[50] I. Steinwart and A. Christmann. *Support Vector Machines*. 1st. Springer Publishing Company, Incorporated, 2008.

[51] D. J. Sutherland, H. Strathmann, M. Arbel, and A. Gretton. “Efficient and principled score estimation with Nyström kernel exponential families.” In: *AISTATS* (2018).

[52] J. Tugaut. “Phase transitions of McKean–Vlasov processes in double-wells landscape.” In: *Stochastics An International Journal of Probability and Stochastic Processes* 86.2 (2014), pp. 257–284.

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

[54] C. Villani. *Topics in Optimal Transportation*. en. Google-Books-ID: R_nWqjq89oEC. American Mathematical Soc., 2003.

[55] C. Villani. “Trend to equilibrium for dissipative equations, functional inequalities and mass transportation.” In: *Contemporary Mathematics* 353 (2004), p. 95.

This appendix is organized as follows. In Appendix A, the mathematical background needed for this paper is given and mainly concerns kernel and optimal transport theory. In Appendix B, we discuss connexions with other gradient flows in the literature. In Appendix C, we state the assumptions on the kernel on which we rely for the proofs. Appendix D is dedicated to the construction of the gradient flow of the MMD. Appendix E is dedicated to the proofs of the convergence results provided in Section 3. Appendix F is dedicated to the modified gradient flow based on noise injection. Proofs of convergence are developed and a pseudocode is provided. The proofs of many results rely on some preliminary results which are given Appendix G.

A Mathematical background

A.1 Maximum Mean Discrepancy and Reproducing Kernel Hilbert Spaces

We recall here fundamental definitions and properties of reproducing kernel Hilbert spaces (RKHS) (see [48]) and Maximum Mean Discrepancies (MMD). Given a positive semi-definite kernel $(x,y) \mapsto k(x,y) \in \mathbb{R}$ defined for all $x, y \in \mathcal{X}$, we denote by $\mathcal{H}$ its corresponding RKHS (see [48]). The space $\mathcal{H}$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}}$ and corresponding norm $\|.\|_{\mathcal{H}}$. A key property of $\mathcal{H}$ is the reproducing property: for all $f \in \mathcal{H}, f(x) = \langle f, k(x,.)\rangle_{\mathcal{H}}$. Moreover, if k is m-times differentiable w.r.t. each of its coordinates, then any $f \in \mathcal{H}$ is m-times differentiable and $\partial^\alpha f(x) = \langle f, \partial^\alpha k(x,.)\rangle_{\mathcal{H}}$ where α is any multi-index with $\alpha \leq m$ [50, Lemma 4.34]. When k has at most quadratic growth, then for all $\mu \in \mathcal{P}_2(\mathcal{X})$, $\int k(x,x)\,\mathrm{d}\mu(x) < \infty$. In that case, for any $\mu \in \mathcal{P}_2(\mathcal{X})$, $\phi_\mu := \int k(.,x)\,\mathrm{d}\mu(x)$ is a well defined element in $\mathcal{H}$ called the mean embedding of μ. The kernel k is said to be characteristic when such mean embedding is injective, that is any mean embedding is associated to a unique probability distribution. When k is characteristic, it is possible to define a distance between distributions in $\mathcal{P}_2(\mathcal{X})$ called the Maximum Mean Discrepancy:

$$MMD(\mu,\nu) = \|\phi_\mu - \phi_\nu\|_{\mathcal{H}} \qquad \forall\, \mu,\nu \in \mathcal{P}_2(\mathcal{X}). \tag{22}$$

The difference between the mean embeddings of μ and ν is an element in $\mathcal{H}$ called the witness function between μ and ν: $f_{\mu,\nu} = \phi_\nu - \phi_\mu$. The MMD can also be seen as an *Integral Probability Metric*:

$$MMD(\mu,\nu) = \sup_{g\in\mathcal{B}} \int g\,\mathrm{d}\mu - \int g\,\mathrm{d}\nu \tag{23}$$

where $\mathcal{B} = \{g \in \mathcal{H} : \ \|g\|_{\mathcal{H}} \leq 1\}$ is the unit ball in the RKHS.

A.2 2-Wasserstein geometry

Let $\mathcal{P}_2(\mathcal{X})$ the set of probability distributions on $\mathcal{X}$ with finite second moment. For two given probability distributions ν and μ in $\mathcal{P}_2(\mathcal{X})$, we denote by $\Pi(\nu,\mu)$ the set of possible couplings between ν and μ. In other words $\Pi(\nu,\mu)$ contains all possible distributions π on $\mathcal{X}\times\mathcal{X}$ such that if $(X,Y)\sim\pi$ then $X\sim\nu$ and $Y\sim\mu$. The 2-Wasserstein distance on $\mathcal{P}_2(\mathcal{X})$ is defined by means of an optimal coupling between ν and μ in the following way:

$$W_2^2(\nu,\mu) := \inf_{\pi\in\Pi(\nu,\mu)} \int \|x-y\|^2 d\pi(x,y) \qquad \forall \nu,\mu \in \mathcal{P}_2(\mathcal{X}) \tag{24}$$

It is a well established fact that such optimal coupling π^* exists. Moreover, it can be used to define a path $(\rho_t)_{t\in[0,1]}$ between ν and μ in $\mathcal{P}_2(\mathcal{X})$. For a given time t in $[0,1]$ and given a sample (x,y) from π^*, it possible to construct a sample z_t from ρ_t by taking the convex combination of x and y: $z_t = s_t(x,y)$ where s_t is given by:

$$s_t(x,y) = (1-t)x + ty \qquad \forall x,y\in\mathcal{X},\ \forall t\in[0,1]. \tag{25}$$

The function s_t is well defined since $\mathcal{X}$ is a convex set. More formally, ρ_t can be written as the projection or push-forward of the optimal coupling π^* by s_t:

$$\rho_t = (s_t)_{\#}\pi^* \tag{26}$$

We recall that for any $T : \mathcal{X} \to \mathcal{X}$ a measurable map, and any $\rho \in \mathcal{P}(\mathcal{X})$, the push-forward measure $T_{\#}\rho$ is characterized by:

$$\int_{y\in\mathcal{X}} \phi(y) d(T_{\#}\rho)(y) = \int_{x\in\mathcal{X}} \phi(T(x)) d\rho(x) \text{ for every measurable function } \phi. \tag{27}$$

It is easy to see that (26) satisfies the following boundary conditions at $t = 0, 1$:

$$\rho_0 = \nu \qquad \rho_1 = \mu. \tag{28}$$

Paths of the form of (26) are called *displacement geodesics*. They can be seen as the shortest paths from ν to μ in terms of mass transport ([44] Theorem 5.27). It can be shown that there exists a *velocity vector field* $(t, x) \mapsto v_t(x)$ with values in $\mathbb{R}^d$ such that ρ_t satisfies the continuity equation:

$$\partial_t \rho_t + div(\rho_t v_t) = 0 \qquad \forall t \in [0, 1]. \tag{29}$$

This equation expresses two facts, the first one is that $-div(\rho_t v_t)$ reflects the infinitesimal changes in ρ_t as dictated by the vector field (also referred to as velocity field) v_t, the second one is that the total mass of ρ_t does not vary in time as a consequence of the divergence theorem. Equation (29) is well defined in the distribution sense even when ρ_t does not have a density. At each time t, v_t can be interpreted as a tangent vector to the curve $(\rho_t)_{t\in[0,1]}$ so that the length $l((\rho_t)_{t\in[0,1]})$ of the curve $(\rho_t)_{t\in[0,1]}$ would be given by:

$$l((\rho_t)_{t\in[0,1]})^2 = \int_0^1 \|v_t\|^2_{L_2(\rho_t)} \,\mathrm{d}t \quad \text{where} \quad \|v_t\|^2_{L_2(\rho_t)} = \int \|v_t(x)\|^2 \,\mathrm{d}\rho_t(x) \tag{30}$$

This perspective allows to provide a dynamical interpretation of the W_2 as the length of the shortest path from ν to μ and is summarized by the celebrated Benamou-Brenier formula ([5]):

$$W_2(\nu, \mu) = \inf_{(\rho_t, v_t)_{t\in[0,1]}} l((\rho_t)_{t\in[0,1]}) \tag{31}$$

where the infimum is taken over all couples ρ and v satisfying (29) with boundary conditions given by (28). If $(\rho_t, v_t)_{t\in[0,1]}$ satisfying (29) and (28) realizes the infimum in (31), is simply called a geodesic between ν and μ; moreover it is called a constant-speed geodesic if the norm of v_t is constant for all $t \in [0, 1]$. In consequence, (26) is a constant-speed displacement geodesic.

Remark 1. *Such paths should not be confused with another kind of paths called mixture geodesics. The mixture geodesic $(m_t)_{t\in[0,1]}$ from ν to μ is obtained by first choosing either ν or μ according to a Bernoulli distribution of parameter t and then sampling from the chosen distribution:*

$$m_t = (1 - t)\nu + t\mu \qquad \forall t \in [0, 1]. \tag{32}$$

Paths of the form (32) can be thought as the shortest paths between two distributions when distances on $\mathcal{P}_2(\mathcal{X})$ are measured using the MMD (see [8] Theorem 5.3). We refer to [8] for an overview of the notion of shortest paths in probability spaces and for the differences between mixture geodesics and displacement geodesics. Although, we will be interested in the MMD as a loss function, we will not consider the geodesics that are naturally associated to it and we will rather consider the displacement geodesics defined in (26) for reasons that will become clear in Appendix A.4 and Appendix E.

Linearization of the W_2. Given a probability distribution ν, the *weighted Sobolev semi-norm* is defined for all squared integrable functions f in $L_2(\nu)$ as $\|f\|_{\dot{H}(\nu)} = \left(\int \|\nabla f(x)\|^2 \,\mathrm{d}\nu(x)\right)^{\frac{1}{2}}$ with the convention $\|f\|_{\dot{H}(\nu)} = +\infty$ if f does not have a square integrable gradient. The *Negative weighted Soboelv distance* $\|.\|_{\dot{H}^{-1}(\nu)}$ is then defined on distributions as the dual norm of $\|.\|_{\dot{H}(\nu)}$. Interestingly, $\|.\|_{\dot{H}^{-1}(\nu)}$ linearizes the W_2 distance (see [54, Theorem 7.26]).

A.3 Gradient flows on the space of probability measures

Consider a functional over the space of distributions:

$$\begin{aligned} \mathcal{F} \colon \mathcal{P}(\mathcal{X}) &\to \mathbb{R} \cup \infty \\ \nu &\mapsto \mathcal{F}(\nu). \end{aligned}$$

We call $\frac{\partial \mathcal{F}}{\partial \nu}$ if it exists, the unique (up to additive constants) function such that $\frac{d}{d\epsilon}\mathcal{F}(\nu + \epsilon\chi)_{\epsilon=0} = \int \frac{\partial \mathcal{F}}{\partial \nu}(\nu) d\chi$ for every perturbation $\chi \in \mathcal{P}_2(\mathcal{X})$ such that, at least for ϵ small enough, the measure $\nu + \epsilon\chi$ belongs to $\mathcal{P}_2(\mathcal{X})$. The function $\frac{\partial \mathcal{F}}{\partial \nu}$ is called first variation of the functional $\mathcal{F}$ at ν. A celebrated class of functionals over the space of probability measures $\mathcal{P}(\mathcal{X})$, called free energies, are of the form:

$$\mathcal{F}(\nu) = \int U(\nu(x))\nu(x)dx + \int V(x)\nu(x)dx + \int W(x,y)\nu(x)\nu(y)dxdy \tag{33}$$

where U is the internal energy, V the potential (or confinement) energy and W the interaction energy. The formal gradient flow equation associated to such a functional can be written (see [9], Lemma 8 to 10):

$$\frac{\partial \nu}{\partial t} = div(\nu \nabla \frac{\partial \mathcal{F}}{\partial \nu}) = div(\nu \nabla (U'(\nu) + V + W * \nu)) \tag{34}$$

where div is the divergence operator and $\nabla \frac{\partial \mathcal{F}}{\partial \nu}$ is the strong subdifferential of $\mathcal{F}$ associated with the W_2 metric (see [1], Lemma 10.4.1). Indeed, for some generalized notion of gradient ∇_{W_2}, and for sufficiently regular ν and $\mathcal{F}$, the r.h.s. of (34) corresponds to $-\nabla_{W_2}\mathcal{F}(\nu)$. The dissipation of energy along the flow is then given by (see [55]):

$$\frac{d\mathcal{F}(\nu)}{dt} = -D(\nu) \quad \text{with } D(\nu) = \int |\nabla \frac{\partial \mathcal{F}(\nu(x))}{\partial \nu}|^2 \nu(x)dx \tag{35}$$

Standard considerations from fluid mechanics tell us that the continuity equation (34) may be interpreted as the equation ruling the evolution of the density ν_t of a family of particles initially distributed according to some ν_0, and each particle follows the velocity vector field $V_t = \nabla \frac{\partial \mathcal{F}}{\partial \nu_t}(\nu_t)$.

A.4 Displacement convexity

Just as for Euclidian spaces, an important criterion to characterize the convergence of the Wasserstein gradient flow of a functional $\mathcal{F}$ is given by displacement convexity (see[55, Definition 16.5])):

Definition 2. *[Displacement convexity] We say that a functional $\nu \mapsto \mathcal{F}(\nu)$ is displacement convex if for any ν and ν' and a constant speed geodesic $(\rho_t)_{t\in[0,1]}$ between ν and ν' with velocity vector field $(v_t)_{t\in[0,1]}$ as defined by (29), the following holds:*

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_0) + t\mathcal{F}(\nu_1) \qquad \forall\, t \in [0,1]. \tag{36}$$

Definition 2 can be relaxed to a more general notion of convexity called Λ-displacement convexity (see [53, Definition 16.5]). We first define an admissible functional Λ:

Definition 3. *[Admissible Λ functional] A functional $(\rho, v) \mapsto \Lambda(\rho, v) \in \mathbb{R}$ defined for any probability distribution $\rho \in \mathcal{P}_2(\mathcal{X})$ and v any square integrable vector field in $L_2(\rho)$ is admissible, if it satisfies:*

- *For any $\rho \in \mathcal{P}_2(\mathcal{X})$, $v \mapsto \Lambda(\rho, v)$ is a quadratic form on $L_2(\mathcal{X}, \mathcal{X}, \rho)$.*
- *For any minimizing geodesic $(\rho_t)_{0\leq t\leq 1}$ between two distributions ν and ν' with corresponding vector fields $(v_t)_{t\in[0,1]}$ it holds that $\inf_{0\leq t\leq 1} \Lambda(\rho_t, v_t)/\|v_t\|^2_{L_2(\rho_t)} > -\infty$*

We can now define the notion of Λ-convexity:

Definition 4. *[Λ convexity] We say that a functional $\nu \mapsto \mathcal{F}(\nu)$ is Λ-convex if for any $\nu, \nu' \in \mathcal{P}_2(\mathcal{X})^2$ and a constant speed geodesic $(\rho_t)_{t\in[0,1]}$ between ν and ν' with velocity vector field $(v_t)_{t\in[0,1]}$ as defined by (29), the following holds:*

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_0) + t\mathcal{F}(\nu_1) - \int_0^1 \Lambda(\rho_s, v_s)G(s,t)ds \qquad \forall\, t \in [0,1]. \tag{37}$$

where $(\rho, v) \mapsto \Lambda(\rho, v)$ satisfies Definition 3, and $G(s,t) = s(1-t)\mathbb{I}\{s \leq t\} + t(1-s)\mathbb{I}\{s \geq t\}$. A particular case is when $\Lambda(\rho, v) = \lambda \int \|v(x)\|^2 \, \mathrm{d}\rho(x)$ for some $\lambda \in \mathbb{R}$. In that case, (37) becomes:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu_0) + t\mathcal{F}(\nu_1) - \frac{\lambda}{2}t(1-t)W_2^2(\nu_0, \nu_1) \qquad \forall\, t \in [0,1]. \tag{38}$$

Definition 2 is a particular case of Definition 4, where in (38) one has $\lambda = 0$.

B Related Work

B.1 Connection with Neural Networks

In this sub-section we establish a formal connection between the MMD gradient flow defined in (5) and neural networks optimization in the limit of infinitely many neurons based on the formulation in [43]. To remain consistent with the rest of the paper, the parameters of a network will be denoted by $x \in \mathcal{X}$ while the input and outputs will be denoted as z and y. Given a neural network or any parametric function $(z, x) \mapsto \psi(z, x)$ with parameter $x \in \mathcal{X}$ and input data z we consider the supervised learning problem:

$$\min_{(x_1,\dots,x_m)\in\mathcal{X}} \frac{1}{2}\mathbb{E}_{(y,z)\sim p}\left[\left\|y - \frac{1}{m}\sum_{i=1}^{m}\psi(z, x_i)\right\|^2\right] \tag{39}$$

where $(y, z) \sim p$ are samples from the data distribution and the regression function is an average of m different networks. The formulation in (39) includes any type of networks. Indeed, the averaged function can itself be seen as one network with augmented parameters $(x_1, ..., x_m)$ and any network can be written as an average of sub-networks with potentially shared weights. In the limit $m \to \infty$, the average can be seen as an expectation over the parameters under some probability distribution ν. This leads to an expected network $\Psi(z, \nu) = \int \psi(z, x)\, \mathrm{d}\nu(x)$ and the optimization problem in (39) can be lifted to an optimization problem in $\mathcal{P}_2(\mathcal{X})$ the space of probability distributions:

$$\min_{\nu\in\mathcal{P}_2(\mathcal{X})} \mathcal{L}(\nu) := \mathbb{E}_{(y,z)\sim p}\left[\left\|y - \int \psi(z, x)\, \mathrm{d}\nu(x)\right\|^2\right] \tag{40}$$

For convenience, we consider $\bar{\mathcal{L}}(\nu)$ the function obtained by subtracting the variance of y from $\mathcal{L}(\nu)$, i.e.: $\bar{\mathcal{L}}(\nu) = \mathcal{L}(\nu) - var(y)$. When the model is well specified, there exists $\mu \in \mathcal{P}_2(\mathcal{X})$ such that $\mathbb{E}_{y\sim\mathbb{P}(.|z)}[y] = \int \psi(z, x)\, \mathrm{d}\mu(x)$. In that case, the cost function $\bar{\mathcal{L}}$ matches the functional $\mathcal{F}$ defined in (3) for a particular choice of the kernel k. More generally, as soon as a global minimizer for (40) exists, Proposition 10 relates the two losses $\bar{\mathcal{L}}$ and $\mathcal{F}$.

Proposition 10. *Assuming a global minimizer of (40) is achieved by some $\mu \in \mathcal{P}_2(\mathcal{X})$, the following inequality holds for any $\nu \in \mathcal{P}_2(\mathcal{X})$:*

$$\left(\bar{\mathcal{L}}(\mu)^{\frac{1}{2}} + \mathcal{F}^{\frac{1}{2}}(\nu)\right)^2 \geq \bar{\mathcal{L}}(\nu) \geq \mathcal{F}(\nu) + \bar{\mathcal{L}}(\mu) \tag{41}$$

where $\mathcal{F}(\nu)$ is defined by (3) with a kernel k constructed from the data as an expected product of networks:

$$k(x, x') = \mathbb{E}_{z\sim\mathbb{P}}[\psi(z, x)^T\psi(z, x')] \tag{42}$$

Moreover, $\bar{\mathcal{L}} = \mathcal{F}$ iif $\bar{\mathcal{L}}(\mu) = 0$, which means that the model is well-specified.

Proof of Proposition 10. Let $\phi(z, \nu)$=$\int \psi(z, x)\, \mathrm{d}\nu(x)$ through the computations. By definition (42), we have: $k(x, x') = \int_z \psi(z, x)^T\psi(z, x')\, \mathrm{d}s(z)$ where s denotes the distribution of z. It is easy to see that $\mathcal{F}(\nu) = \frac{1}{2}\int \|\phi(z, \nu) - \phi(z, \mu)\|^2\, \mathrm{d}s(z)$. Indeed expanding the square in the l.h.s and exchanging the order of integrations w.r.t s and $(\mu \otimes \nu)$ one get $\mathcal{F}(\nu)$. Now, introducing $\phi(z, \mu)$ in the expression of $\mathcal{L}(\nu)$, it follows by a simple calculation that:

$$\mathcal{L}(\nu) = \mathcal{L}(\mu) + \mathcal{F}(\nu) + \int \langle\phi(z, \mu) - m(z), \phi(z, \nu) - \phi(z, \mu)\rangle\, \mathrm{d}s(z) \tag{43}$$

where $m(z)$ is the conditional mean of y, i.e.: $m(z) = \int y\, \mathrm{d}p(y|z)$. On the other hand we have that $2\mathcal{L}(\mu) = var(y) + \int \|\phi(z, \mu) - m(z)\|^2\, \mathrm{d}s(z)$, so that $\int \|\phi(z, \mu) - m(z)\|^2\, \mathrm{d}s(z) = 2\bar{\mathcal{L}}(\mu)$. Hence, using Cauchy-Schwartz for the last term in (43), one gets the upper-bound:

$$\mathcal{L}(\nu) \leq \mathcal{L}(\mu) + \mathcal{F}(\nu) + 2\bar{\mathcal{L}}(\mu)^{\frac{1}{2}}\mathcal{F}(\nu)^{\frac{1}{2}}$$

which gives an upper-bound on $\bar{\mathcal{L}}(\nu)$ after subtracting $1/2var(y)$ on both sides of the inequality. To get the lower bound on $\bar{\mathcal{L}}$ one needs to use the global optimality condition of μ for $\mathcal{L}$ from [12, Porposition 3.1]. Indeed, for any $0 < \epsilon \leq 1$ it is easy to see that:

$$\frac{1}{\epsilon}\mathcal{L}(\mu + \epsilon(\nu - \mu)) - \mathcal{L}(\mu) = \int \langle\phi(z, \mu) - m(z), \phi(z, \nu) - \phi(z, \mu)\rangle\, \mathrm{d}s(z)$$

taking the limit $\epsilon \to 0$ and recalling that the l.h.s is always non-negative by optimality of μ is follows that $\int \langle \phi(z,\mu) - m(z), \phi(z,\nu) - \phi(z,\mu) \rangle \, \mathrm{d}s(z)$ must also be non-negative. Therefore, from (43) one get that $\mathcal{L}(\nu) \geq \mathcal{L}(\mu) + \mathcal{F}(\nu)$. The final bound is obtained by again subtracting $1/2var(y)$ form both sides of the inequality. □

The framing (41) implies that optimizing $\mathcal{F}$ can decrease $\bar{\mathcal{L}}$ and vice-versa. However, the two functionals do not generally share the same local minima although they share the same global optima in general. One interesting class of problems where (40) corresponds exactly to minimizing the MMD is the student-teacher problem or the problem of distilling a pre-trained network into another network with the same architecture (see [42]). In this case the gradient flow of the MMD defined in (5) corresponds to the population limit of the usual gradient flow of (39) when the final layer becomes infinitely wide. Indeed, solving (39) is usually done using gradient descent. When the step-size approaches 0, the parameters $(x_1, ..., x_m)$ satisfy the continuous-time system of equations:

$$\dot{x}_i(t) = -\nabla \mathcal{L}(x_1(t), ..., x_m(t)) \text{ for } i = 1, \ldots, m \tag{44}$$

As pointed out in [12, 43], the dynamics in (44) can be analyzed in the "mean-field" limit when $m \to \infty$. For (44), this leads to the continuity equation (5).

B.2 Comparison with the Kullback Leilber divergence flow

Continuity equation and McKean Vlasov process. A famous example of a free energy (33) is the Kullback-Leibler divergence, defined for $\nu, \mu \in \mathcal{P}(\mathcal{X})$ by $KL(\nu,\mu) = \int log(\frac{\nu(x)}{\mu(x)})\nu(x)dx$. Indeed, $KL(\nu,\mu) = \int U(\nu(x))dx + \int V(x)\nu(x)dx$ with $U(s) = s\log(s)$ the entropy function and $V(x) = -log(\mu(x))$. In this case, $\nabla \frac{\partial \mathcal{F}}{\partial \nu} = \nabla \log(\nu) + \nabla V = \nabla \log(\frac{\nu}{\mu})$ and equation (34) leads to the classical Fokker-Planck equation:

$$\frac{\partial \nu}{\partial t} = div(\nu \nabla V) + \Delta \nu \tag{45}$$

where Δ is the Laplacian operator. It is well-known (see for instance [26]) that the distribution of the Langevin diffusion:

$$dX_t = -\nabla \log \mu(X_t)dt + \sqrt{2}dB_t \tag{46}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion, satisfies (45). While the entropy term in the KL functional prevents the particles from "crashing" onto the mode of μ, this role could be played by the interaction energy W for MMD^2 defined in (4). Indeed, consider for instance the gaussian kernel $k(x,x') = e^{-\|x-x'\|^2}$. It is convex thus attractive at long distances ($\|x - x'\| > 1$) but not at small distances (so repulsive).

Convergence to a global minimum. The solution to the Fokker-Planck equation describing the gradient flow of the KL can be shown to converge towards μ under mild assumptions. This follows from the displacement convexity of the KL along the Wasserstein geodesics. Unfortunately the MMD^2 is not displacement convex in general as it is shown in Section 3.1 or Appendix E.2. This makes the task of proving the convergence of the gradient flow of the MMD^2 to the global optimum μ much harder. Moreover, we show in Section 3.2 that local minima which are not global exist and that it is rather easy to reach them. Interestingly, it was shown for some free energies (33) that when the external potential V is not convex, the diffusion may admit several local minima, i.e. there exists several invariant measures to (29), under specific assumptions on V (see [24, 52]). Such assumptions do not apply to the confinement potential V (4) here, but the study of invariant measures of (5) is left for future work.

Sampling algorithms derived from gradient flows. Two settings are usually encountered in the sampling literature: *density-based*, i.e. the target μ is known up to a constant, or *sample-based*, i.e. we only have access to a set of samples $X \sim \mu$. The Unadjusted Langevin Algorithm (ULA), which involves a time-discretized version of the Langevin diffusion, seems much more suitable for first setting, since it only requires the knowledge of $\nabla \log \mu$, whereas our algorithm requires the knowledge of μ (since $\nabla f_{\mu,\nu_n}$ involves an integration over μ). However, in the sample-based setting, it may be difficult to adapt the ULA algorithm, since it would require firstly to estimate $\nabla \log(\mu)$ based on a set of samples of μ, before plugging this estimate in the update of the algorithm. This problem, sometimes referred to as *score estimation* in the literature, has been the subject of a lot of work but remains hard especially in high dimensions (see [51],[31],[45]). In contrast, the discretized flow (in time and space) of the MMD^2 presented Section 4.2 seems naturally adapted to the sample-based setting. Indeed, given samples $(X_n^i)_{1\leq i\leq N}$ of ν_n and samples $(Y^m)_{1\leq m\leq M}$ of μ, $\nabla f_{\hat{\mu},\hat{\nu}_n}(.)$ can be evaluated easily by:

$$\nabla f_{\hat{\mu},\hat{\nu}_n}(z) = \frac{1}{M}\sum_{m=1}^{M}\nabla_2 k(Y^m, z) - \frac{1}{N}\sum_{j=1}^{N}\nabla_2 k(X_n^j, z) \qquad \forall z \in \mathcal{X} \tag{47}$$

where $\nabla_2 k(x,z)$ denotes the gradient of k w.r.t. z.

C Main assumptions

We state here all the assumptions on the kernel k used to prove all the results:

(A) k is continuously differentiable on $\mathcal{X}$ with L-Lipschitz gradient: $\|\nabla k(x,x') - \nabla k(y,y')\| \leq L(\|x-y\| + \|x'-y'\|)$ for all $x,x',y,y' \in \mathcal{X}$.

(B) k is twice differentiable on $\mathcal{X}$.

(C) $\|Dk(x,y)\| \leq \lambda$ for all $x,y \in \mathcal{X}$, where $Dk(x,y)$ is an $\mathbb{R}^{d^2}\times\mathbb{R}^{d^2}$ matrix with entries given by $\partial_{x_i}\partial_{x_j}\partial_{x'_i}\partial_{x'_j}k(x,y)$.

(D) $\sum_{i=1}^{d}\|\partial_i k(x,.) - \partial_i k(y,.)\|^2_{\mathcal{H}} \leq \lambda^2\|x-y\|^2$ for all $x,y\in\mathcal{X}$.

D Construction of the gradient flow of the MMD

D.1 Continuous time flow

Existence and uniqueness of a solution to (5) and (6) is guaranteed under Lipschitz regularity of ∇k.

Proof of Proposition 1. [Existence and uniqueness] Under Assumption (A), the map $(x,\nu)\mapsto \nabla f_{\mu,\nu}(x) = \int\nabla k(x,.)d\nu - \int\nabla k(x,.)d\mu$ is Lipschitz continuous on $\mathcal{X}\times\mathcal{P}_2(\mathcal{X})$ (endowed with the product of the canonical metric on $\mathcal{X}$ and W_2 on $\mathcal{P}_2(\mathcal{X})$), see Proposition 19. Hence, we benefit from standard existence and uniqueness results of McKean-Vlasov processes (see [27]). Then, it is straightforward to verify that the distribution of (6) is solution of (5) by Itô's formula (see [25]). The uniqueness of the gradient flow, given a starting distribution ν_0, results from the λ-convexity of $\mathcal{F}$ which is given by Lemma 15, and then from Theorem 11.1.4 of [1]. The existence derive from the fact that the subdifferential of $\mathcal{F}$ is single-valued, as stated by (2), and that any ν_0 in $\mathcal{P}_2(\mathcal{X})$ is in the domain of $\mathcal{F}$ ([19]). The existence then results from Theorem 11.1.6 and Corollary 11.1.8 from [1]. □

Proof of Proposition 2. [Decay of the MMD] By (2), we have that the differential of $\mathcal{F}(\nu)$ is given by $f_{\mu,\nu}$. The strong subdifferential of F associated with the W_2 metric is thus $\nabla f_{\mu,\nu}$. Finally, since, $\mathcal{F}$ is $3L$-convex by Lemma 15 it follows by the energy identity in [1, Theorem 11.3.2] that for all $0\leq s\leq t$:

$$\int_s^t\int\|\nabla f_{\mu,\nu_u}(x)\|^2\,\mathrm{d}\nu_u(x)\,\mathrm{d}u = \mathcal{F}(\nu_s) - \mathcal{F}(\nu_t).$$

The result follows by dividing by $t-s$ and taking the limit when s got to t. □

D.2 Time-discretized flow

We start by showing that (8) decreases the functional $\mathcal{F}$. In all the proofs, the step-size γ is fixed.

Proof of Proposition 4. Consider a path between ν_n and ν_{n+1} of the form $\rho_t = (I - \gamma t\nabla f_{\mu,\nu_n})_{\#}\nu_n$. We know by Proposition 19 that $\nabla f_{\mu,\nu_n}$ is $2L$ Lipschitz, thus by Lemma 20 and using $\phi(x) = -\gamma\nabla f_{\mu,\nu_n}(x)$ and $\psi(x) = x$, it follows that $\mathcal{F}(\rho_t)$ is differentiable and hence absolutely continuous. Therefore one can write:

$$\mathcal{F}(\rho_1) - \mathcal{F}(\rho_0) = \dot{\mathcal{F}}(\rho_0) + \int_0^1 \dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)dt. \tag{48}$$

Moreover, Lemma 20 with $q = \nu_n$ allows to write:

$$\dot{\mathcal{F}}(\rho_0) = -\gamma \int \|\nabla f_{\mu,\nu_n}(x)\|^2 d\nu_n(x); \qquad |\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)| \leq 3Lt\gamma^2 \int \|\nabla f_{\mu,\nu_n}(X)\|^2 d\nu_n(X).$$

where $t \leq 1$. Hence, the result follows directly by applying the above expression to (48). $\square$

We prove now that (8) approximates (5). To explicit the dependence of the latter sequence on the (fixed) step-size γ, we will write it as : $\nu_{n+1}^\gamma = (I - \gamma \nabla f_{\mu,\nu_n^\gamma})_\# \nu_n^\gamma$ (so $\nu_n^\gamma = \nu_n$ for any $n \geq 0$). We start by introducing a sequence $\bar{\nu}_n^\gamma$ built by iteratively applying:

$$\bar{\nu}_{n+1}^\gamma = (I - \gamma \nabla f_{\mu,\nu_{\gamma n}})_\# \bar{\nu}_n^\gamma \tag{49}$$

with $\bar{\nu}_0 = \nu_0$. Notice that the latter sequence involves the continuous process ν_t of (5) where $t = \gamma n$. Using ν_n^γ, we also consider the interpolation path $\rho_t^\gamma = (I - (t - n\gamma)\nabla f_{\mu,\nu_n^\gamma})_\# \nu_n^\gamma$ for all $t \in [n\gamma, (n+1)\gamma)$ and $n \in \mathbb{N}$, which is the same as in Proposition 3.

Proof of Proposition 3. Let π be an optimal coupling between ν_n^γ and $\nu_{\gamma n}$, and (x, y) a sample from π. For $t \in [n\gamma, (n+1)\gamma)$ we write $y_t = y - \int_{n\gamma}^t \nabla f_{\mu,\nu_s}(u) \, \mathrm{d}u$ and $x_t = x - (t - n\gamma)\nabla f_{\mu,\nu_n^\gamma}(x)$. We also introduce the approximation error $E(t, n\gamma) := y_t - y + (t - n\gamma)\nabla f_{\mu,\nu_{\gamma n}}(y)$ for which we know by Lemma 13 that $\mathcal{E}(t, n\gamma) := \mathbb{E}[E(t, n\gamma)^2]^{\frac{1}{2}}$ is upper-bounded by $(t - n\gamma)^2 C$ for some positive constant C that depends only on T and the Lipschitz constant L. This allows to write:

$$\begin{aligned}
W_2(\rho_t^\gamma, \nu_t) &\leq \mathbb{E}\left[\|y - x + (t - n\gamma)(\nabla f_{\mu,\nu_n^\gamma}(x) - \nabla f_{\mu,\nu_{\gamma n}}(y)) + E(t, n\gamma)\|^2\right]^{\frac{1}{2}} \\
&\leq W_2(\nu_n^\gamma, \nu_{\gamma n}) + 4L(t - n\gamma)W_2(\nu_n^\gamma, \nu_{\gamma n}) + \mathcal{E}(t, n\gamma) \\
&\leq (1 + 4\gamma L)W_2(\nu_n^\gamma, \nu_{\gamma n}) + (t - \gamma n)^2 C \\
&\leq (1 + 4\gamma L)(W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma) + W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma)) + \gamma^2 C \\
&\leq \gamma\left[(1 + 4\gamma L)M(T) + \gamma C\right]
\end{aligned}$$

The second line is obtained using that $\nabla f_{\mu,\nu_{\gamma n}}(x)$ is jointly $2L$-Lipschitz in x and ν (see Proposition 19) and by the fact that $W_2(\nu_n^\gamma, \nu_{\gamma n}) = \mathbb{E}_\pi[\|y - x\|^2]^{\frac{1}{2}}$. The third one is obtained using $t - n\gamma \leq \gamma$. For the last inequality, we used Lemmas 11 and 12 where $M(T)$ a constant that depends only on T. Hence for $\gamma \leq \frac{1}{4L}$ we get $W_2(\rho_t^\gamma, \nu_t) \leq \gamma(\frac{C}{4L} + 2M(T))$. $\square$

Lemma 11. *For any $n \geq 0$:*

$$W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma) \leq \gamma \frac{C}{2L}(e^{n\gamma 2L} - 1)$$

Proof. Let π be an optimal coupling between $\bar{\nu}_n^\gamma$ and $\nu_{\gamma n}$ and $(\bar{x}, x)$ a joint sample from π. Consider also the joint sample $(\bar{y}, y)$ obtained from $(\bar{x}$,$x)$ by applying the gradient flow of $\mathcal{F}$ in continuous time to get $y = x - \int_{n\gamma}^{(n+1)\gamma} \nabla f_{\mu,\nu_s}(u) \, \mathrm{d}u$ and by taking a discrete step from $\bar{x}$ to write $\bar{y} = \bar{x} - \gamma \nabla f_{\mu,\nu_{\gamma n}}(\bar{x})$. It is easy to see that $y \sim \nu_{\gamma(n+1)}$ (i.e. a sample from the continous process (5) at time $t = (n+1)\gamma$) and $\bar{y} \sim \bar{\nu}_{n+1}^\gamma$ (i.e. a sample from (49)). Moreover, we introduce the approximation error $E((n+1)\gamma, n\gamma) := y - x + \gamma \nabla f_{\mu,\nu_{\gamma n}}(x)$ for which we know by Lemma 13 that $\mathcal{E}((n+1)\gamma, n\gamma) := \mathbb{E}[E((n+1)\gamma, n\gamma)^2]^{\frac{1}{2}}$ is upper-bounded by $\gamma^2 C$ for some positive constant C that depends only on T and the Lipschitz constant L. Denoting by $a_n = W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma)$, one can therefore write:

$$\begin{aligned}
a_{n+1} \leq& \mathbb{E}_\pi\left[\|x - \gamma \nabla f_{\mu,\nu_{\gamma n}}(x) - \bar{x} + \gamma \nabla f_{\mu,\nu_{\gamma n}}(\bar{x}) + E((n+1)\gamma, n\gamma)\|^2\right]^{\frac{1}{2}} \\
\leq& \mathbb{E}_\pi\left[\|x - \bar{x}\|^2\right]^{\frac{1}{2}} + \gamma \mathbb{E}_\pi\left[\|\nabla f_{\mu,\nu_{\gamma n}}(x) - \nabla f_{\mu,\nu_{\gamma n}}(\bar{x}))\|^2\right]^{\frac{1}{2}} + \gamma^2 C
\end{aligned}$$

Using that $\nabla f_{\mu,\nu_{\gamma n}}$ is $2L$-Lipschitz by Proposition 19 and recalling that $\mathbb{E}_\pi\left[\|x - \bar{x}\|^2\right]^{\frac{1}{2}} = W_2(\nu_{\gamma n}, \bar{\nu}_n^\gamma)$, we get the recursive inequality $a_{n+1} \leq (1 + 2\gamma L)a_n + \gamma^2 C$. Finally, using Lemma 24 and recalling that $a_0 = 0$, since by definition $\bar{\nu}_0^\gamma = \nu_0^\gamma$, we conclude that $a_n \leq \gamma \frac{C}{2L}(e^{n\gamma 2L} - 1)$. $\square$

Lemma 12. *For any $T > 0$ and n such that $n\gamma \leq T$*

$$W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma) \leq \gamma \frac{C}{8L^2}(e^{4TL} - 1)^2 \tag{50}$$

Proof. Consider now an optimal coupling π between $\bar{\nu}_n^\gamma$ and ν_n^γ. Similarly to Lemma 11, we denote by $(\bar{x}, x)$ a joint sample from π and $(\bar{y}, y)$ is obtained from $(\bar{x}, x)$ by applying the discrete updates : $\bar{y} = \bar{x} - \gamma\nabla f_{\mu,\nu_{\gamma n}}(\bar{x})$ and $y = x - \gamma\nabla f_{\mu,\nu_n^\gamma}(x)$. We again have that $y \sim \nu_{n+1}^\gamma$ (i.e. a sample from the time discretized process (8)) and $\bar{y} \sim \bar{\nu}_{n+1}^\gamma$ (i.e. a sample from (49)). Now, denoting by $b_n = W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma)$, it is easy to see from the definition of $\bar{y}$ and y that we have:

$$\begin{aligned}
b_{n+1} &\leq \mathbb{E}_\pi\left[\|x - \gamma\nabla f_{\mu,\nu_n^\gamma}(x) - \bar{x} + \gamma\nabla f_{\mu,\nu_{\gamma n}}(\bar{x})\|^2\right]^{\frac{1}{2}} \\
&\leq (1 + 2\gamma L)\mathbb{E}_\pi\left[\|x - \bar{x}\|^2\right]^{\frac{1}{2}} + 2\gamma L W_2(\nu_n^\gamma, \nu_{\gamma n})) \\
&\leq (1 + 4\gamma L)b_n + \gamma L W_2(\bar{\nu}_n^\gamma, \nu_{\gamma n})
\end{aligned}$$

The second line is obtained recalling that $\nabla f_{\mu,\nu}(x)$ is $2L$-Lipschitz in both x and ν by Proposition 19. The third line follows by triangular inequality and recalling that $\mathbb{E}_\pi\left[\|x - \bar{x}\|^2\right]^{\frac{1}{2}} = W_2(\nu_n^\gamma, \bar{\nu}_n^\gamma) = b_n$ since π is an optimal coupling between $\bar{\nu}_n^\gamma$ and ν_n^γ. By Lemma 11, we know that $W_2(\bar{\nu}_n^\gamma, \nu_{\gamma n}) \leq \gamma\frac{C}{2L}(e^{2n\gamma L} - 1)$, hence, for any n such that $n\gamma \leq T$ we get the recursive inequality: $b_{n+1} \leq (1 + 4\gamma L)b_n + (C/2L)\gamma^2(e^{2TL} - 1)$. Finally, using again Lemma 24, it follows that $b_n \leq \gamma\frac{C}{8L^2}(e^{4TL} - 1)^2$. ∎

Lemma 13. *[Taylor expansion] Consider the process $\dot{x}_t = -\nabla f_{\mu,\nu_t}(x_t)$, and denote by $\mathcal{E}(t, s) = \mathbb{E}[\|x_t - x_s + (t - s)\nabla f_{\mu,\nu_s}(x_s)\|^2]^{\frac{1}{2}}$ for $0 \leq s \leq t \leq T$. Then one has:*

$$\mathcal{E}(t, s) \leq 4L^2 r_0 e^{LT} \int_s^t \int_s^u \mathrm{d}l\, \mathrm{d}u = 2L^2 r_0 e^{LT}(t - s)^2 \tag{51}$$

Proof. By definition of x_t and $\mathcal{E}(t, s)$ one can write:

$$\begin{aligned}
\mathcal{E}(t, s) &= \mathbb{E}[\|\int_s^t (\nabla f_{\mu,\nu_s}(x_s) - \nabla f_{\mu,\nu_u}(x_u))\,\mathrm{d}u\|^2]^{\frac{1}{2}} \\
&\leq \int_s^t \mathbb{E}[\|(\nabla f_{\mu,\nu_s}(x_s) - \nabla f_{\mu,\nu_u}(x_u))\|^2]^{\frac{1}{2}}\,\mathrm{d}u \\
&\leq 2L \int_s^t \mathbb{E}[(\|x_s - x_u\| + W_2(\nu_s, \nu_u))^2]^{\frac{1}{2}}\,\mathrm{d}u \leq 4L \int_s^t \mathbb{E}[(\|x_s - x_u\|^2]^{\frac{1}{2}}\,\mathrm{d}u
\end{aligned}$$

Where we used an integral expression for x_t in the first line then applied a triangular inequality for the second line. The last line is obtained recalling that $\nabla f_{\mu,\nu}(x)$ is jointly $2L$-Lipschitz in x and ν by Proposition 19 and that $W_2(\nu_s, \nu_u) \leq \mathbb{E}[(\|x_s - x_u\|^2]^{\frac{1}{2}}$. Now we use again an integral expression for x_u which further gives:

$$\begin{aligned}
\mathcal{E}(t, s) \leq& 4L \int_s^t \mathbb{E}[\|\int_s^u \nabla f_{\mu,\nu_l}(x_l)\,\mathrm{d}l\|^2]^{\frac{1}{2}}\,\mathrm{d}u \\
\leq& 4L \int_s^t \int_s^u \mathbb{E}[\|\mathbb{E}[\nabla_1 k(x_l, x_l') - \nabla_1 k(x_l, z)]\|^2]^{\frac{1}{2}}\,\mathrm{d}l\,\mathrm{d}u \\
\leq& 4L^2 \int_s^t \int_s^u \mathbb{E}[\|x_l' - z\|]\,\mathrm{d}l\,\mathrm{d}u
\end{aligned}$$

Again, the second line is obtained using a triangular inequality and recalling the expression of $\nabla f_{\mu,\nu}(x)$ from Proposition 19. The last line uses that ∇k is L-Lipschitz by Assumption (A). Now we need to make sure that $\|x_l' - z\|$ remains bounded at finite times. For this we will first show that $r_t = \mathbb{E}[\|x_t - z\|]$ satisfies an integro-differential inequality:

$$\begin{aligned}
r_t \leq& \mathbb{E}[\|x_0 - z - \int_0^t \nabla f_{\mu,\nu_s}(x_s)\,\mathrm{d}s\|] \\
\leq& r_0 + \int_0^t \mathbb{E}[\|\nabla_1 k(x_s, x_s') - \nabla_1 k(x_s, z)\|\,\mathrm{d}s] \leq r_0 + L\int_0^t r_s\,\mathrm{d}s
\end{aligned}$$

Again, we used an integral expression for x_t in the first line, then a triangular inequality recalling the expression of $\nabla f_{\mu,\nu_s}$. The last line uses again that ∇k is L-Lipschitz. By Gronwall's lemma it is easy to see that $r_t \leq r_0 e^{Lt}$ at all times. Moreover, for all $t \leq T$ we have a fortiori that $r_t \leq r_0 e^{LT}$. Recalling back the upper-bound on $\mathcal{E}(t,s)$ we have finally:

$$\mathcal{E}(t,s) \leq 4L^2 r_0 e^{LT} \int_s^t \int_s^u \mathrm{d}l \, \mathrm{d}u = 2L^2 r_0 e^{LT}(t-s)^2$$

$\square$

E Convergence of the gradient flow of the MMD

E.1 Equilibrium condition

We discuss here the equilibrium condition (11) and relate it to [36, Assumption A]. Recall that (11) is given by: $\int \|\nabla f_{\mu,\nu^*}(x)\|^2 \, \mathrm{d}\nu^*(x) = 0$. Under some mild assumptions on the kernel which are states in [36, Appendix C.1] it is possible to write (11) as:

$$\int \|\nabla f_{\mu,\nu^*}(x)\|^2 \, \mathrm{d}\nu^*(x) = \langle f_{\mu,\nu^*}, D_{\nu^*} f_{\mu,\nu^*} \rangle_{\mathcal{H}} = 0$$

where D_{ν^*} is a Hilbert-Schmidt operator given by:

$$D_{\nu^*} = \int \sum_{i=1}^{d} \partial_i k(x,.) \otimes \partial_i k(x,.) \, \mathrm{d}\nu^*(x)$$

Hence (11) is equivalent to say that f_{μ,ν^*} belongs to the null space of D_{ν^*}. In [36, Theorem 2], a similar equilibrium condition is derived by considering the time derivative of the MMD along the KSD gradient flow:

$$\frac{1}{2}\frac{d}{dt} MMD^2(\mu, \nu_t) = -\lambda \langle f_{\mu,\nu_t}, (\frac{1}{\lambda} I - (D_{\nu_t} + \lambda I)^{-1}) f_{\mu,\nu_t} \rangle_{\mathcal{H}}$$

The r.h.s is shown to be always negative and thus the MMD decreases in time. Hence, as t approaches ∞, the r.h.s tends to 0 since the MMD converges to some limit value l. This provides the equilibrium condition:

$$\lambda \langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*} \rangle_{\mathcal{H}} = 0$$

It is further shown in [36, Lemma 2] that the above equation is also equivalent to having f_{μ,ν^*} in the null space of D_{ν^*} in the case when D_{ν^*} has finite dimensions. We generalize this statement to infinite dimension in Proposition 14. In [36, Assumption A], it is simply assumed that if $f_{\mu,\nu^*} \neq 0$ then $D_{\nu^*} f_{\mu,\nu^*} \neq 0$ which exactly amounts to assuming that local optima which are not global don't exist.

Proposition 14.

$$\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*} \rangle_{\mathcal{H}} = 0 \iff f_{\mu,\nu^*} \in null(D_{\nu^*})$$

Proof. This follows simply by recalling D_{ν^*} is a symmetric non-negative Hilbert-Schmidt operator it has therefore an eigen-decomposition of the form:

$$D_{\nu^*} = \sum_{i=1}^{\infty} \lambda_i e_i \otimes e_i$$

where e_i is an ortho-norrmal basis of $\mathcal{H}$ and λ_i are non-negative. Moreover, f_{μ,ν^*} can be decomposed in $(e_i)_{1 \leq i}$ in the form:

$$f_{\mu,\nu^*} = \sum_{i=0}^{\infty} \alpha_i e_i$$

where α_i is a squared integrable sequence. It follows that $\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*} \rangle_{\mathcal{H}}$ can be written as:

$$\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*} \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} \alpha_i^2$$

Hence, if $f_{\mu,\nu^*} \in null(D_{\nu^*})$ then $\langle f_{\mu,\nu^*}, D_{\nu^*} f_{\mu,\nu^*}\rangle_{\mathcal{H}} = 0$, so that $\sum_{i=1}^{\infty} \lambda_i \alpha_i^2 = 0$. Since λ_i are non-negative, this implies that $\lambda_i \alpha_i^2 = 0$ for all i. Therefore, it must be that $\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*}\rangle_{\mathcal{H}} = 0$. Similarly, if $\langle f_{\mu,\nu^*}, (\frac{1}{\lambda} I - (D_{\nu^*} + \lambda I)^{-1}) f_{\mu,\nu*}\rangle_{\mathcal{H}} = 0$ then $\frac{\lambda_i \alpha_i^2}{\lambda_i + \lambda} = 0$ hence $\langle f_{\mu,\nu*}, D_{\nu^*} f_{\mu,\nu*}\rangle_{\mathcal{H}} = 0$. This means that $f_{\mu,\nu*}$ belongs to $null(D_{\nu^*})$. □

E.2 Λ-displacement convexity of the MMD

We provide now a proof of Proposition 5:

Proof of Proposition 5. [Λ- displacement convexity of the MMD] To prove that $\nu \mapsto \mathcal{F}(\nu)$ is Λ-convex we need to compute the second time derivative $\ddot{\mathcal{F}}(\rho_t)$ where $(\rho_t)_{t\in[0,1]}$ is a displacement geodesic between two probability distributions ν_0 and ν_1 as defined in (26). Such a minimizing geodesic always exists and can be written as $\rho_t = (s_t)_{\#}\pi$ with $s_t = x + t(y - x)$ for all $t \in [0, 1]$ and π is an optimal coupling between ν_0 and ν_1 ([44], Theorem 5.27). Moreover, we denote by v_t the corresponding velocity vector as defined in (29). Recall that $\mathcal{F}(\rho_t) = \frac{1}{2}\|f_{\mu,\rho_t}\|^2_{\mathcal{H}}$, with f_{μ,ρ_t} defined in (1). We start by computing the first derivative of $t \mapsto \mathcal{F}(\rho_t)$. Since Assumptions **(A)** and **(B)** hold, Lemma 21 applies for $\phi(x, y) = y - x$, $\psi(x, y) = x$ and $q = \pi$, thus we know that $\ddot{\mathcal{F}}(\rho_t)$ is well defined and given by:

$$\begin{aligned}\ddot{\mathcal{F}}(\rho_t) =& \mathbb{E}\left[(y - x)^T \nabla_1 \nabla_2 k(s_t(x, y), s_t(x', y'))(y' - x')\right] \\ &+ \mathbb{E}\left[(y - x)^T (H_1 k(s_t(x, y), s_t(x', y')) - H_1 k(s_t(x, y), z))(y - x)\right]\end{aligned} \tag{52}$$

Moreover, Assumption **(C)** also holds which means by Lemma 21 that the second term in (52) can be lower-bounded by $-\sqrt{2}\lambda d\mathcal{F}(\rho_t)\mathbb{E}[\|y - x\|^2]$ so that:

$$\ddot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[(y - x)^T \nabla_1 \nabla_2 k(s_t(x, y), s_t(x', y'))(y' - x')\right] - \sqrt{2}\lambda d\mathcal{F}(\rho_t)\mathbb{E}[\|y - x\|^2]$$

Recall now that $(\rho_t)_{t\in[0,1]}$ is a constant speed geodesic with velocity vector $(v_t)_{t\in[0,1]}$ thus by a change of variable, one further has:

$$\ddot{\mathcal{F}}(\rho_t) \geq \int \left[v_t^T(x) \nabla_1 \nabla_2 k(x, x') v_t(x')\right] \mathrm{d}\rho_t(x) - \sqrt{2}\lambda d\mathcal{F}(\rho_t) \int \|v_t(x)\|^2 \,\mathrm{d}\rho_t(x).$$

Now we can introduce the function $\Lambda(\rho, v) = \langle v, (C_\rho - \sqrt{2}\lambda d\mathcal{F}(\rho)^{\frac{1}{2}} I)v\rangle_{L_2(\rho)}$ which is defined for any pair (ρ, v) with $\rho \in \mathcal{P}_2(\mathcal{X})$ and v a square integrable vector field in $L_2(\rho)$ and where C_ρ is a non-negative operator given by $(C_\rho v)(x) = \int \nabla_x \nabla_{x'} k(x, x')v(x')d\rho(x')$ for any $x \in \mathcal{X}$. This allows to write $\ddot{\mathcal{F}}(\rho_t) \geq \Lambda(\rho_t, v_t)$. It is clear that $\Lambda(\rho, .)$ is a quadratic form on $L_2(\rho)$ and satisfies the requirement in Definition 3. Finally, using Lemma 22 and Definition 4 we conclude that $\mathcal{F}$ is Λ-convex. Moreover, by the reproducing property we also know that for all $\rho \in \mathcal{P}_2(\mathcal{X})$:

$$\mathbb{E}_\rho[v(x)^T \nabla_1 \nabla_2 k(x, x') v(x')] = \mathbb{E}_\rho[\langle v(x)^T \nabla_1 k(x, .), v(x')^T \nabla_1 k(x', .)\rangle_{\mathcal{H}}].$$

By Bochner integrability of $v(x)^T \nabla_1 k(x, .)$ it is possible to exchange the order of the integral and the inner-product [41, Theorem 6]. This leads to the expression $\|\mathbb{E}[v(x)^T \nabla_1 k(x, .)]\|^2_{\mathcal{H}}$. Hence $\Lambda(\rho, v)$ has a second expression of the form:

$$\Lambda(\rho, v) = \|\mathbb{E}_\rho[v(x)^T \nabla_1 k(x, .)]\|^2_{\mathcal{H}} - \sqrt{2}\lambda d\mathcal{F}(\rho)^{\frac{1}{2}} \mathbb{E}_\rho[\|v(x)\|^2].$$

□

We also provide a result showing Λ convexity for $\mathcal{F}$ only under Assumption **(A)**:

Lemma 15 (Λ-discplacement convexity)**.** *Under Assumption* **(A)***, for any $\nu, \nu' \mathcal{P}_2(\mathcal{X})$ and any constant speed geodesic ρ_t from ν to ν', $\mathcal{F}$ satisfies for all $0 \leq t \leq 1$:*

$$\mathcal{F}(\rho_t) \leq (1 - t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) + 3LW_2^2(\nu, \nu')$$

Proof. Let ρ_t be a constant speed geodesic of the form $\rho_t = s_t \# \pi$ where π is an optimal coupling between ν and ν' and $s_t(x, y) = x + t(y - x)$. Since Assumption **(A)** holds, one can apply Lemma 20 with $\psi(x,y) = x$, $\phi(x,y) = y - x$ and $q = \pi$. Hence, one has that $\mathcal{F}(\rho_t)$ is differentiable and its differential satisfies:

$$|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_t)| \leq 3L|t-s| \int \|y-x\|^2 \, \mathrm{d}\pi(x,y)$$

This implies that $\dot{\mathcal{F}}(\rho_t)$ is Lipschitz continuous and therefore is differentiable for almost all $t \in [0,1]$ by Rademacher theorem. Moreover $\ddot{\mathcal{F}}(\rho_t)$ satisfies $\ddot{\mathcal{F}}(\rho_t) \geq -3L \int \|y-x\|^2 \, \mathrm{d}\pi(x,y) = -3LW_2^2(\nu,\nu')$ for almost all $t \in [0,1]$. Using Lemma 22 it follows directly that $\mathcal{F}$ satisfies the desired inequality. □

E.3 Descent up to a barrier

To provide a proof of Theorem 6, we need the following preliminary results. Firstly, an upper-bound on a scalar product involving $\nabla f_{\mu,\nu}$ for any $\mu,\nu \in \mathcal{P}_2(\mathcal{X})$ in terms of the loss functional $\mathcal{F}$, is obtained using the Λ-displacement convexity $\mathcal{F}$ in Lemma 16. Then, an EVI (Evolution Variational Inequality) is obtained in Proposition 17 on the gradient flow of $\mathcal{F}$ in W_2. The proof of the theorem is given afterwards.

Lemma 16. *Let ν be a distribution in $\mathcal{P}_2(\mathcal{X})$ and μ the target distribution such that $\mathcal{F}(\mu) = 0$. Let π be an optimal coupling between ν and μ, and $(\rho_t)_{t\in[0,1]}$ the displacement geodesic defined by (26) with its corresponding velocity vector $(v_t)_{t\in[0,1]}$ as defined in (29). Finally let $\nabla f_{\nu,\mu}(X)$ be the gradient of the witness function between μ and ν. The following inequality holds:*

$$\int \nabla f_{\mu,\nu}(x).(y-x)d\pi(x,y) \leq \mathcal{F}(\mu) - \mathcal{F}(\nu) - \int_0^1 \Lambda(\rho_s, v_s)(1-s)ds$$

where Λ is defined Proposition 5.

Proof. Recall that for all $t \in [0,1]$, ρ_t is given by $\rho_t = (s_t)_{\#}\pi$ with $s_t = x + t(y-x)$. By Λ-convexity of $\mathcal{F}$ the following inequality holds:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\nu) + t\mathcal{F}(\mu) - \int_0^1 \Lambda(\rho_s, v_s)G(s,t)ds$$

Hence by bringing $\mathcal{F}(\nu)$ to the l.h.s and dividing by t and then taking its limit at 0 it follows that:

$$\dot{\mathcal{F}}(\rho_t)|_{t=0} \leq \mathcal{F}(\mu) - \mathcal{F}(\nu) - \int_0^1 \Lambda(\rho_s, v_s)(1-s)ds. \tag{53}$$

where $\dot{\mathcal{F}}(\rho_t) = d\mathcal{F}(\rho_t)/dt$ and since $\lim_{t\to 0} G(s,t) = (1-s)$. Moreover, under Assumption (A), Lemma 20 applies for $\phi(x,y) = y - x$, $\psi(x,y) = x$ and $q = \pi$. It follows therefore that $\dot{\mathcal{F}}(\rho_t)$ is differentiable with time derivative given by: $\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(s_t(x,y)).(y-x) \, \mathrm{d}\pi(x,y)$. Hence at $t = 0$ we get: $\dot{\mathcal{F}}(\rho_t)|_{t=0} = \int \nabla f_{\mu,\nu}(x).(y-x) \, \mathrm{d}\pi(x,y)$ which shows the desired result when used in (53). □

Proposition 17. *Consider the sequence of distributions ν_n obtained from (8). For $n \geq 0$, consider the scalar $K(\rho^n) := \int_0^1 \Lambda(\rho_s^n, V_s^n)(1-s) \, \mathrm{d}s$ where $(\rho_s^n)_{0\leq s\leq 1}$ is a constant speed displacement geodesic from ν_n to the optimal value μ with velocity vectors $(V_s^n)_{0\leq s\leq 1}$. If $\gamma \leq 1/L$, where L is the Lispchitz constant of ∇k in Assumption (A), then:*

$$2\gamma(\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\mu)) \leq W_2^2(\nu_n, \mu) - W_2^2(\nu_{n+1}, \mu) - 2\gamma K(\rho^n). \tag{54}$$

Proof. Let Π^n be the optimal coupling between ν_n and μ, then the optimal transport between ν_n and μ is given by:

$$W_2^2(\mu, \nu_n) = \int \|X - Y\|^2 d\Pi^n(\nu_n, \mu) \tag{55}$$

Moreover, consider $Z = X - \gamma \nabla f_{\mu,\nu_n}(X)$ where (X,Y) are samples from π^n. It is easy to see that (Z,Y) is a coupling between ν_{n+1} and μ, therefore, by definition of the optimal transport map between ν_{n+1} and μ it follows that:

$$W_2^2(\nu_{n+1}, \mu) \leq \int \|X - \gamma\nabla f_{\mu,\nu_n}(X) - Y\|^2 d\pi^n(\nu_n, \mu) \tag{56}$$

By expanding the r.h.s in (56), the following inequality holds:

$$W_2^2(\nu_{n+1},\mu) \le W_2^2(\nu_n,\mu) - 2\gamma \int \langle \nabla f_{\mu,\nu_n}(X), X - Y\rangle d\pi^n(\nu_n,\mu) + \gamma^2 D(\nu_n) \tag{57}$$

where $D(\nu_n) = \int \|\nabla f_{\mu,\nu_n}(X)\|^2 d\nu_n$. By Lemma 16 it holds that:

$$-2\gamma \int \nabla f_{\mu,\nu_n}(X).(X-Y)d\pi(\nu,\mu) \le -2\gamma\left(\mathcal{F}(\nu_n) - \mathcal{F}(\mu) + K(\rho^n)\right) \tag{58}$$

where $(\rho_t^n)_{0\le t\le 1}$ is a constant-speed geodesic from ν_n to μ and $K(\rho^n) := \int_0^1 \Lambda(\rho_s^n, v_s^n)(1-s)ds$. Note that when $K(\rho^n) \le 0$ it falls back to the convex setting. Therefore, the following inequality holds:

$$W_2^2(\nu_{n+1},\mu) \le W_2^2(\nu_n,\mu) - 2\gamma\left(\mathcal{F}(\nu_n) - \mathcal{F}(\mu) + K(\rho^n)\right) + \gamma^2 D(\nu_n) \tag{59}$$

Now we introduce a term involving $\mathcal{F}(\nu_{n+1})$. The above inequality becomes:

$$W_2^2(\nu_{n+1},\mu) \le W_2^2(\nu_n,\mu) - 2\gamma\left(\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\mu) + K(\rho^n)\right) \tag{60}$$

$$+ \gamma^2 D(\nu_n) - 2\gamma(\mathcal{F}(\nu_n) - \mathcal{F}(\nu_{n+1})) \tag{61}$$

It is possible to upper-bound the last two terms on the r.h.s. by a negative quantity when the step-size is small enough. This is mainly a consequence of the smoothness of the functional $\mathcal{F}$ and the fact that ν_{n+1} is obtained by following the steepest direction of $\mathcal{F}$ starting from ν_n. Proposition 4 makes this statement more precise and enables to get the following inequality:

$$\gamma^2 D(\nu_n) - 2\gamma(\mathcal{F}(\nu_n) - \mathcal{F}(\nu_{n+1}) \le -\frac{3}{2}\gamma^2(1-\gamma L)D(\nu_n), \tag{62}$$

where L is the Lispchitz constant of ∇k. Combining (61) and (62) we finally get:

$$2\gamma(\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\mu)) + \frac{3}{2}\gamma^2(1-\gamma L)D(\nu_n) \le W_2^2(\nu_n,\mu) - W_2^2(\nu_{n+1},\mu) - 2\gamma K(\rho^n). \tag{63}$$

and under the condition $\gamma \le 1/L$ we recover the desired result. □

We can now give the proof of the Theorem 6.

Proof of Theorem 6. Consider the Lyapunov function $L_j = j\gamma(\mathcal{F}(\nu_j) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_j,\mu)$ for any iteration j. At iteration $j+1$, we have:

$$\begin{aligned}
L_{j+1} &= j\gamma(\mathcal{F}(\nu_{j+1}) - \mathcal{F}(\mu)) + \gamma(\mathcal{F}(\nu_{j+1}) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_{j+1},\mu) \\
&\le j\gamma(\mathcal{F}(\nu_{j+1}) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_j,\mu) - \gamma K(\rho^j) \\
&\le j\gamma(\mathcal{F}(\nu_j) - \mathcal{F}(\mu)) + \frac{1}{2}W_2^2(\nu_j,\mu) - \gamma K(\rho^j) - j\gamma^2(1 - \frac{3}{2}\gamma L)\int \|\nabla f_{\mu,\nu_j}(X)\|^2 d\nu_j \\
&\le L_j - \gamma K(\rho^j).
\end{aligned}$$

where we used Proposition 17 and Proposition 4 successively for the two first inequalities. We thus get by telescopic summation:

$$L_n \le L_0 - \gamma \sum_{j=0}^{n-1} K(\rho^j) \tag{64}$$

Let us denote $\bar{K}$ the average value of $(K(\rho^j))_{0\le j\le n}$ over iterations up to n. We can now write the final result:

$$\mathcal{F}(\nu_n) - \mathcal{F}(\mu) \le \frac{W_2^2(\nu_0,\mu)}{2\gamma n} - \bar{K} \tag{65}$$

□

E.4 Lojasiewicz type inequalities

Proof of Proposition 7. This proof follows simply from the definition of the negative Sobolev distance. Under Assumption (A), the kernel has at most quadratic growth hence, for any $\mu, \nu \in \mathcal{P}_2(\mathcal{X})^2$, $f_{\mu,\nu} \in L_2(\nu)$. Consider $g = \|f_{\mu,\nu_t}\|^{-1}_{\dot{H}^1(\nu_t)} f_{\mu,\nu_t}$, then $g \in L_2(\nu_t)$ and $\|g\|_{\dot{H}(\nu_t)} \leq 1$. Therefore, we directly have:

$$|\int g \, \mathrm{d}\nu_t - \int g \, \mathrm{d}\mu| \leq \|\nu_t - \mu\|_{\dot{H}^{-1}(\nu_t)} \tag{66}$$

Now, recall the definition of g, which implies that

$$|\int g \, \mathrm{d}\nu_t - \int g \, \mathrm{d}\mu| = \|\nabla f_{\mu,\nu_t}\|^{-1}_{L_2(\nu_t)} |\int f_{\mu,\nu_t} \, \mathrm{d}\nu_t - \int f_{\mu,\nu_t} \, \mathrm{d}\mu|. \tag{67}$$

Moreover, we have that $\int f_{\mu,\nu_t} \, \mathrm{d}\nu_t - \int f_{\mu,\nu_t} \, \mathrm{d}\mu = \|f_{\mu,\nu_t}\|^2_{\mathcal{H}}$, since f_{μ,ν_t} is the witness functions between ν_t and μ. Combining (66) and (67) we thus get the desired Lojasiewicz inequality on f_{μ,ν_t}:

$$\|f_{\mu,\nu_t}\|^2_{\mathcal{H}} \leq \|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)} \|\mu - \nu_t\|_{\dot{H}^{-1}(\nu_t)} \tag{68}$$

where $\|f_{\mu,\nu_t}\|_{\dot{H}(\nu_t)} = \|\nabla f_{\mu,\nu_t}\|^2_{L_2(\nu_t)}$ by definition. Then, using Proposition 2 and recalling by assumption that: $\|\mu - \nu_t\|^2_{\dot{H}^{-1}(\nu_t)} \leq C$, we have:

$$\dot{\mathcal{F}}(\nu_t) = -\|\nabla f_{\mu,\nu_t}\|^2_{L_2(\nu_t)} \leq -\frac{1}{C}\|f_{\mu,\nu_t}\|^2_{\mathcal{H}} = -\frac{4}{C}\mathcal{F}(\nu_t)^2 \tag{69}$$

It is clear that if $\mathcal{F}(\nu_0) > 0$ then $\mathcal{F}(\nu_t) > 0$ at all times by uniqueness of the solution. Hence, one can divide by $\mathcal{F}(\nu_t)^2$ and integrate the inequality from 0 to some time t. The desired inequality is obtained by simple calculations.

Then, using Proposition Proposition 4 and (69) where ν_t is replaced by ν_n it follows:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma\left(1 - \frac{3}{2}L\gamma\right)\|\phi_n\|^2_{L_2(\nu_n)} \leq -\frac{4}{C}\gamma\left(1 - \frac{3}{2}\gamma L\right)\mathcal{F}(\nu_n)^2.$$

Dividing by both sides of the inequality by $\mathcal{F}(\nu_n)\mathcal{F}(\nu_{n+1})$ and recalling that $\mathcal{F}(\nu_{n+1}) \leq \mathcal{F}(\nu_n)$ it follows directly that:

$$\frac{1}{\mathcal{F}(\nu_n)} - \frac{1}{\mathcal{F}(\nu_{n+1})} \leq -\frac{4}{C}\gamma\left(1 - \frac{3}{2}\gamma L\right).$$

The proof is concluded by summing over n and rearranging the terms. □

E.5 Lojasiewicz-type inequalities for $\mathcal{F}$ under different metrics

The Wasserstein gradient flow of $\mathcal{F}$ can be seen as the continuous-time limit of the so called minimizing movement scheme [1]. Such proximal scheme is defined using an initial distribution ν_0, a step-size τ, and an iterative update equation:

$$\nu_{n+1} \in \arg\min_{\nu} \mathcal{F}(\nu) + \frac{1}{2\tau}W_2^2(\nu, \nu_n). \tag{70}$$

In [1], it is shown that the continuity equation (5) can be obtained as the limit when $\tau \to 0$ of (70) using suitable interpolations between the elements ν_n. In [42], a different proximal scheme is used where $W_2^2(\nu, \nu_n)$ is replaced by $\beta W_2^2(\nu, \nu_n) + \alpha KL(\nu\|\nu_n)$ with $\beta = 1$. Here, we keep $\beta > 0$ for more generality. It is shown at least formally that such scheme corresponds to a transport equation with a birth-death dynamics:

$$\partial_t \nu_t = \beta div(\nu_t \nabla f_{\mu,\nu_t}) + \alpha(f_{\mu,\nu_t} - \int f_{\mu,\nu_t}(x) \, \mathrm{d}\nu_t(x))\nu_t$$

Under such dynamics, [42, Proposition 3.1] states that the time evolution of $\mathcal{F}$ can be written as:

$$\dot{\mathcal{F}}(\nu_t) = -\beta \int \|\nabla f_{\mu,\nu_t}\|^2 \, \mathrm{d}\nu_t(x) - \alpha \int \left| f_{\mu,\nu_t}(x) - \int f_{\mu,\nu_t}(x') \, \mathrm{d}\nu_t(x') \right|^2 \mathrm{d}\nu_t(x) \tag{71}$$

We would like to apply the same approach as in Section 3.2 to provide a condition on the convergence of (71). Hence we first introduce an analogue to the Negative Sobolev distance in Definition 1 by duality:

$$D_\nu(p,q) = \sup_{\substack{g\in L_2(\nu)\\ \beta\|\nabla g\|^2_{L_2(\nu)}+\alpha\|g-\bar{g}\|^2_{L_2(\nu)}\leq 1}} |\int g(x)\,\mathrm{d}p(x) - \int g(x)\,\mathrm{d}q(x)| \tag{72}$$

where $\bar{g}$ is simply the expectation of g under ν. Such quantity defines a distance, since it is the dual of a semi-norm. Now using the particular structure of the MMD, we recall that $f_{\mu,\nu}\in L_2(\nu)$ and that $\beta\|\nabla f\|^2_{L_2(\nu)} + \alpha\|f-\bar{f}\|^2_{L_2(\nu)} < \infty$. Hence for a particular g of the form:

$$g = \frac{f_{\mu,\nu}}{(\beta\|\nabla f_{\mu,\nu}\|^2_{L_2(\nu)} + \alpha\|f_{\mu,\nu}-\bar{f}_{\mu,\nu}\|^2_{L_2(\nu)})^{\frac{1}{2}}}$$

the following inequality holds:

$$D_\nu(\mu,\nu) \geq \frac{|\int f_{\mu,\nu}\,\mathrm{d}\nu(x) - \int f_{\mu,\nu}\,\mathrm{d}\mu(x)|}{(\beta\|\nabla f_{\mu,\nu}\|^2_{L_2(\nu)} + \alpha\|f_{\mu,\nu}-\bar{f}_{\mu,\nu}\|^2_{L_2(\nu)})^{\frac{1}{2}}}.$$

But since $f_{\mu,\nu}$ is the witness function between μ and ν we have that $2\mathcal{F}(\nu) = |\int f_{\mu,\nu}\,\mathrm{d}\nu(x) - \int f_{\mu,\nu}\,\mathrm{d}\mu(x)|$. Hence one can write that:

$$D^2_\nu(\mu,\nu)(\beta\|\nabla f_{\mu,\nu}\|^2_{L_2(\nu)}) \geq 4\mathcal{F}^2(\nu) \tag{73}$$

Now provided that $D^2_\nu(\mu,\nu_t)$ remains bounded at all time t by some constant $C>0$ one can easily deduce a rate of convergence for $\mathcal{F}(\nu_t)$ just as in Proposition 7. In fact, in the case when $\beta = 1$ and $\alpha = 0$ one recovers Proposition 7. Another interesting case is when $\beta = 0$ and $\alpha = 1$. In this case, $D_\nu(p,q)$ is defined for p and q such that the difference $p-q$ is absolutely continuous w.r.t. ν. Moreover, $D_\nu(p,q)$ has the simple expression:

$$D_\nu(p,q) = \int (\frac{p-q}{\nu}(x))^2\,\mathrm{d}\nu(x)$$

where $\frac{p-q}{\nu}$ denotes the radon nikodym density of $p-q$ w.r.t. ν. More importantly, $D^2_\nu(\mu,\nu)$ is exactly equal to $\chi^2(\mu\|\nu)^{\frac{1}{2}}$. As we will show now, $(\chi^2)^{\frac{1}{2}}$ turns out to be a linearization of $\sqrt{2}KL^{\frac{1}{2}}$. For $0<\epsilon<1$ and μ absolutely continuous w.r.t to ν set $G(\epsilon) = KL(\nu\|(\nu+\epsilon(\mu-\nu))$. Exchanging the derivatives and the integral, $\dot{G}(\epsilon)$ and $\ddot{G}(\epsilon)$ are both given by:

$$\dot{G}(\epsilon) = -\int \frac{\mu-\nu}{\nu+\epsilon(\mu-\nu)}\,\mathrm{d}\nu$$
$$\ddot{G}(\epsilon) = \int \frac{(\nu-\mu)^2}{(\nu+\epsilon(\mu-\nu))^2}\,\mathrm{d}\nu$$

Hence, we have for $\epsilon = 0$: $\dot{G}(0) = 0$ and $\ddot{G}(0) = \chi^2(\mu\|\nu)$. Therefore, using a Taylor expansion of the second order in ϵ it follows: $G(\epsilon) = \frac{1}{2}\chi^2(\mu\|\nu)\epsilon^2 + o(\epsilon^2)$, which means that

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}(2KL(\nu'\|(\nu'+\epsilon(\nu-\nu')))^{\frac{1}{2}} = \chi^2(\nu\|\nu')^{\frac{1}{2}}. \tag{74}$$

Whereas the Negative Sobolev Distance is a linearization of W_2 and its boundedness alongs the Wasserstein flow of $\mathcal{F}$ guarantees convergence towards the global optimum. The square-root of the χ^2 divergence is a linearization of $KL^{\frac{1}{2}}$ and its boundedness along the birth-death dynamics of $\mathcal{F}$ also guarantees convergence towards the global optimum.

F Algorithms

F.1 Noisy Gradient flow of the MMD

Proof of Proposition 8. To simplify notations, we write $V = \nabla f_{\mu,\nu_n}$ and $\mathcal{D}_{\beta_n}(\nu_n) = \int \|V(x+\beta_n u)\|^2 g(u)\,\mathrm{d}\nu_n\,\mathrm{d}u$ where g is the density of a standard gaussian. The symbol $\otimes$ denotes the standard convolution. Recall that a sample x_{n+1} from ν_{n+1} is obtained using $x_{n+1} = x_n - \gamma V(x_n + \beta_n u_n)$ where x_n is a sample from ν_n and u_n is a sample from a standard gaussian distribution that is independent from x_n. Moreover, by assumption β_n is a non-negative scalar satisfying:

$$8\lambda^2\beta_n^2\mathcal{F}(\nu_n) \leq \mathcal{D}_{\beta_n}(\nu_n) \tag{75}$$

Consider now the map $(x, u) \mapsto s_t(x) = x - \gamma t V(x + \beta_n u)$ for $0 \leq t \leq 1$, then ν_{n+1} is obtained as a push-forward of $\nu_n \otimes g$ by s_1: $\nu_{n+1} = (s_1)_{\#}(\nu_n \otimes g)$. Moreover, the curve $\rho_t = (s_t)_{\#}(\nu_n \otimes g)$ is a path from ν_n to ν_{n+1}. We know by Proposition 19 that $\nabla f_{\mu,\nu_n}$ is $2L$-Lipschitz, thus using $\phi(x, u) = -\gamma V(x + \beta_n u)$, $\psi(x, u) = x$ and $q = \nu_n \otimes g$ in Lemma 20 it follows that $\mathcal{F}(\rho_t)$ is differentiable in t with:

$$\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(s_t(x)).(-\gamma V(x + \beta_n u))g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u$$

Moreover, $\dot{\mathcal{F}}(\rho_0)$ is given by $\dot{\mathcal{F}}(\rho_0) = -\gamma \int V(x).V(x + \beta_n u)g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u$ and the following estimate holds:

$$|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)| \leq 3\gamma^2 Lt \int \|V(x + \beta_n u)\|^2 g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u = 3\gamma^2 Lt\mathcal{D}_{\beta_n}(\nu_n). \tag{76}$$

Using the absolute continuity of $\mathcal{F}(\rho_t)$, one has $\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) = \dot{\mathcal{F}}(\rho_0) + \int_0^1 \dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_0)\,\mathrm{d}t$. Combining with (76) and using the expression of $\dot{\mathcal{F}}(\rho_0)$, it follows that:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma \int V(x).V(x + \beta_n u)g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u + \frac{3}{2}\gamma^2 L\mathcal{D}_{\beta_n}(\nu_n). \tag{77}$$

Adding and subtracting $\gamma\mathcal{D}_{\beta_n}(\nu_n)$ in (77) it follows directly that:

$$\begin{aligned}\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq& -\gamma(1 - \frac{3}{2}\gamma L)\mathcal{D}_{\beta_n}(\nu_n)\\ &+ \gamma \int (V(x + \beta_n u) - V(x)).V(x + \beta_n u)g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u\end{aligned} \tag{78}$$

We shall control now the last term in (78). Recall now that for all $1 \leq i \leq d$, $V_i(x) = \partial_i f_{\mu,\nu_n}(x) = \langle f_{\mu,\nu_n}, \partial_i k(x, .)\rangle$ where we used the reproducing property for the derivatives of f_{μ,ν_n} in $\mathcal{H}$ (see Appendix A.1). Therefore, it follows by Cauchy-Schwartz in $\mathcal{H}$ and using Assumption (D):

$$\begin{aligned}\|V(x + \beta_n u) - V(x)\|^2 &\leq \|f_{\mu,\nu_n}\|^2_{\mathcal{H}} \left(\sum_{i=1}^{d} \|\partial_i k(x + \beta_n u, .) - \partial_i k(x, .)\|^2_{\mathcal{H}}\right)\\ &\leq \lambda^2\beta_n^2\|f_{\mu,\nu_n}\|^2_{\mathcal{H}}\|u\|^2\end{aligned}$$

for all $x, u \in \mathcal{X}$. Now integrating both sides w.r.t. ν_n and g and recalling that g is a standard gaussian, we have:

$$\int \|V(x + \beta_n u) - V(x)\|^2 g(u)\,\mathrm{d}\nu_n(x)\,\mathrm{d}u \leq \lambda^2\beta_n^2\|f_{\mu,\nu_n}\|^2_{\mathcal{H}} \tag{79}$$

Getting back to (78) and applying Cauchy-Schwarz in $L_2(\nu_n \otimes g)$ it follows:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\gamma(1 - \frac{3}{2}\gamma L)\mathcal{D}_{\beta_n}(\nu_n) + \gamma\lambda\beta_n\|f_{\mu,\nu_n}\|_{\mathcal{H}}\mathcal{D}^{\frac{1}{2}}_{\beta_n}(\nu_n) \tag{80}$$

It remains to notice that $\|f_{\mu,\nu_n}\|^2_{\mathcal{H}} = 2\mathcal{F}(\nu_n)$ and that β_n satisfies (75) to get:

$$\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\frac{\gamma}{2}(1 - \frac{3}{2}\gamma L)\mathcal{D}_{\beta_n}(\nu_n).$$

We introduce now $\Gamma = 4\gamma(1 - \frac{3}{2}\gamma L)\lambda^2$ to simplify notation and prove the second inequality. Using (75) again in the above inequality we directly have: $\mathcal{F}(\nu_{n+1}) - \mathcal{F}(\nu_n) \leq -\Gamma\beta_n^2\mathcal{F}(\nu_n)$. One can already deduce that $\Gamma\beta_n^2$ is necessarily smaller than 1. Hence, taking $\mathcal{F}(\nu_n)$ to the r.h. side and iterating over n it follows that:

$$\mathcal{F}(\nu_n) \leq \mathcal{F}(\nu_0)\prod_{i=0}^{n-1}(1 - \Gamma\beta_n^2)$$

Simply using that $1 - \Gamma\beta_n^2 \leq e^{-\Gamma\beta_n^2}$ leads to the desired upper-bound $\mathcal{F}(\nu_n) \leq \mathcal{F}(\nu_0)e^{-\Gamma\sum_{i=0}^{n-1}\beta_n^2}$.

□

F.2 Sample-based approximate scheme

Proof of Theorem 9. Let $(u_n^i)_{1\leq i\leq N}$ be i.i.d standard gaussian variables and $(x_0^i)_{1\leq i\leq N}$ i.i.d. samples from ν_0. We consider $(x_n^i)_{1\leq i\leq N}$ the particles obtained using the approximate scheme (21): $x_{n+1}^i = x_n^i - \gamma\nabla f_{\hat{\mu},\hat{\nu}_n}(x_n^i + \beta_n u_n^i)$ starting from $(x_0^i)_{1\leq i\leq N}$, where $\hat{\nu_n}$ is the empirical distribution of these N interacting particles. Similarly, we denote by $(\bar{x}_n^i)_{1\leq i\leq N}$ the particles obtained using the exact update equation (17): $\bar{x}_{n+1}^i = \bar{x}_n^i - \gamma\nabla f_{\mu,\nu_n}(\bar{x}_n^i + \beta_n u_n^i)$ also starting from $(x_0^i)_{1\leq i\leq N}$. By definition of ν_n we have that $(\bar{x}_n^i)_{1\leq i\leq N}$ are i.i.d. samples drawn from ν_n with empirical distribution denoted by $\bar{\nu}_n$. We will control the expected error c_n defined as $c_n^2 = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\|x_n^i - \bar{x}_n^i\|^2\right]$. By recursion, we have:

$$
\begin{aligned}
c_{n+1} =& \frac{1}{\sqrt{N}}\left(\sum_{i=1}^N \mathbb{E}\left[\|x_n^i - \bar{x}_n^i - \gamma(\nabla f_{\hat{\mu},\hat{\nu}_n}(x_n^i+\beta_n u_n^i) - \nabla f_{\mu,\nu_n}(\bar{x}_n^i + \beta_n u_n^i))\|^2\right]\right)^{\frac{1}{2}}\\
\leq& c_n + \frac{\gamma}{\sqrt{N}}(\sum_{i=1}^N \mathcal{E}_i)^{\frac{1}{2}} + \frac{\gamma}{\sqrt{N}}(\sum_{i=1}^N \mathcal{G}_i)^{\frac{1}{2}})\\
&+\frac{\gamma}{\sqrt{N}}(\sum_{i=1}^N \mathbb{E}[\|\nabla f_{\mu,\hat{\nu}_n}(x_n^i+\beta_n u_n^i) - \nabla f_{\mu,\bar{\nu}_n}(\bar{x}_n^i+\beta_n u_n^i)\|^2])^{\frac{1}{2}}\\
\leq& c_n + 2\gamma L(c_n + \mathbb{E}[W_2(\hat{\nu}_n,\bar{\nu}_n)^2]^{\frac{1}{2}}) + \frac{\gamma}{\sqrt{N}}(\sum_{i=1}^N \mathcal{E}_i)^{\frac{1}{2}} + \frac{\gamma}{\sqrt{N}}(\sum_{i=1}^N \mathcal{G}_i)^{\frac{1}{2}})
\end{aligned}
$$

where the second line follows from a simple triangular inequality and the last line is obtained recalling that $\nabla f_{\mu,\nu}(x)$ is jointly $2L$ Lipschitz in x and ν by Proposition 19. Here, $\mathcal{E}_i$ represents the error between $\bar{\nu}_n$ and ν_n while $\mathcal{G}_i$ represents the error between $\hat{\mu}$ and μ and are given by:

$$
\begin{aligned}
\mathcal{E}_i &= \mathbb{E}[\|\nabla f_{\mu,\bar{\nu}_n}(\bar{x}_n^i+\beta_n u_n^i) - \nabla f_{\mu,\nu_n}(\bar{x}_n^i+\beta_n u_n^i)\|^2]\\
\mathcal{G}_i &= \mathbb{E}[\|\nabla f_{\hat{\mu},\hat{\nu}_n}(x_n^i+\beta_n u_n^i) - \nabla f_{\mu,\hat{\nu}_n}(x_n^i+\beta_n u_n^i)\|^2]
\end{aligned}
$$

We will first control the error term $\mathcal{E}_i$. To simplify notations, we write $y^i = \bar{x}_n^i + \beta_n u_n^i$. Recalling the expression of $\nabla f_{\mu,\nu}$ from Proposition 19 and expanding the squared norm in $\mathcal{E}_i$, it follows:

$$
\begin{aligned}
\mathcal{E}_i &= \mathbb{E}[\|\frac{1}{N}\sum_{j=1}^N \nabla k(y^i,\bar{x}_n^j) - \int \nabla k(y^i,x)d\nu_n(x)\|^2]\\
&= \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}\left[\|\nabla k(y^i,\bar{x}_n^j) - \int \nabla k(y^i,x)d\nu_n(x)\|^2\right]\\
&\leq \frac{L^2}{N^2}\sum_{j=1}^N \mathbb{E}[\|\bar{x}_n^j - \int x d\nu_n(x)\|^2] = \frac{L^2}{N}var(\nu_n).
\end{aligned}
$$

The second line is obtained using the independence of the auxiliary samples $(\bar{x}_n^i)_{1\leq i\leq N}$ and recalling that they are distributed according to ν_n. The last line uses the fact that $\nabla k(y,x)$ is L-Lipshitz in x by Assumption (A). To control the variance $var(\nu_n)$ we use Lemma 18 which implies that $var(\nu_n)^{\frac{1}{2}} \leq (B + var(\nu_0)^{\frac{1}{2}})e^{LT}$ for all $n \leq \frac{2T}{\gamma}$. For $\mathcal{G}_i$, it is sufficient to expand again the squared norm and recall that $\nabla k(y,x)$ is L-Lipschitz in x which then implies that $\mathcal{G}_i \leq \frac{L^2}{M}var(\mu)$. Finally, one can observe that $\mathbb{E}[W_2^2(\hat{\nu}_n,\bar{\nu}_n)] \leq \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\|x_n^i - \bar{x}_n^i\|^2\right] = c_n^2$, hence c_n satisfies the recursion:

$$
c_{n+1} \leq (1+4\gamma L)c_n + \frac{\gamma L}{\sqrt{N}}(B + var(\nu_0)^{\frac{1}{2}})e^{2LT} + \frac{\gamma L}{\sqrt{M}}var(\mu).
$$

Using Lemma 24 to solve the above inequality, it follows that:

$$
c_n \leq \frac{1}{4}\left(\frac{1}{\sqrt{N}}(B + var(\nu_0)^{\frac{1}{2}})e^{2LT} + \frac{1}{\sqrt{M}}var(\mu))\right)(e^{4LT}-1)
$$

□

Lemma 18. *Consider an initial distribution ν_0 with finite variance, a sequence $(\beta_n)_{n\geq 0}$ of non-negative numbers bounded by $B < \infty$ and define the sequence of probability distributions ν_n of the process (17):*

$$x_{n+1} = x_n - \gamma \nabla f_{\mu,\nu_n}(x_n + \beta_n u_n) \qquad x_0 \sim \nu_0$$

where $(u_n)_{n\geq 0}$ are standard gaussian variables. Under Assumption (A), the variance of ν_n satisfies for all $T > 0$ and $n \leq \frac{T}{\gamma}$ the following inequality:

$$var(\nu_n)^{\frac{1}{2}} \leq (B + var(\nu_0)^{\frac{1}{2}})e^{2TL}$$

Proof. Let g be the density of a standard gaussian. The idea is to find a recursion from $var(\nu_n)$ to $var(\nu_{n+1})$:

$$\begin{aligned}
var(\nu_{n+1})^{\frac{1}{2}} &= (\mathbb{E}[\|x_n - \gamma \nabla f_{\mu,\nu_n}(x_n + \beta_n u_n) - \int x d\nu_n(x) + \gamma \mathbb{E}[\nabla f_{\mu,\nu_n}(x + \beta_n u)]])^{\frac{1}{2}} \\
&\leq var(\nu_n)^{\frac{1}{2}} + \gamma(\mathbb{E}[\|\nabla f_{\mu,\nu_n}(x_n + \beta_n u_n) - \mathbb{E}[\nabla f_{\mu,\nu_n}(x + \beta_n u)]\|^2])^{\frac{1}{2}} \\
&\leq var(\nu_n)^{\frac{1}{2}} + 2\gamma L \mathbb{E}_{\substack{x,x' \sim \nu_n \\ u,u' \sim g}}[\|x + \beta_n u - x' + \beta_n u'\|^2]^{\frac{1}{2}} \\
&\leq var(\nu_n)^{\frac{1}{2}} + 2\gamma L(var(\nu_n)^{\frac{1}{2}} + \beta_n)
\end{aligned}$$

The second and last lines are obtained using a triangular inequality while the third line uses that $\nabla f_{\mu,\nu_n}(x)$ is $2L$-Lipschitz in x by Proposition 19. Recalling that β_n is bounded by B it is easy to conclude using Lemma 24. □

F.3 Pseudocode for the algorithm of Section 4.2

Algorithm 1 Noisy particle approximation of the MMD flow

Input ν_0, N, n_{iter}, h, $(Y^m)_{1\leq m\leq M}$, β_0, γ
Output $(X^i_{n_{iter}})_{1\leq i\leq N}$
Initialize the particles
$X^i_0 \overset{\text{i.i.d}}{\sim} \nu_0$
Initialize the level of noise
$\beta = \beta_0$
for $n = 0, \dots, n_{iter}$ **do**
  Update the level of noise
  $\beta_n = h(\beta, n)$
  Compute the current empirical distribution of the particles
  $\widehat{\nu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^i_n}$
  Compute the current vector field
  $\nabla f_{\hat{\mu},\hat{\nu}_n}(.) = \frac{1}{M}\sum\limits_{m=1}^{M} \nabla_2 k(Y^m, .) - \frac{1}{N}\sum\limits_{i=1}^{N} \nabla_2 k(X^i_n, .)$
  for $i = 1, \dots, N$ **do**
    $X^i_{n+1} = X^i_n - \gamma \nabla f_{\hat{\mu},\hat{\nu}_n}(X^i_n + \beta_n U^i_n)$

In the experiments, we choose the update function as $h(\beta, n) = \beta \mathbb{I}\{n \leq 0.5 * n_{iter}\}$.

G Auxiliary results

Proposition 19. *Under Assumption (A), the witness function $f_{\mu,\nu}$ between any probability distributions μ and ν in $\mathcal{P}_2(\mathcal{X})$ is differentiable and satisfies:*

$$\nabla f_{\mu,\nu}(z) = \int \nabla_1 k(z,x)\, \mathrm{d}\mu(x) - \int \nabla_1 k(z,x)\, \mathrm{d}\nu(x) \qquad \forall z \in \mathcal{X} \tag{81}$$

where $z \mapsto \nabla_1 k(x, z)$ denotes the gradient of $z \mapsto k(x, z)$ for a fixed $x \in \mathcal{X}$. Moreover, the map $(z, \mu, \nu) \mapsto f_{\mu,\nu}(z)$ is Lipschitz with:

$$\|\nabla f_{\mu,\nu}(z) - \nabla f_{\mu',\nu'}(z')\| \leq 2L(\|z - z'\| + W_2(\mu, \mu') + W_2(\nu, \nu')) \tag{82}$$

Finally, each component of $\nabla f_{\mu,\nu}$ belongs to $\mathcal{H}$.

Proof. The expression of the witness function is given in (1). To establish (81), we simply need to apply the differentiation lemma [29, Theorem 6.28]. By Assumption **(A)**, it follows that $(x, z) \mapsto \nabla_1 k(z, x)$ has at most a linear growth. Hence on any bounded neighborhood of z, $x \mapsto \|\nabla_1 k(z, x)\|$ is upper-bounded by an integrable function w.r.t. μ and ν. Therefore, the differentiation lemma applies and $\nabla f_{\mu,\nu}(z)$ is differentiable with gradient given by (81).

To prove the second statement, we will consider two optimal couplings: π_1 with marginals μ and μ' and π_2 with marginals ν and ν'. We use (81) to write:

$$\begin{aligned}
\|\nabla f_{\mu,\nu}(z) - \nabla f_{\mu',\nu'}(z')\| &= \|\mathbb{E}_{\pi_1}[\nabla_1 k(z, x) - \nabla_1 k(z', x')] - \mathbb{E}_{\pi_2}[\nabla_1 k(z, y) - \nabla_1 k(z', y')]\| \\
&\leq \mathbb{E}_{\pi_1}[\|\nabla_1 k(z, x) - \nabla_1 k(z', x')\|] + \mathbb{E}_{\pi_2}[\|\nabla_1 k(z, y) - \nabla_1 k(z', y')\|] \\
&\leq L\left(\|z - z'\| + \mathbb{E}_{\pi_1}[\|x - x'\|] + \|z - z'\| + \mathbb{E}_{\pi_2}[\|y - y'\|]\right) \\
&\leq L(2\|z - z'\| + W_2(\mu, \mu') + W_2(\nu, \nu'))
\end{aligned}$$

The second line is obtained by convexity while the third one uses Assumption **(A)** and finally the last line relies on π_1 and π_2 being optimal. The desired bound is obtained by further upper-bounding the last two terms by twice their amount. □

Lemma 20. *Let U be an open set, q a probability distribution in $\mathcal{P}_2(\mathcal{X} \times \mathcal{U})$ and ψ and ϕ two measurable maps from $\mathcal{X} \times \mathcal{U}$ to $\mathcal{X}$ which are square-integrable w.r.t q. Consider the path ρ_t from $(\psi)_{\#}q$ and $(\psi + \phi)_{\#}q$ given by: $\rho_t = (\psi + t\phi)_{\#}q \quad \forall t \in [0, 1]$. Under Assumption **(A)**, $\mathcal{F}(\rho_t)$ is differentiable in t with*

$$\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(\psi(x, u) + t\phi(x, u))\phi(x, u)\, \mathrm{d}q(x, u)$$

where f_{μ,ρ_t} is the witness function between μ and ρ_t as defined in (1). Moreover:

$$|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_s)| \leq 3L|t - s| \int \|\phi(x, u)\|^2 dq(x, u)$$

Proof. For simplicity, we write f_t instead of f_{μ,ρ_t} and denote by $s_t(x, u) = \psi(x, u) + t\phi(x, u)$ The function $h : t \mapsto k(s_t(x, u), s_t(x', u')) - k(s_t(x, u), z) - k(s_t(x', u'), z)$ is differentiable for all (x, u),(x', u') in $\mathcal{X} \times \mathcal{U}$ and $z \in \mathcal{X}$. Moreover, by Assumption **(A)**, a simple computation shows that for all $0 \leq t \leq 1$:

$$|\dot{h}| \leq L\left[(\|z - \phi(x, u)\| + \|\psi(x, u)\|)\|\phi(x', u')\| + (\|z - \phi(x', u')\| + \|\psi(x', u')\|)\|\phi(x, u)\|\right]$$

The right hand side of the above inequality is integrable when z, (x, u) and (x', u') are independent and such that $z \sim \mu$ and both (x, u) and (x', u') are distributed according to q. Therefore, by the differentiation lemma [29, Theorem 6.28] it follows that $\mathcal{F}(\rho_t)$ is differentiable and:

$$\dot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[(\nabla_1 k(s_t(x, u), s_t(x', u')) - \nabla_1 k(s_t(x, u), z)).\phi(x, u)\right]. \tag{83}$$

By Proposition 19, we directly get $\dot{\mathcal{F}}(\rho_t) = \int \nabla f_{\mu,\rho_t}(\psi(x, u) + t\phi(x, u))\phi(x, u)\, \mathrm{d}q(x, u)$. We shall control now the difference $|\dot{F}(\rho_t) - \dot{\mathcal{F}}(\rho_{t'})|$ for $0 \leq t, t' \leq 1$. Using Assumption **(A)** and recalling that $s_t(x, u) - s_{t'}(x, u) = (t - t')\phi(x, u)$ a simple computation shows:

$$\begin{aligned}
|\dot{\mathcal{F}}(\rho_t) - \dot{\mathcal{F}}(\rho_{t'})| &\leq L|t - t'|\mathbb{E}\left[(2\|\phi(x, u)\| + \|\phi(x', u')\|)\|\phi(x, u)\|\right] \\
&\leq L|t - t'|(2\mathbb{E}\left[\|\phi(x, u)\|^2\right] + \mathbb{E}\left[\|\phi(x, u)\|\right]^2) \\
&\leq 3L|t - t'| \int \|\phi(x, u)\|^2\, \mathrm{d}q(x, u).
\end{aligned}$$

which gives the desired upper-bound. □

Lemma 21. *Let q be a probability distribution in $\mathcal{P}_2(\mathcal{X} \times \mathcal{X})$ and ψ and ϕ two measurable maps from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{X}$ which are square-integrable w.r.t q. Consider the path ρ_t from $(\psi)_{\#}q$ and $(\psi + \phi)_{\#}q$ given by: $\rho_t = (\psi + t\phi)_{\#}q \quad \forall t \in [0,1]$. Under Assumptions (A) and (B), $\mathcal{F}(\rho_t)$ is twice differentiable in t with*

$$\begin{aligned}\ddot{\mathcal{F}}(\rho_t) =& \mathbb{E}\left[\phi(x,y)^T \nabla_1 \nabla_2 k(s_t(x,y), s_t(x',y'))\phi(x',y')\right] \\ &+ \mathbb{E}\left[\phi(x,y)^T (H_1 k(s_t(x,y), y'_t) - H_1 k(s_t(x,y), z))\phi(x,y)\right]\end{aligned}$$

where (x,y) and (x',y') are independent samples from q, z is a sample from μ and $s_t(x,y) = \psi(x,y) + t\phi(x,y)$. Moreover, if Assumption (C) also holds then:

$$\ddot{\mathcal{F}}(\rho_t) \geq \mathbb{E}\left[\phi(x,y)^T \nabla_1 \nabla_2 k(s_t(x,y), s_t(x',y'))\phi(x',y')\right] - \sqrt{2}\lambda d \mathcal{F}(\rho_t)^{\frac{1}{2}} \mathbb{E}[\|\phi(x,y)\|^2]$$

where we recall that $\mathcal{X} \subset \mathbb{R}^d$.

Proof. The first part is similar to Lemma 20. In fact we already know by Lemma 20 that $\dot{\mathcal{F}}(\rho_t)$ exists and is given by:

$$\dot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[(\nabla_1 k(s_t(x,y), s_t(x',y')) - \nabla_1 k(s_t(x,y), z)).\phi(x,y)\right]$$

Define now the function $\xi : t \mapsto (\nabla_1 k(s_t(x,y), s_t(x',y')) - \nabla_1 k(s_t(x,y), z)).\phi(x,y)$ which is differentiable for all (x,y),(x',y') in $\mathcal{X} \times \mathcal{X}$ and $z \in \mathcal{X}$ by Assumption (B). Moreover, its time derivative is given by:

$$\begin{aligned}\dot{\xi} =& \phi(x',y')^T \nabla_2 \nabla_1 k(s_t(x,y), s_t(x',y'))\phi(x,y) &(84)\\ &+ \phi(x,y)^T (H_1 k(s_t(x,y), s_t(x',y')) - H_1 k(s_t(x,y), z))\phi(x,y) &(85)\end{aligned}$$

where $H_1 k$ denotes the Hessian of k. By Assumption (A) it follows in particular that $\nabla_2 \nabla_1 k$ and $H_1 k$ are bounded hence $|\dot{\xi}|$ is upper-bounded by $(\|\phi(x,y)\| + \|\phi(x',u')\|)\|\phi(x,y)\|$ which is integrable. Therefore, by the differentiation lemma [29, Theorem 6.28] it follows that $\dot{\mathcal{F}}(\rho_t)$ is differentiable and $\ddot{\mathcal{F}}(\rho_t) = \mathbb{E}\left[\dot{\xi}\right]$. We prove now the second statement. Bu the reproducing property, it is easy to see that the last term in the expression of $\dot{\xi}$ can be written as:

$$\langle \phi(x,y)^T H_1 k(s_t(x,y), .)\phi(x,y), k(s_t(x',y'), .) - k(z, .)\rangle_{\mathcal{H}}$$

Now, taking the expectation w.r.t x' ,y' and z which can be exchanged with the inner-product in $\mathcal{H}$ since $(x',y',z) \mapsto k(s_t(x',y'), .) - k(z, .)$ is Bochner integrable [41, Definition 1, Theorem 6] and recalling that such integral is given by f_{μ,ρ_t} one gets the following expression:

$$\langle \phi(x,y)^T H_1 k(s_t(x,y), .)\phi(x,y), f_{\mu,\rho_t}\rangle_{\mathcal{H}}$$

Using Cauchy-Schwartz and Assumption (C) it follows that:

$$|\langle \phi(x,y)^T H_1 k(s_t(x,y), .)\phi(x,y), f_{\mu,\rho_t}\rangle_{\mathcal{H}}| \leq \lambda d \|\phi(x,y)\|^2 \|f_{\mu,\rho_t}\|$$

One then concludes using the expression of $\ddot{\mathcal{F}}(\rho_t)$ and recalling that $\mathcal{F}(\rho_t) = \frac{1}{2}\|f_{\mu,\rho_t}\|^2$. □

Lemma 22. *Assume that for any geodesic $(\rho_t)_{t\in[0,1]}$ between ρ_0 and ρ_1 in $\mathcal{P}(\mathcal{X})$ with velocity vectors $(v_t)_{t\in[0,1]}$ the following holds:*

$$\ddot{\mathcal{F}}(\rho_t) \geq \Lambda(\rho_t, v_t)$$

for some admissible functional Λ as defined in Definition 3, then:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \int_0^1 \Lambda(\rho_s, v_s) G(s,t) ds$$

with $G(s,t) = s(1-t)\mathbb{1}\{s \leq t\} + t(1-s)\mathbb{1}\{s \geq t\}$ for $0 \leq s,t \leq 1$.

Proof. This is a direct consequence of the general identity ([53], Proposition 16.2). Indeed, for any continuous function ϕ on $[0,1]$ with second derivative $\ddot{\phi}$ that is bounded below in distribution sense the following identity holds:

$$\phi(t) = (1-t)\phi(0) + t\phi(1) - \int_0^1 \ddot{\phi}(s) G(s,t) ds.$$

This holds a fortiori for $\mathcal{F}(\rho_t)$ since $\mathcal{F}$ is smooth. By assumption, we have that $\ddot{\mathcal{F}}(\rho_t) \geq \Lambda(\rho_t, V_t)$, hence, it follows that:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \int_0^1 \Lambda(\rho_s, v_s) G(s,t) ds.$$

□

Lemma 23. *[Mixture convexity] The functional $\mathcal{F}$ is mixture convex: for any probability distributions ν_1 and ν_2 and scalar $1 \leq \lambda \leq 1$:*

$$\mathcal{F}(\lambda\nu_1 + (1-\lambda)\nu_2) \leq \lambda\mathcal{F}(\nu_1) + (1-\lambda)\mathcal{F}(\nu_2)$$

Proof. Let ν and ν' be two probability distributions and $0 \leq \lambda \leq 1$. Expanding the RKHS norm in $\mathcal{F}$ it follows directly that:

$$\mathcal{F}(\lambda\nu + (1-\lambda)\nu') - \lambda\mathcal{F}(\nu) - (1-\lambda)\mathcal{F}(\nu') = -\frac{1}{2}\lambda(1-\lambda)MMD(\nu,\nu')^2 \leq 0.$$

which concludes the proof. □

Lemma 24. *[Discrete Gronwall lemma] Let $a_{n+1} \leq (1+\gamma A)a_n + b$ with $\gamma > 0$, $A > 0$, $b > 0$ and $a_0 = 0$, then:*

$$a_n \leq \frac{b}{\gamma A}(e^{n\gamma A} - 1).$$

Proof. Using the recursion, it is easy to see that for any $n > 0$:

$$a_n \leq (1+\gamma A)^n a_0 + b(\sum_{i=0}^{n-1}(1+\gamma A)^k)$$

One concludes using the identity $\sum_{i=0}^{n-1}(1+\gamma A)^k = \frac{1}{\gamma A}((1+\gamma A)^n - 1)$ and recalling that $(1+\gamma A)^n \leq e^{n\gamma A}$. □