherwise, by optimality of $h^\star_0$, we have:

$$\begin{aligned}
&\int h^\star_\lambda \mathrm{d}\mathbb{P} - \int e^{h^\star_\lambda} \mathrm{d}\mathbb{Q} \leq \int h^\star_0 \mathrm{d}\mathbb{P} - \int e^{h^\star_0} \mathrm{d}\mathbb{Q} \\
\implies &\int h^\star_\lambda \mathrm{d}\mathbb{P} - \int e^{h^\star_\lambda} \mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h^\star\|^2_{\mathcal{H}} \leq \int h^\star_0 \mathrm{d}\mathbb{P} - \int e^{h^\star_0} \mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\|h^\star_0\|^2_{\mathcal{H}},
\end{aligned}$$

contradicting the optimality of $h^\star_\lambda$. As a consequence, we have that $\lim_{\lambda \to 0} \frac{\lambda}{2}\|h^\star_\lambda\|^2 (\leq \frac{\lambda}{2}\|h^\star_0\|^2) = 0$. To bound the first two terms, we use [6] (Lemma 11), ensuring that:

$$\lim_{\lambda \to 0} \|h^\star_\lambda - h^\star_0\| = 0. \tag{28}$$

As a consequence:

- For all $x \in \mathbb{R}^d$, $\lim_{\lambda \to 0} h^\star_\lambda(x) - h^\star_0(x) = 0$.
- $h^\star_\lambda$ is a bounded function.

We conclude that the first two terms tend to 0 as $\lambda \to 0$ by the dominated convergence theorem. We thus have: $\lim_{\lambda \to 0} \Big| \mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) - \mathrm{KL}(\mathbb{P} \mid\mid \mathbb{Q}) \Big| = 0$. □

D Proof of Proposition 2

As explained in the introduction, the Wasserstein gradient flow of the KALE does not have a known expression, other than the abstract one given by Definition 3, applied to the KALE. Relying on the formalism introduced in [3], we first show that KALE's gradient flow admits the "traditional" form:

$$\partial_t \mathbb{P}_t - \operatorname{div}\left(\mathbb{P}_t \nabla \frac{\delta \mathrm{KALE}}{\delta \mathbb{P}}\right) = 0$$

We start by giving an expression of the *first variation* of the KALE. This proof is the first in the appendix that involves an implicit function theorem argument, which we justify at length. For brevity, the same justifications will be skipped in other proofs relying on small variations around the same implicit function theorem argument.

Lemma 2 (Differentiability of KALE)**.**
Let $\mathbb{Q} \in \mathcal{P}_2(\mathbb{R}^d)$, and $\lambda > 0$. Then, the function $\mathbb{P} \in \mathcal{P}_2(\mathbb{R}^d) \longmapsto KALE(\mathbb{P} \mid\mid \mathbb{Q})$ is Gâteaux differentiable w.r.t. $\mathbb{P}$ and admits the following first variation:

$$\frac{\delta KALE(\mathbb{P} \mid\mid \mathbb{Q})}{\delta \mathbb{P}} = (1+\lambda)\, h^\star, \quad h^\star = \arg\max_{h \in \mathcal{H}} \mathcal{K}(h, \mathbb{P}).$$

Proof. Informally, computing the first variation of KALE w.r.t $\mathbb{P}$ can be done using a chain rule argument:

$$\frac{\delta \mathrm{KALE}}{\delta \mathbb{P}} = \frac{\delta \mathrm{KALE}(h^\star(\mathbb{P}), \mathbb{P})}{\delta \mathbb{P}} = \frac{\partial \mathrm{KALE}}{\partial \mathbb{P}} + \left.\frac{\partial \mathrm{KALE}}{\partial h}\right|_{h^\star} \frac{\partial h^\star}{\partial \mathbb{P}} = \frac{\partial \mathrm{KALE}}{\partial \mathbb{P}}$$

where the second term is 0 given that $h^\star$ is defined as $\max_{h \in \mathcal{H}} \mathcal{K}(h, \mathbb{P})$. To make this discussion rigorous, we need to make sure that "$\frac{\partial h^\star}{\partial \mathbb{P}}$" (formally, the Gâteaux derivative of the map $\mathbb{P} \longmapsto h^\star(\mathbb{P})$) exists.

We recall that given two topologically convex vector spaces X and Y, and a function $f : X \to Y$, the Gâteaux derivative of f at x in the direction $\chi \in X$ is defined as:

$$Df(x; \chi) = \lim_{t \to 0} \frac{f(x + t\chi) - f(x)}{t}.$$

A complete argument for the differentiability of both $\mathcal{K}$ and $\mathbb{P} \longmapsto h^\star(\mathbb{P})$ would require augmenting the domains of functionals of interest from $\mathcal{P}_2(\mathbb{R}^d)$ (which is not a vector space) by the vector space of signed radon measures $\mathcal{M}(\mathbb{R}^d)$. We circumvent this additional step by simply considering "admissible" directions χ, such that $\int d\chi = 0$. Noting $h_t^\star = \arg\max_h \mathcal{K}(h, \mathbb{P} + t\chi)$, we know given Lemma 1 that $h_t^\star$ verifies:

$$\mathcal{F}_\chi(h_t^\star, t) \triangleq \nabla_h \mathcal{K}(h_t^\star, \mathbb{P}) = \int k(x, \cdot) \mathrm{d}(\mathbb{P}(x) + t\chi(x)) - \int k(x, \cdot) \exp\left(h_t^\star(x)\right) \mathrm{d}\mathbb{Q}(x) - \lambda h_t^\star = 0.$$

Thus, $h_t^\star$ is defined *implicitly* through $\mathcal{K}$'s optimality at $h_t^\star$. To study the differentiability of the mapping $t \longmapsto h_t^\star$, it is natural to rely on an implicit function theorem argument on $\mathcal{F} : (\mathcal{H} \times \mathbb{R}) \to \mathcal{H}$. Similarly to implicit function theorems on euclidean spaces, we will need to invert $D_h \mathcal{F}_\chi(h, t)$, the (Fréchet) differential of $\mathcal{F}$ w.r.t h. This differential is given by:

$$D_h \mathcal{F}_\chi(h, t) = -\underbrace{\int k(x, \cdot) \otimes k(x, \cdot) e^{h(x)} \mathrm{d}\mathbb{Q}(x)}_{\triangleq \boldsymbol{L}(h)} - \lambda \mathrm{I},$$

which is an invertible operator on $\mathcal{H}$, given that $\boldsymbol{L}(h)$ is self-adjoint and positive for all h. We can now apply an implicit function theorem on Banach spaces [30] (Theorem 5.9): For all χ, there exists a neighborhood of 0, $\mathcal{V}(0)$, such that the mapping $t \in \mathcal{V}(0) \longmapsto h_t^\star$ is differentiable. The derivative of $h_t^\star$ at 0 is then the Gâteaux derivative of $h^\star(\mathbb{P})$ in the direction χ:

$$D_{\mathbb{P}} h^\star(\mathbb{P}; \chi) = \int (\boldsymbol{L}(h^\star) + \lambda I)^{-1} k(x, \cdot) d\chi.$$

To conclude on KALE's first variation, we can rigorously write, using the chain rule of Gâteaux derivatives,

$$D_{\mathbb{P}}\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}; \chi) = \int h^{\star}(\mathbb{P})(x) d\chi(x) + \langle \underbrace{\nabla_h \mathcal{K}(h^{\star}(\mathbb{P}), \mathbb{P})}_{=0}, Dh^{\star}(\mathbb{P}; \chi) \rangle_{\mathcal{H}} = \int h^{\star}(\mathbb{P}) d\chi$$

which concludes the proof. $\square$

We now show that the KALE admits strong Fréchet subgradients, and that they are equal to the gradient of KALE's first variation.

Lemma 3. *A coupling γ of the form $(I \times v)_{\#}\mathbb{P}_0$ belongs to the extended (strong) Fréchet subdifferential of KALE at $\mathbb{P} = \mathbb{P}_0$ if and only if $v = \nabla \frac{\delta KALE}{\delta \mathbb{P}} = (1+\lambda)\nabla h_0^{\star}$ $\mathbb{P}_0$-a.e, where $h_0^{\star} = \arg\max_h \mathcal{K}(h, \mathbb{P}_0)$ is the first variation of KALE at $\mathbb{P} = \mathbb{P}_0$.*

Proof. Using an analogue of [3, Equation 10.3.13] for the extended *strong* Fréchet subdifferential, we have that:

$$\gamma = (I \times v)_{\#}\mathbb{P}_0 \in \boldsymbol{\partial}\mathrm{KALE}(\mathbb{P}_0 \mid\mid \mathbb{Q}) \iff$$
$$\mathrm{KALE}(\mathbb{P}_1 \mid\mid \mathbb{Q}) - \mathrm{KALE}(\mathbb{P}_0 \mid\mid \mathbb{Q}) \geq \int (y-x)^{\top} v(x) \mathrm{d}\tilde{\gamma}(x,y) + o(C_2(\tilde{\gamma}))$$

for any $\mathbb{P}_1 \in \mathcal{P}_2(\mathbb{R}^d)$, $\tilde{\gamma} \in \Gamma(\mathbb{P}_0, \mathbb{P}_1)$. Note that without loss of generality, we switched the coupling $\boldsymbol{\mu} \in \Gamma((I \times v)_{\#}\mathbb{P}_0, \mathbb{P}_1)$ present in Definition 4 with a coupling $\tilde{\gamma} \in \Gamma(\mathbb{P}_0, \mathbb{P}_1)$, a switch that is made possible because of the specific form of γ considered above, which is the one needed in Definition 3. Our goal is to show that $(I \times v)_{\#}\mathbb{P}_0 \in \boldsymbol{\partial}\mathrm{KALE}(\mathbb{P}_0 \mid\mid \mathbb{Q}) \iff v = (1+\lambda)\nabla h_0^{\star}$.
We first show the reverse implication, e.g. $(I \times (1+\lambda)\nabla h_0^{\star})_{\#}\mathbb{P}_0 \in \boldsymbol{\partial}\mathrm{KALE}(\mathbb{P}_0 \mid\mid \mathbb{Q})$. To do so, we consider the following interpolation scheme between $\mathbb{P}_0$ and $\mathbb{P}_1$:

$$\mathbb{P}_t = \left(t\pi^2 + (1-t)\pi^1\right)_{\#} \tilde{\gamma}.$$

And note for each $\mathbb{P}_t$, $h_t^{\star} = \arg\max_h \mathcal{K}(h, \mathbb{P}_t)$. Noting $g(t) = \mathrm{KALE}(\mathbb{P}_t \mid\mid \mathbb{Q})$, we have:

$$g'(t) = (1+\lambda) \int (y-x)^{\top} \nabla h_t^{\star}(ty + (1-t)x) \mathrm{d}\tilde{\gamma}(x,y), \quad g''(t) = (1+\lambda)((I) + (II))$$

where

$$(I) = \int (y-x)^{\top} \left(\mathbf{H}h_t^{\star}(ty + (1-t)x)(y-x)\right) \mathrm{d}\tilde{\gamma}(x,y)$$
$$(II) = \int (y-x)^{\top} \left(\nabla \frac{\mathrm{d}h_t^{\star}}{\mathrm{d}t}(ty + (1-t)x)\right) \mathrm{d}\tilde{\gamma}(x,y)$$

(and we exchanged the t-derivative and ∇ in (II)). From Assumption 2 we have that $\|\boldsymbol{H}h\| \leq \|h\| \sqrt{K_{2d}} \leq \frac{4\sqrt{KK_{2d}}}{\lambda}$, implying $(I) \leq \frac{4\sqrt{KK_{2d}}}{\lambda} C_2^2(\tilde{\gamma})$. Using an implicit function theorem argument, we have:

$$\frac{\mathrm{d}h_t^{\star}}{\mathrm{d}t} = -(\boldsymbol{L}(h_t^{\star}) + \lambda I)^{-1}(y-x)^{\top} \nabla_1 k_{ty+(1-t)x},$$

implying

$$(II) = \int \left\langle \sum_{i=1}^{d} (y_i - x_i)\partial_i k_{ty+(1-t)x}, \sum_{i=1}^{d} (y_i - x_i)(\mathbf{L}(h_t^{\star}) + \lambda I)^{-1} \partial_i k_{ty+(1-t)x} \right\rangle d\tilde{\gamma}(x,y)$$
$$\leq \frac{K_{1d}}{\lambda} C_2^2(\tilde{\gamma})$$

where the last line was obtained using the Cauchy-Schwarz inequality on $\mathcal{H}$, RKHS norm homogeneity, the $\frac{1}{\lambda}$-bound on $\|(L + \lambda I)^{-1}\|$, and then the Cauchy-Schwarz inequality on $\mathbb{R}^d$. Using now Taylor's inequality upper bounding the second derivative of g between $t=0$ and $t=1$, we have that:

$$g(1) - g(0) \geq \int (y-x)^{\top} \nabla h_0^{\star}(x) \mathrm{d}\tilde{\gamma}(x,y) + \mathcal{O}(C_2^2(\tilde{\gamma})).$$

Since $\mathcal{O}(C_2^2(\tilde{\gamma})) = o(C_2(\tilde{\gamma}))$, it follows that $(I \times (1+\lambda)\nabla h_0^\star)_\# \mathbb{P}_0 \in \boldsymbol{\partial}\text{KALE}(\mathbb{P}_0 \,||\, \mathbb{Q})$.

To prove the reverse implication, assume $v(\overset{\Delta}{=} (1+\lambda)\tilde{v}) \neq (1+\lambda)\nabla h_0^\star$. Fix $u > 0$, and choose an "adversarial" $\mathbb{P}_{1,u}$ defined as $\mathbb{P}_{1,u} = (x \longmapsto x + u(1+\lambda)(\tilde{v}(x) - \nabla h_0^\star(x)))_\# \mathbb{P}_0$, with an associated coupling $\tilde{\gamma} = (x \times (x \longmapsto x + (1+\lambda)u(\tilde{v}(x) - \nabla h_0^\star(x)))_\# \mathbb{P}_0$. We then have, using a Taylor inequality *lower bounding* the second derivative of g:

$$\begin{aligned} g(1) - g(0) - \int (y-x)^\top (1+\lambda)\tilde{v}(x) d\tilde{\gamma}(x,y) &\leq \int (y-x)^\top (1+\lambda)(\nabla h_0^\star(x) - \tilde{v}(x)) d\tilde{\gamma}(x,y) \\ &\quad + \mathcal{O}(C_2^2(\tilde{\gamma})) \\ &\leq -u(1+\lambda) \int \|\tilde{v}(x) - \nabla h_0^\star(x))\|^2 \,\mathrm{d}\mathbb{P}_0(x) \\ &\quad + \mathcal{O}(C_2^2(\tilde{\gamma})). \end{aligned}$$

In the limit $\mathbb{P}_{1,u} \rightharpoonup \mathbb{P}_0$, e.g. $u \to 0$), the right-hand side scales in u, which is the same scaling as $C_2(\tilde{\gamma}) = (\int \|x_1 - x_2\|^2 d\tilde{\gamma}(x_1, x_2))^{1/2} = u(1+\lambda)(\int \|\tilde{v}(x) - \nabla h_0^\star(x)\|^2 d\mathbb{P}_0(x))^{1/2}$. Thus, it follows that the inequality:

$$g(1) - g(0) - \int (y-x)^\top (1+\lambda)\tilde{v}(x) d\tilde{\gamma}(x,y) \geq o(C_2(\tilde{\gamma}))$$

cannot be verified unless $\tilde{v} = \nabla h_0^\star,\ \mathbb{P}_0$a-e. ☐

We are now ready to make the following claim:

Proposition 7 (KALE's gradient flow)**.** *The Wasserstein-2 KALE's gradient flow of KALE on $\mathcal{P}_2(\mathbb{R}^d)$ follows:*

$$\partial_t \mathbb{P}_t - \mathit{div}\left(\mathbb{P}_t \nabla \frac{\delta \mathit{KALE}}{\partial \mathbb{P}}\right) = 0$$

Proof. This is a direct application of [3, Definition 11.1.1] using the expression of KALE's strong subdifferential of the form $(i \times v)_\# \,\mathbb{P}$. ☐

Now that we identified the expression of the KALE gradient flow, we will show that the KALE gradient flow admits a unique solution. To prove that the KALE gradient flow admits a unique solution is to prove that KALE is $-M$-semiconvex, for some $M > 0$.

Lemma 4. $\mathbb{P} \longmapsto$ *KALE*$(\mathbb{P} \,||\, \mathbb{Q})$ *is* $-\frac{K_{1d} + 4\sqrt{KK_{2d}}}{\lambda}$*-geodesically convex.*

Proof. Let $\mathbb{P}_a, \mathbb{P}_b \in \mathcal{P}_2(\mathbb{R}^d)$, and consider an admissible coupling $\gamma \in \Gamma(\mathbb{P}_a, \mathbb{P}_b)$ with associated transport costs (for various p) $C_p(\gamma) = (\int \|x-y\|^p \,\mathrm{d}\gamma(x,y))^{\frac{1}{p}}$. We consider $(\mathbb{P}_t)_{0 \leq t \leq 1}$ (where $\mathbb{P}_t = \left(t\pi^2 + (1-t)\,\pi^1\right)_\# \gamma)$ a constant-speed geodesic between $\mathbb{P}_a$ and $\mathbb{P}_b$. To prove the geodesic convexity of the KALE, we follow a similar approach as in [16] (Lemma B.2). In particular, we show that $t \longmapsto g(t) = \text{KALE}(\mathbb{P}_t \,||\, \mathbb{Q})$ has an $MC_2^2(\gamma)$-Lipschitz derivative, with some M to be determined. Using a similar implicit function theorem argument as in the proof of Lemma 2, we have:

$$g'(t) = \int (x-y)^\top \nabla h_t^\star(ty + (1-t)x) d\gamma(x,y).$$

Given t_1, t_2, we thus have:

$$|g'(t_1) - g'(t_2)| \leq (I) + (II),$$

where:

$$\begin{aligned} (I) &= \left| \int (x-y)^\top \left(\nabla h_{t_1}^\star(t_1 y + (1-t_1)\,x) - \nabla h_{t_1}^\star(t_2 y + (1-t_2)\,x)\right) d\gamma(x,y) \right| \\ &\leq \left\|h_{t_1}^\star\right\| \sqrt{K_{2d}}(t_2 - t_1) \int \|x-y\|^2 \,\mathrm{d}\gamma(x,y) \leq \frac{4\sqrt{KK_{2d}}}{\lambda}(t_1 - t_2) C_2^2(\gamma) \end{aligned}$$

and:

$$
\begin{aligned}
(II) &= \int (x-y)^\top \left(\nabla h_{t_1}^\star(t_2 y + (1-t_2)\,x) - \nabla h_{t_2}^\star(t_2 y + (1-t_2)\,x)\right) d\gamma(x,y) \\
&= \int \sum_{i=1}^{d} (x_i - y_i) \left\langle h_{t_1}^\star - h_{t_2}^\star, \frac{\partial k_{t_2 y + (1-t_2)x}}{\partial x_i} \right\rangle d\gamma(x,y) \\
&\overset{(i)}{\leq} \int \left\| h_{t_1}^\star - h_{t_2}^\star \right\| \sum_{i=1}^{d} |x_i - y_i| \left\| \frac{\partial k_{t_2 y + (1-t_2)x}}{\partial x_i} \right\| d\gamma(x,y) \\
&\overset{(ii)}{\leq} \sqrt{K_{1d}} \left\| h_{t_1}^\star - h_{t_2}^\star \right\| \int \sqrt{\|x-y\|}^2 d\gamma(x,y) \\
&\overset{(iii)}{\leq} \frac{(t_2 - t_1) K_{1d} C_2^2(\mathbb{P}_a, \mathbb{P}_b)}{\lambda}
\end{aligned}
$$

where (i) follows from Cauchy-Schwarz on $\mathcal{H}$, (ii) uses Cauchy-Schwarz on $\mathbb{R}^d$ and (iii) relies on Lemma 8 and Jensen inequality. We thus conclude that $g'(t)$ is $MC_2^2(\gamma)$-Lipschitz, with $M = \frac{K_{1d} + 4\sqrt{KK_{2d}}}{\lambda}$, and thus that KALE is $-M$-geodesically semiconvex □

The geodesic convexity of the KALE allows to conclude the proof of Proposition 2: indeed, since the KALE is geodesically semiconvex in $\mathbb{P}$, and admits strong extended Fréchet subdifferentials, we conclude that the KALE gradient flow solutions exist and are unique, as guaranteed by [3, Theorem 11.2.1]. □

E Proof of Proposition 3

We recall the following definitions: given a positive measure $\mathbb{P}$, and a function $f \in \mathcal{C}^1(\mathbb{R}^d)$, the weighted Sobolev *semi-norm* of f is given by:

$$
\|f\|_{\dot{H}(\mathbb{P})} = \left(\int \|\nabla f\|^2 \, \mathrm{d}\mathbb{P} \right)^{\frac{1}{2}}.
$$

Note the important role of the weighted Sobolev semi-norm in the energy dissipation formula of KALE's gradient flow:

$$
\frac{\mathrm{dKALE}(\mathbb{P}_t \,||\, \mathbb{Q})}{\mathrm{d}t} = -\int (1+\lambda)^2 \|\nabla h^\star\|^2 \, \mathrm{d}\mathbb{P}_t = -(1+\lambda)^2 \|h^\star\|^2_{\dot{H}(\mathbb{P}_t)}. \tag{29}
$$

By duality, one can define the (possibly infinite) negative *weighted negative Sobolev distance* [5] between μ and ν:

$$
\|\mu - \nu\|_{\dot{H}^{-1}(\mathbb{P})} = \sup_{\|f\|_{\dot{H}(\mathbb{P})} \leq 1} \left| \int f \mathrm{d}(\mu - \nu) \right|.
$$

As proven in [48], the weighted negative Sobolev distance linearizes the Wasserstein distance, and one can formally write:

$$
W_2(\mu, \mu + d\mu) = \|d\mu\|_{\dot{H}^{-1}(\mathbb{P})} + o(d\mu).
$$

Moreover, for all $f \in \mathcal{C}^1(\mathbb{R}^d)$, and $\mu \in \mathcal{M}(\mathbb{R}^d)$, one has:

$$
\int f \mathrm{d}\mu \leq \|f\|_{\dot{H}(\mathbb{P})} \|\mu\|_{\dot{H}^{-1}(\mathbb{P})}. \tag{30}
$$

To prove Proposition 3, we use the λ-strong concavity of $\mathcal{K}(h, \mathbb{P})$ w.r.t. h :

$$
\begin{aligned}
\mathrm{KALE}(\mathbb{P} \,||\, \mathbb{Q}) = (1+\lambda)\mathcal{K}(h^\star, \mathbb{P}) &\leq (1+\lambda)(\mathcal{K}(0, \mathbb{P}) + \langle h^\star, \nabla_h \mathcal{K}(0, \mathbb{P}) \rangle - \frac{\lambda}{2} \|h^\star\|^2) \\
&\leq (1+\lambda) \langle h^\star, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle = (1+\lambda) \int h^\star(x) \mathrm{d}\mathbb{P} - \int h^\star(x) \mathrm{d}\mathbb{Q} \\
&\leq (1+\lambda) \|h\|_{\dot{H}(\mathbb{P})} \|\mathbb{P} - \mathbb{Q}\|_{\dot{H}^{-1}(\mathbb{P})} \leq (1+\lambda) C \|h\|_{\dot{H}(\mathbb{P})}.
\end{aligned}
$$

Here we successively applied Eq. (30) and the hypothesis $\|\mathbb{P}-\mathbb{Q}\|_{\dot{H}(\mathbb{P})} \le C$. Recalling Eq. (29), one has:

$$\frac{d\mathrm{KALE}(\mathbb{P}_t,\mathbb{Q})}{\mathrm{d}t} \le -\frac{\mathrm{KALE}(\mathbb{P}_t\,||\,\mathbb{Q})^2}{C^2} \implies \frac{d(1/\mathrm{KALE}(\mathbb{P}_t,\mathbb{Q}))}{\mathrm{d}t} \ge \frac{1}{C},$$

from which the desired inequality follows. □

Proof of Proposition 5 We rely on the proof technique used in [5, E.1]. From Lemma 7, we get that assumptions A, D of [5] hold with $L=\sqrt{KK_{2d}}$ and $\lambda^2=K_{2d}$. Moreover, we know from Lemma 5 that $h^\star$ is $\frac{4K}{\lambda}$-Lipschitz. From these smoothness conditions, all steps in [5, E.1], follow until:

$$\mathrm{KALE}(\mathbb{P}_{n+1}\,||\,\mathbb{Q}) - \mathrm{KALE}(\mathbb{P}_n\,||\,\mathbb{Q}) \le -\gamma\left(1-\frac{3}{2}\gamma\sqrt{KK_{2d}}\right)\mathcal{D}_{\beta_n}(\mathbb{P}_n) + \gamma\sqrt{K_{2d}}\beta_n\|h^\star\|\,\mathcal{D}_{\beta_n}(\mathbb{P}_n)^{\frac{1}{2}}.$$

Now, given that $\|h^\star\|^2 \le \frac{2\mathrm{KALE}(\mathbb{P}_n,\mathbb{Q})}{\lambda}$ and that $\frac{8K_{2d}\beta_n^2}{\lambda^2}\mathrm{KALE}(\mathbb{P}_n,\mathbb{Q}) \le \mathcal{D}_{\beta_n}(\mathbb{P}_n)$ we have:

$$\begin{aligned}
\mathrm{KALE}(\mathbb{P}_{n+1}\,||\,\mathbb{Q}) - \mathrm{KALE}(\mathbb{P}_n\,||\,\mathbb{Q}) &\le -\gamma\left(1-\frac{3}{2}\gamma\sqrt{KK_{2d}}\right)\mathcal{D}_{\beta_n}(\mathbb{P}_n) + \gamma\sqrt{\frac{2}{8}}D_{\beta_n}(\mathbb{P}_n) \\
&\le -\frac{\gamma}{2}\left(1-3\gamma\sqrt{KK_{2d}}\right)\mathcal{D}_{\beta_n}(\mathbb{P}_n) \\
&\overset{(iv)}{\le} -4\gamma\left(1-3\gamma\sqrt{KK_{2d}}\right)\frac{K_{2d}}{\lambda^2}\beta_n^2\mathrm{KALE}(\mathbb{P}\,||\,\mathbb{Q}) \\
&\overset{(v)}{\le} -\Gamma\beta_n^2\mathrm{KALE}(\mathbb{P}_n\,||\,\mathbb{Q}),
\end{aligned}$$

where (iv) uses the noise schedule assumption and in (v) we noted $\Gamma = 4\gamma\left(1-3\gamma\sqrt{KK_{2d}}\right)\frac{K_{2d}}{\lambda^2}$, and the result follows as in [5].

F Proof of Proposition 4

We recall the update equations defining the trajectories $(Y_n^{(i)})_{n\le n_{\max}}$ and $(\bar{Y}_n^{(i)})_{n\le n_{\max}}$:

$$\begin{aligned}
Y_{n+1}^{(i)} &= Y_n^{(i)} - \gamma(1+\lambda)\nabla\widehat{h}_n^\star(Y_n^{(i)}), \\
\bar{Y}_{n+1}^{(i)} &= \bar{Y}_n^{(i)} - \gamma(1+\lambda)\nabla h_n^\star(\bar{Y}_n^{(i)}).
\end{aligned} \tag{31}$$

We denote $c_n = \sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|\bar{Y}_n^{(i)} - Y_n^{(i)}\right\|^2}$. Note that

$$\mathbb{E}W_2(\overline{\mathbb{P}}_n^N,\widehat{\mathbb{P}}_n^N)^2 \le \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|Y_{n+1}^{(i)} - \bar{Y}_{n+1}^{(i)}\right\|^2\right] = c_n^2.$$

The iterates c_n satisfy the following recursion:

$$\begin{aligned}
c_{n+1} &= \sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|Y_{n+1}^{(i)} - \bar{Y}_{n+1}^{(i)}\right\|^2\right]} \\
&\le \sqrt{\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\left\|Y_n^{(i)} - \bar{Y}_n^{(i)} - \gamma(1+\lambda)\left(\nabla\widehat{h}_n^\star(Y_n^{(i)}) - \nabla h_n^\star(\bar{Y}_n^{(i)})\right)\right\|^2\right]} \\
&\le c_n + \underbrace{\frac{\gamma(1+\lambda)}{\sqrt{N}}\sqrt{\sum_{i=1}^N \mathbb{E}\left[\left\|\nabla\widehat{h}_n^\star(Y_n^{(i)}) - \nabla h_n^\star(\bar{Y}_n^{(i)})\right\|^2\right]}}_{\triangleq A}.
\end{aligned}$$

Using a triangular inequality, we now split (A) into terms that will be handled differently:

$$c_{n+1} \leq c_n + \gamma(1+\lambda)\left(\underbrace{\frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^N \mathbb{E}\left[\left\|\nabla \widehat{h}_n^\star(Y_n^{(i)}) - \nabla \widehat{h}_n^\star(\bar{Y}_n^{(i)})\right\|^2\right]}}_{(i)}\right.$$

$$\left. + \underbrace{\frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^N \mathbb{E}\left[\left\|\nabla \widehat{h}_n^\star(\bar{Y}_n^{(i)}) - \nabla \bar{h}_n^\star(\bar{Y}_n^{(i)})\right\|^2\right]}}_{(ii)} + \underbrace{\frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^N \mathbb{E}\left[\left\|\nabla \bar{h}_n^\star(\bar{Y}_n^{(i)}) - \nabla h_n^\star(\bar{Y}_n^{(i)})\right\|^2\right]}}_{(iii)}\right)$$

Where we introduced the notation $\bar{h}_n^\star = \arg\max_h \mathcal{K}(h, \overline{\mathbb{P}}_n^N)$, the witness function that estimates the *true* witness function $h_n^\star$ using $\overline{\mathbb{P}}_n^N$, the empirical version of $\mathbb{P}_n$, instead of $\mathbb{P}_n$. Let us explain the source of each of the terms in the last inequality:

- (i) comes from evaluating the velocity field $\widehat{h}_n^\star$ at different points $Y_n^{(i)}$ and $\bar{Y}_n^{(i)}$,
- (ii) comes from using *biased* samples $\{Y_n^{(i)}\}_{i=1}^N$ to compute $\widehat{h}_n^\star$, and unbiased samples $\{\bar{Y}_n^{(i)}\}_{i=1}^N$ to compute $\bar{h}_n^\star$.
- (iii) comes from the use of a finite number of unbiased samples to compute $\bar{h}_n^\star$.

After controlling (i), (ii), (iii), as detailed below, we get the following upper bound:

$$c_{n+1} \leq c_n\gamma(1+\lambda)\left(1 + \frac{4\sqrt{KK_{2d}} + K_{2d}}{\lambda}\right) + \frac{\gamma(1+\lambda)}{\lambda}\sqrt{\frac{KK_{2d}(1+e^{\frac{8K}{\lambda}})}{N}}.$$

We use [5, Lemma 26] to conclude:

$$c_n = \sqrt{\frac{2KK_{1d}(1+e^{\frac{8K}{\lambda}})}{N}} \times \frac{1}{4\sqrt{KK_{1d}} + K_{2d}}(e^{\gamma(1+\lambda)\frac{4\sqrt{KK_{1d}}+K_{2d}}{\lambda}n} - 1).$$

The result on $\mathbb{E}W_2(\bar{\mathbb{P}}_n, \widehat{\mathbb{P}}_n)$ follows by noting that $\mathbb{E}W_2(\bar{\mathbb{P}}_n^N, \widehat{\mathbb{P}}_n^N) \leq \sqrt{\mathbb{E}W_2^2(\bar{\mathbb{P}}_n^N, \widehat{\mathbb{P}}_n^N)}$ by Jensen's inequality. □

F.1 Control of the 3 error terms

Controlling (i) To control the first term, we rely on the RKHS derivative reproducing property [65]: $\frac{\partial h}{\partial x_i} = \langle \partial_i k_x, h \rangle$, Assumption 2, and on the uniform bound on $\|h^\star\|$ (for all $\mathbb{P}$, $\mathbb{Q}$) given by (Lemma 5) :

$$\left\|\nabla \widehat{h}_n^\star(Y_n^{(i)}) - \nabla \widehat{h}_n^\star(\bar{Y}_n^{(i)})\right\|^2 \leq \sum_{i=1}^d \left\|\partial_i k_{Y_n^{(i)}} - \partial_i k_{\bar{Y}_n^{(i)}}\right\|^2 \left\|\hat{h}_n\right\|^2 = \frac{16KK_{2d}}{\lambda^2}\left\|Y_n^{(i)} - \bar{Y}_i^{(n)}\right\|^2.$$

Consequently, we have

$$(i) = \frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^N \mathbb{E}\left\|\nabla \widehat{h}_n^\star(Y_n^{(i)}) - \nabla \widehat{h}_n^\star(\bar{Y}_n^{(i)})\right\|^2} \leq \frac{4\sqrt{KK_{2d}}}{\lambda\sqrt{N}}c_n.$$

Controlling (ii) To control (ii), we rely on Lemma 8, that guarantees that KALE($\mathbb{P}$ || $\mathbb{Q}$) is $\frac{\sqrt{K_{1d}}}{\lambda}$-Lipschitz in $\mathbb{P}$ and $\mathbb{Q}$, when $\mathcal{P}(\mathbb{R}^d)$ is endowed with the Wasserstein-2 metric:

$$\left\|\nabla \widehat{h}_n^\star(\bar{Y}_n^{(i)}) - \nabla \bar{h}_n^\star(\bar{Y}_n^{(i)})\right\|^2 = \sum_{j=1}^{d} \left(\partial_j \widehat{h}_n^\star(\bar{Y}_n^{(i)}) - \partial_j \bar{h}_n^\star(\bar{Y}_n^{(i)})\right)^2$$
$$\leq K_{1d} \left\|\hat{h}_n^\star - \bar{h}_n^\star\right\|^2.$$

Consequently, using Lemma 8, we have:

$$\left\|\nabla \widehat{h}_n^\star(\bar{Y}_n^{(i)}) - \nabla \bar{h}_n^\star(\bar{Y}_n^{(i)})\right\|^2 \leq \frac{K_{1d}^2}{\lambda^2} W_2(\widehat{\mathbb{P}}_n^N, \bar{\mathbb{P}}_n^N)^2$$

$$\implies (ii) = \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^{N} \mathbb{E}\left\|\nabla \widehat{h}^\star(\bar{Y}_n^{(i)}) - \nabla h^\star(\bar{Y}_n^{(i)})\right\|^2} \leq \frac{K_{1d}}{\lambda\sqrt{N}} \sqrt{\mathbb{E} W_2^2(\widehat{\mathbb{P}}_n^N, \bar{\mathbb{P}}_n^N)} \leq \frac{K_{1d}}{\lambda\sqrt{N}} c_n.$$

Controlling (iii) In (iii), the witness function $\bar{h}_n^\star$ is an empirical version of $h_n^\star$. Repeating the first lines of (ii), we have:

$$\left\|\nabla \bar{h}_n^\star(\bar{x}_n^{(i)}) - \nabla h_n^\star(\bar{x}_n^{(i)})\right\|^2 \leq K_{1d} \left\|\bar{h}_n^\star - h_n^\star\right\|^2.$$

We could use the bound given in (ii) to get a bound on $\left\|\bar{h}_n^\star - h_n^\star\right\|$, but the sample complexity of the Wasserstein distances scales in $\mathcal{O}(n^{-1/d})$, which is much slower than our target rate $1/\sqrt{N}$ [63]. Instead, we rely on the concentration inequality given by Lemma 6, ensuring that $\mathbb{E}\left\|\bar{h}_n^\star - h_n^\star\right\|^2 \leq \frac{2K(1+e^{\frac{8K}{\lambda}})}{N\lambda^2}$. Following this, we have:

$$(iii) = \frac{1}{\sqrt{N}} \sqrt{\sum_{i=1}^{N} \mathbb{E}\left\|\nabla \bar{h}_n^\star(\bar{x}_n^{(i)}) - \nabla h^\star(\bar{x}_n^{(i)})\right\|^2} \leq \frac{1}{\lambda} \sqrt{\frac{2K K_{1d}(1+e^{\frac{8K}{\lambda}})}{N}}.$$

G Auxiliary Lemmas

Lemma 5 (Uniform smoothness of the KALE witness function)**.** *Under Assumption 1, and for all $\mathbb{P}$, $\mathbb{Q}$, the following inequalities hold:*

$$\frac{\lambda}{2}\left\|h^\star\right\|^2 \leq \mathit{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) \leq 2\sqrt{K}\left\|h^\star\right\|,$$

implying $\|h^\star\| \leq \frac{4\sqrt{K}}{\lambda}$. We also have the finer estimate $\|h^\star\| \leq \frac{2MMD(\mathbb{P}||\mathbb{Q})}{\lambda}$.

Proof. The right inequality follows from the proof of Proposition 3. Indeed, we have:

$$\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) \leq \langle h^\star, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle \leq \|h^\star\| \left(\|\mu_{\mathbb{P}}\| + \|\mu_{\mathbb{Q}}\|\right) \leq 2\sqrt{K}\left\|h^\star\right\|.$$

The left inequality can be noticed using KALE's *dual formulation* Eq. (6)

$$\begin{aligned}\mathrm{KALE}(\mathbb{P} \mid\mid \mathbb{Q}) &= \underbrace{\int (f^\star(\log f^\star - 1) + 1)\,\mathrm{d}\mathbb{Q}}_{\geq 0} + \frac{1}{2\lambda}\left\|\int f^\star(x)\mathrm{d}\mathbb{Q}(x) - \mu_{\mathbb{P}}\right\|^2 \\ &\geq \frac{1}{2\lambda}\left\|\int f^\star(x)\mathrm{d}\mathbb{Q}(x) - \mu_{\mathbb{P}}\right\|^2 = \frac{\lambda}{2}\left\|h^\star\right\|^2.\end{aligned} \tag{32}$$

To get the finer estimate, we keep track of $\frac{\lambda}{2}\left\|h^\star\right\|_{\mathcal{H}}^2$ term. By convexity of $\exp$, we have:

$$\underbrace{1 + \int h\mathrm{d}\mathbb{P} - \int e^h \mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\left\|h\right\|^2}_{\mathcal{K}(h,\mathbb{P})} \leq \int h\mathrm{d}\mathbb{P} - \int h\mathrm{d}\mathbb{Q} - \frac{\lambda}{2}\left\|h\right\|^2.$$

Recalling now that $\mathcal{K}(h^\star, \mathbb{P}) \geq \mathcal{K}(0, \mathbb{P}) = 0$, we must have:

$$\int h^\star \mathrm{d}\mathbb{P} - \int h^\star \mathrm{d}\mathbb{Q} - \frac{\lambda}{2} \|h^\star\|^2 \geq 0$$
$$\implies \|h^\star\| \leq \frac{2\|f_{\mathbb{P},\mathbb{Q}}\|}{\lambda}$$

Where the last line used the Cauchy-Schwarz inequality. ☐

Lemma 6. *Under Assumption 1, and using the notations of Appendix F, we have:*

$$\mathbb{E}\left\|\bar{h}_n^\star - h_n^\star\right\|^2 \leq \frac{2K(1+e^{\frac{8K}{\lambda}})}{N\lambda^2}$$

Proof. We first notice, as explained in [6] (Proposition 12), that $\left\|\bar{h}_n^\star - h_n^\star\right\| \leq \frac{1}{\lambda}\left\|\nabla\widehat{\mathcal{L}}(h_n^\star) - \nabla\mathcal{L}(h_n^\star)\right\|$ where $\mathcal{L} = 1 + \int h \mathrm{d}\mathbb{P} - \int e^h \mathrm{d}\mathbb{Q}$ is the KL objective, and $\widehat{\mathcal{L}}(h) = \int h \mathrm{d}\bar{\mathbb{P}}_n^N - \int e^h \mathrm{d}\widehat{\mathbb{Q}}^N + 1$ is its empirical equivalent. We then use [59] (Proposition A.1, notice that their statement also holds for $(\mathbb{E}\|\int r\mathrm{d}\mathbb{P}_n - \int r \mathrm{d}\mathbb{P}\|^2)^{1/2}$), to get:

$$\begin{aligned}
E\left\|\nabla\widehat{\mathcal{L}}(h_n^\star) - \nabla\mathcal{L}(h_n^\star)\right\|^2 \leq & \mathbb{E}\left\|\int k(x,\cdot)\mathrm{d}\mathbb{P}_n - \int k(x,\cdot)d\bar{\mathbb{P}}_n^N\right\|^2 \\
& + \mathbb{E}\left\|\int k(x,\cdot)e^{h^\star_n}\mathrm{d}\mathbb{Q} - \int k(x,\cdot)e^{h^\star_n}\mathrm{d}\widehat{\mathbb{Q}}^N\right\|^2 \\
& \leq \frac{K(1+e^{\frac{8K}{\lambda}})}{N},
\end{aligned}$$

where we used the Cauchy-Schwarz inequality on $\mathcal{H}$ and Lemma 5 to bound the squared norm of $x \longmapsto k(x,\cdot)e^{h^\star(x)}$. ☐

Lemma 7. *Under Assumption 2, The maps $x \longmapsto k_x (\overset{\Delta}{=} k(x,\cdot))$ and $x \longmapsto \nabla k_x$ are differentiable. Moreover, we have*

$$\|k_x - k_y\| \leq \sqrt{K_{1d}}\,\|x-y\|$$
$$\|\nabla k_x - \nabla k_y\| \leq \sqrt{K_{2d}}\,\|x-y\|$$

Proof. We prove the differentiability and the Lispchitzness property for the map $x \longmapsto k_x$; the arguments can be straightforwardly adapted to the case of $x \longmapsto \nabla k_x$. To prove the differentiability, we build upon [58, Lemma 4.34], that guarantees that $x \longmapsto k(x,\cdot)$ admits partial derivatives for all i, noted $\partial_i\phi(x)$. We finish the proof by construction: let $D\phi(x) : \mathbb{R}^d \longmapsto \mathcal{H}$ our candidate differential, defined as $D\phi(x)(\Delta) = \sum_{i=1}^d \Delta_i \partial_i \phi(x)$ for all $\Delta \in \mathbb{R}^d$. We show that $D\phi(x)$ is the differential of ϕ at x using a simple telescopic argument: let us note $(x+\Delta)_{:i} = (x_1+\Delta_1, \ldots, x_i+\Delta_i, x_{i+1}, \ldots, x_d)$ for any $i \in \{0, \ldots, d\}$ with $(x+\Delta)_{:0} = x$ by convention. Then:

$$\phi(x+\Delta) - \phi(x) = \sum_{i=d}^{1} \phi(x+\Delta)_{:i} - \phi((x+\Delta)_{:i-1})$$

Knowing that $\phi((x+\Delta)_{:i}) - \phi((x+\Delta)_{:(i-1)}) = \partial_i\phi((x+\Delta)_{:i-1})\Delta_i + o(|\Delta_i|)$, we have:

$$\phi(x+\Delta) - \phi(x) - D\phi(x)(\Delta) = \sum_{i=1}^d \left(\partial_i\phi((x+\Delta)_{:i-1}) - \partial_i\phi(x)\right)\Delta_i + o(|\Delta_i|)$$
$$\implies \|\phi(x+\Delta) - \phi(x) - D\phi(x)(\Delta)\| \leq \sum_{i=1}^d |\Delta_i|\,\|\partial_i\phi((x+\Delta)_{:i-1}) - \partial_i\phi(x)\| + o(\|\Delta\|_1)$$

From [58, Lemma 4.34], we have: that:

$$\|\partial_i\phi((x+\Delta)_{:i-1}) - \partial_i\phi(x)\|^2 = A - B$$

where

$$A = \partial_i \partial_{i+d} k((x+\Delta)_{:i-1}, (x+\Delta)_{:i-1}) - \partial_i \partial_{i+d} k((x+\Delta)_{:i-1}, x))$$
$$B = \partial_i \partial_{i+d} k((x+\Delta)_{:i-1}, x) - \partial_i \partial_{i+d} k(x, x)$$

Since $\partial_i \partial_{i+d} k(x, x')$ is continuous, both A and B tend to 0 as $\|\Delta\|$ tends to 0. Thus, we have:

$$\|\phi(x+\Delta) - \phi(x) - D\phi(x)(\Delta)\| \leq \sum_{i=1}^{d} o(|\Delta_i|) + o(\|\Delta\|_1) = o(\|\Delta\|_2)$$

by equivalency of $\|\cdot\|_1$ and $\|\cdot\|_2$ in $\mathbb{R}^d$.

Lipschitzness is guaranteed by bounding the operator norm of $D\phi(x)$:

$$D\phi(x) = \sup_{\|\Delta\|=1} \|D\phi(x)\Delta\| \leq \sum_{i=1}^{n} |\Delta_i| \, \|\partial_i \phi(x)\| \leq \sqrt{\|\Delta\|^2} \sqrt{\sum_{i=1}^{n} \|\partial_i \phi(x)\|^2} = \sqrt{K_{1d}}$$

□

Lemma 8. *For any $\mathbb{P}_0$, $\mathbb{P}_1 \in \mathcal{P}_2(\mathbb{R}^d)$, with associated KALE witness functions $h_0^\star, h_1^\star$, we have:*

$$\|h_1^\star - h_0^\star\|^2 \leq \frac{K_{1d}}{\lambda^2} W_2(\mathbb{P}_0, \mathbb{P}_1)^2.$$

Proof. The optimal functions $h_0^\star$ and $h_1^\star$ are characterized by the following optimality condition:

$$\int k(x, \cdot) \mathrm{d}\mathbb{P} - \int k(x, \cdot) e^{h^\star} \mathrm{d}\mathbb{Q} - \lambda h^\star = 0.$$

Let us now pose $\mathrm{d}\mathbb{P}_t = \mathrm{d}\mathbb{P}_0 + t d\chi$ with $\mathrm{d}\chi = \mathrm{d}\mathbb{P}_1 - \mathrm{d}\mathbb{P}_0$, and its associated witness function $h_t^\star$. Using an implicit function theorem argument [30] between t and $h_t^\star$, we can write, with notations of Appendix D:

$$\frac{dh_t^\star}{dt} = (\boldsymbol{L}(h_t^\star) + \lambda I)^{-1} \int k(x, \cdot)(\mathrm{d}\mathbb{P}_1 - \mathrm{d}\mathbb{P}_0).$$

The operator $\boldsymbol{L}$ is the covariance operator of the measure $\tilde{\mathbb{Q}} = e^{h^\star}\mathbb{Q}$. This operator is compact given that k is bounded by Assumption 1. Using the spectral theorem on Hilbert spaces, we know that there exists a complete orthonormal system of eigenvectors of $\boldsymbol{L}$, with associated eigenvalues $\{\mu_{i,t}\}_{i \in \mathbb{N}}$ for any t. The operator $(\boldsymbol{L} + \lambda I)^{-1}$ admits an identical eigendecomposition, with eigenvalues $\left\{\frac{1}{\lambda + \mu_{i,t}}\right\}_{i \in \mathbb{N}}$: thus, the operator norm of $(\boldsymbol{L} + \lambda I)^{-1}$ is upper-bounded by $1/\lambda$. We can thus extract a bound on $\|h_1^\star - h_0^\star\|^2$:

$$\begin{aligned}
\|h_1^\star - h_0^\star\|^2 &= \left\| \int_0^1 (\boldsymbol{L}(h_t^\star) + \lambda I)^{-1} \left(\int k(x, \cdot) \left(\mathrm{d}\mathbb{P}_1 - \mathrm{d}\mathbb{P}_0\right) \right) \mathrm{d}t \right\|^2 \\
&\leq \int_0^1 \left\| (L(h_t^\star) + \lambda I) \int k(x, \cdot) \left(\mathrm{d}\mathbb{P}_1 - \mathrm{d}\mathbb{P}_0\right) \right\|^2 \mathrm{d}t \\
&\leq \int_0^1 \frac{1}{\lambda^2} \left\| \int k(x, \cdot) \left(\mathrm{d}\mathbb{P}_1 - \mathrm{d}\mathbb{P}_0\right) \right\|^2 \mathrm{d}t = \frac{1}{\lambda^2} \left\| \int k(x, \cdot) \left(\mathrm{d}\mathbb{P}_1 - \mathrm{d}\mathbb{P}_0\right) \right\|^2.
\end{aligned}$$

Now, let $\nu \in \Gamma(\mathbb{P}_1, \mathbb{P}_0)$. Then one has:

$$\int k(x, \cdot) \left(\mathrm{d}\mathbb{P}_0 - \mathrm{d}\mathbb{P}_1\right) = \int \left(k(x, \cdot) - k(y, \cdot)\right) d\nu(x, y)$$

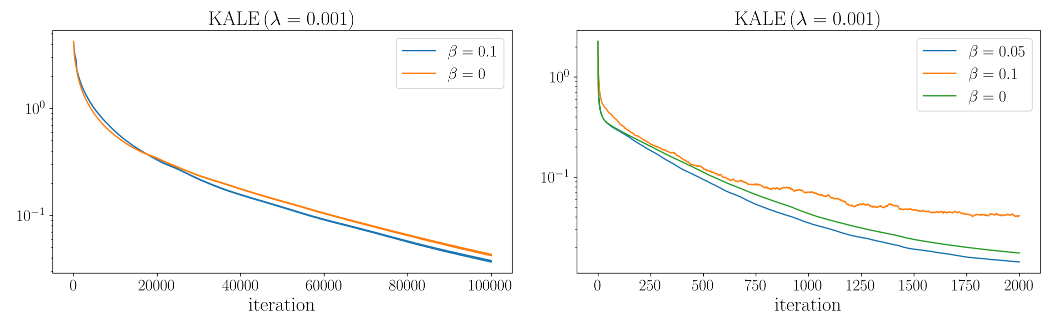
$$\begin{aligned}
\left\| \int k(x, \cdot) \left(\mathrm{d}\mathbb{P}_0 - \mathrm{d}\mathbb{P}_1\right) \right\|^2 &\leq \int \|k(x, \cdot) - k(y, \cdot)\|^2 \mathrm{d}\nu(x, y) \\
&\leq K_{1d} \int \|x - y\|^2 \mathrm{d}\nu(x, y) = K_{1d} W_2(\mathbb{P}_0, \mathbb{P}_1)^2
\end{aligned}$$

Where we applied first Jensen's inequality and Lemma 7. □

H Details of Numerical Experiments and Impact of Noise Injection

In this section, we provide further details on the experiments in the main paper. The step size γ used for the KALE particle descent algorithm scales with λ as $\min(0.1, \frac{\lambda}{10})$. For all experiments, we used a Gaussian kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$. The kernel width σ is described for each experiment set.

"Three rings" experiments For this experiment, the number of particles in each distribution was $N = 300$, and we used the Newton algorithm to compute the KALE. We used a kernel width $\sigma = 0.3$. We show in Fig. 4a the impact of noise injection with a constant noise schedule of $\beta_n = 0.3$.



(a) "Three rings" (b) "Shape transfer"

Figure 4: Impact of noise injection on the KALE value during a KALE particle descent algorithm.

"Shape transfer" experiments For this experiment, we used artificial data from the same source as [41]. We sub-sampled both shapes to $N = 2000$ points, and used a kernel width of $\sigma = 0.3$, as well as $\lambda = 0.001$. Because the number of particles is higher in that case, we used a coordinate descent algorithm to compute KALE, that has a complexity in $\mathcal{O}(N^2)$. We show in Fig. 4b the impact of noise injection with a constant noise schedule of $\beta_n = 0.05$. For this experiment, we also show empirically that while using a small amount of noise *lowers* the final KALE value when compared to the unregularized KALE flow, a too large noise level $\beta_n = 0.1$ results in a *larger* final KALE value. We hypothesize that that noise schedule did not respect the assumptions made in Proposition 5.

"Mixture of Gaussians" experiments For this experiment, we used $N = 240$ particles for each distribution, and a standard deviation of 0.25 for each target Gaussian. We used the Unadjusted Langevin Algorithm [20] to simulate a KL gradient flow with step size 0.001, and the MMD particle descent algorithm of [5] to simulate a MMD flow with step size 0.001. For both the MMD and the KALE, we used the same Gaussian kernel with kernel width $\sigma = 0.35$. We show the impact of noise injection for the KALE flow with a constant noise schedule $\beta_n = 0.3$ to regularize KALE flow with $\lambda = 0.001, 0.1$ and 10000.

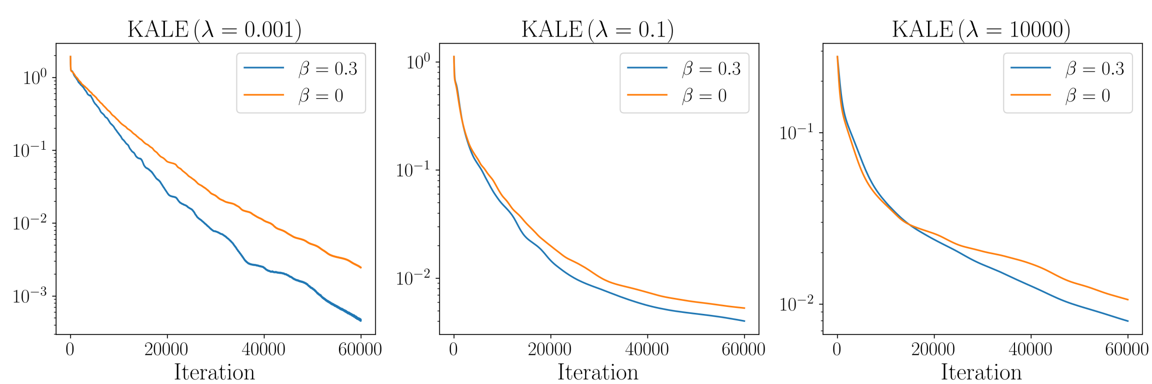


Figure 5: Impact of noise injection: Mixture of Gaussians experiments