


Gradient Flows in the Geometry of the Sinkhorn Divergence

Mathis Hardion

under the supervision of Hugo Lavenant

2024

Abstract

The pioneering work of Jordan, Kinderlehrer, Otto has shown that some meaningful Partial Differential Equations (PDEs) such as the diffusion equation, the continuity equation and more generally Fokker-Planck equations can be interpreted as gradient flows in the space of probability measures endowed with the Wasserstein distance from Optimal Transport (OT). Numerically speaking, the Wasserstein distance can be computationally expensive but its entropic regularization is more accessible thanks to Sinkhorn's algorithm. The debiased version of entropic OT, named the Sinkhorn divergence, has been widely utilized as approximation of the Wasserstein distance, but is emerging as an object of interest in itself due to its smoothness and statistical properties. In this thesis, we study gradient flows in the space of probability measures endowed with the Sinkhorn divergence, by utilizing the recently introduced Riemannian structure based on its local expansion. We focus on the case of a potential energy, and derive the differential equation corresponding to its gradient flow (the validity of the limit of the Sinkhorn JKO scheme is shown for finite spaces as the general case is much more involved). After a change of variables, this equation appears as a constrained rotational motion on the sphere of a Reproducing Kernel Hilbert Space (RKHS), and the particular case of the flow of a single Dirac mass coïncides with the classical gradient flow of a potential when the latter is convex. We obtain existence, uniqueness, contractivity of a solution as well as its convergence in time to the minimum of the energy for a large class of potentials including non-convex ones. We use a simple numerical scheme to obtain illustrations of this theory.

Acknowledgements

I am deeply grateful to my supervisor Hugo Lavenant for his guidance and invaluable assistance on this project, his insightful advice, and for providing me the opportunity to study this subject.

I would also like to thank my professor Gabriel Peyré for being a member of the jury and for his exceptional teaching which got me interested in Optimal Transport.

Many thanks to my friends, family, and girlfriend for their unwavering support which was a tremendous help.

Contents

1 Introduction

1.1 Gradient flows in metric spaces

We begin by briefly introducing gradient flows in the Euclidean case in order to give intuition on how to generalize them to metric spaces. The reader may find a more comprehensive review of gradient flows in metric spaces in [1], and the full theory in the seminal book [2].

Definition 1.1. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. A curve $(x_t)_{t\geq 0} \subset \mathbb{R}^d$ is said follow the **gradient flow** of F starting at $x^0 \in \mathbb{R}^d$ if it is solution of the Cauchy problem*

$$\begin{cases} \forall t \geq 0, \dot{x}_t = -\nabla F(x_t), \\ x_0 = x^0. \end{cases} \tag{1.1}$$

*If F is convex (and potentially not differentiable), a **subgradient flow** is defined analogously as a curve verifying*

$$\begin{cases} \forall t \geq 0, \dot{x}_t \in -\partial F(x_t), \\ x_0 = x^0. \end{cases} \tag{1.2}$$

Existence of a gradient flow in the case where F has Lipschitz gradient is directly given by the Cauchy-Lipschitz theorem, but this assumption is in fact not necessary: one can obtain a solution by discretizing (1.1) or (1.2) appropriately, then taking the limit as the time step goes to 0. This discretization corresponds to the **implicit Euler scheme** widely used for better stability when numerically approximating differential equations, defined through the iterates

$$\frac{x_{k+1}^\tau - x_k^\tau}{\tau} = -\nabla F(x_{k+1}^\tau). \tag{1.3}$$

Observe that the left hand side can be rewritten as the gradient of the function

$$g_k^\tau : x \mapsto \frac{\|x - x_k^\tau\|_2^2}{2\tau} \tag{1.4}$$

at x_{k+1}^τ, and as a result the scheme can be written

$$\nabla \left(g_k^\tau + F\right)\left(x_{k+1}^\tau\right) = 0. \tag{1.5}$$

Fermat's rule gives that the above is verified for

$$x_{k+1}^\tau \in \arg\min_x F(x) + \frac{\|x - x_k^\tau\|_2^2}{2\tau} \tag{1.6}$$

under the extra assumption that F is lower bounded so that such a minimum exists. The formulation (1.6) can in fact be defined with the only assumptions that F is lower semi continuous (l.s.c.) and lower bounded to have existence of a minimizer at each step, and it has the benefit of making no use of the notion of gradient, only that of a metric. We can therefore define, for a metric space $(\mathbb{X}, \mathsf{d})$ and a l.s.c. lower bounded functional $F : \mathbb{X} \to \mathbb{R}$, the scheme

$$x_{k+1}^\tau \in \arg\min_x F(x) + \frac{\mathsf{d}\left(x, x_k^\tau\right)^2}{2\tau} \tag{1.7}$$

and deduce the definition of Generalized Minimizing Movements (GMMs) in the sense of De Giorgi [3].

Definition 1.2. *A curve $(x_t)_t$ in a metric space $(\mathbb{X}, \mathsf{d})$ is said to be a* ***Generalized Minimizing Movement*** *if there exists a sequence of time steps $(\tau_n)_n$ converging to 0 such that the piecewise constant interpolation of the sequence (1.7) given by*

$$\overline{x}_t^n = x_k^{\tau_n} \text{ for } t \in [k\tau_n, (k+1)\tau_n) \tag{1.8}$$

converges uniformly to x.

This can give a preliminary idea of gradient flow, but the drawback is that this definition does not characterize the equation that such GMMs should verify, as the ODE "$\dot{x}_t = -\nabla F(x_t)$" does not make sense in this context since neither the time derivative nor the gradient of a functional are well defined in an arbitrary metric space. To remedy this, the theory developed in [2] finds two alternatives definitions that extend gradient flows in Euclidean space and make use of only metric considerations, namely the Energy Dissipation Equality (EDE) and Evolution Variational Inequality (EVI). We do not dive into this formalism as we shall see that it is possible to get limiting equations in our particular case of probability measures with simpler considerations.

1.2 The Wasserstein space and JKO flows

The theory of gradient flows in metric spaces finds its main applications on the space of probability measures endowed with the Monge-Kantorovitch distance of Optimal Transport (OT), also known as Wasserstein distance, which we now briefly introduce. In this thesis, $\mathcal{X}$ will denote a compact metric space, $\mathcal{P}(\mathcal{X})$ is the set of probability measures on $\mathcal{X}$, $\mathcal{C}(\mathcal{X})$ is the Banach space of continuous functions on $\mathcal{X}$ endowed with the supremum norm $\|\cdot\|_\infty$, whose dual is identified with the set $\mathcal{M}(\mathcal{X})$ of signed Radon measures on $\mathcal{X}$ through the Markov-Riesz-Kakutani theorem [4, Theorem 6.19]. The set of couplings with marginals $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is denoted

$$\Pi(\mu, \nu) \coloneqq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}), (P_1)_\sharp \pi = \mu, (P_2)_\sharp = \nu\}$$

with $(P_i)_\sharp \pi$ the ith marginal of π ($i \in \{1, 2\}$) i.e. its pushforward by the projection P_i onto the ith component of the product space. For a cost $c \in \mathcal{C}(\mathcal{X} \times \mathcal{X})$, the optimal transport problem intuitively consists in finding the minimal total cost to move the mass from one measure to another, which defines the OT cost denoted for $\mu, \nu \in \mathcal{P}(\mathcal{X})$ as [5]

$$\mathrm{OT}_0(\mu, \nu) \coloneqq \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y) d\pi(x, y). \tag{1.9}$$

When the cost c is the square distance on $\mathcal{X} \times \mathcal{X}$, the Wasserstein distance is defined as $\mathbb{W}_2(\mu, \nu) \coloneqq \sqrt{\mathrm{OT}_0(\mu, \nu)}$. It defines a proper metric, metrizes the weak-$*$ topology (i.e. convergence in law), makes the so-called **Wasserstein space** $(\mathcal{P}(\mathcal{X}), \mathbb{W}_2)$ a Polish space [5, Section 6] and is aware of the geometry of the ambient space $\mathcal{X}$ as it is directly influenced by the distance between supports. It additionally has deep ties with partial differential equations (PDEs) through the GMMs presented above: indeed, one can consider the discrete scheme (1.7) in the Wasserstein space i.e. take $(\mathbb{X}, \mathsf{d}) = (\mathcal{P}(\mathcal{X}), \mathbb{W}_2)$. This is referred to as the **JKO scheme**, named after the authors of the seminal paper [6] who have shown that when $\mathcal{X}$ is a convex subset of $\mathbb{R}^d$ endowed with the Euclidean norm, the corresponding GMMs for 'free energy' functionals describe meaningful partial differential equations, the Fokker-Planck equations. A first particular case of importance is the **continuity equation** (or **transport equation**) for a gradient vector field: for a differentiable potential $V \in \mathcal{C}(\mathcal{X})$, it writes

$$\dot{\mu}_t = \mathrm{div}\,(\mu_t \nabla V) \tag{1.10}$$

where the curve $(\mu_t)_t$ is valued in $\mathcal{P}(\mathcal{X})$, and the notion of solution is understood in an appropriate weak sense [7, Definition 4.1]. It is a recurring equation when modeling mass conservation in fluid dynamics [8, Section 1.4], charge conservation in electromagnetism [9, Section 4.2] and others. When thinking of a measure as a distribution of particles, this equation is interpreted as the fact that every particle follows the classical

gradient flow of V [7, Theorem 4.4]. The work [6] further reformulates it as the gradient flow with respect to the Wasserstein distance of the **potential energy** functional

$$\mu \mapsto \langle \mu, V \rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the duality product in a Banach space, here $\langle \mu, V \rangle = \int V d\mu$. A second important case is the **diffusion equation** which writes for a curve of probability density functions $(\rho_t)_t$ with respect to the Lebesgue measure

$$\dot{\rho}_t = \Delta \rho_t \tag{1.11}$$

where Δ denotes the Laplacian. This equation describes the law of a standard Brownian motion and is central in stochastic modeling [10]. It is also prevalent in physics when modeling diffusive phenomena. The paper [6] inteprets this equation as the gradient flow of the **entropy functional**

$$\rho \mapsto \int \log(\rho(x))\rho(x)dx.$$

The more general Fokker-Planck equations are linear combinations of the previous two cases, and are related to a class of stochastic differential equations. They are thus the gradient flow of the **free energy** functional being a linear combination of a potential energy and the entropy.

Multiple other meaningful PDEs have since been interpreted as gradient flows in the Wasserstein space, such as the porous medium equation [11], the Keller-Segel equation [12] and more. Wasserstein gradient flows were also used to model crowd motion by describing a PDE that cannot be analyzed through classical techniques in [13], showing the versatility of this framework. They are also emerging as a useful tool in machine learning, for instance to model particle dynamics [14] or improve Generative Adversarial Network training [15].

1.3 The Sinkhorn divergence

From a numerical perspective, one of the main drawbacks of methods using the Wasserstein distance is its high computational cost. One can compute (1.9) as a linear program when the input measures are discrete [16, Section 3], however this can become untractable for a large number of points, as the size of the matrices c and π are $n \times m$ for n, m the respective numbers of Dirac masses in the input and output measures. The introduction of an entropic regularization allows for a much faster optimization procedure thanks to Sinkhorn's algorithm [17], giving the Entropic OT loss defined as follows: for a regularization parameter $\varepsilon > 0$,

$$\mathrm{OT}_\varepsilon(\mu, \nu) \coloneqq \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x, y)d\pi(x, y) + \varepsilon \mathrm{KL}\left(\pi | \mu \otimes \nu\right) \tag{1.12}$$

where the KL divergence is given by

$$\mathrm{KL}\left(\pi|\gamma\right) \coloneqq \int_{\mathcal{X}\times\mathcal{X}} \log\left(\frac{d\pi}{d\gamma}\right) d\pi - \int_{\mathcal{X}\times\mathcal{X}} d\pi + \int_{\mathcal{X}\times\mathcal{X}} d\gamma$$

when π is absolutely continuous with respect to γ, and $+\infty$ otherwise. This loss approximates OT_0 in the sense that it is recovered when $\varepsilon \to 0$ [18], and it is therefore still aware of the geometry of $\mathcal{X}$ so long as ε does not get large. On top of better numerical tractability [16], other works have shown that this loss has better statistical complexity with regards to the curse of dimensionality [19] and is smooth with efficiently computed gradients [20]. One main drawback is that OT_ε has an inherent bias in the sense that generally speaking $\mathrm{OT}_\varepsilon(\mu, \mu) \neq 0$, and there may even be measures ν such that $\mathrm{OT}_\varepsilon(\mu, \nu) < \mathrm{OT}_\varepsilon(\mu, \mu)$. This motivated [20] to introduce the **Sinkhorn divergence** by removing this bias, namely

$$S_\varepsilon(\mu, \nu) \coloneqq \mathrm{OT}_\varepsilon(\mu, \nu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\mu, \mu) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\nu, \nu). \tag{1.13}$$

This is indeed a divergence in the sense of the following theorem proven by [21], which also provides further desirable properties.

Theorem 1.3 ([21, Theorem 1]). *Assume that $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is a Lipschitz cost function such that $k_c : (x, y) \mapsto e^{-\frac{c(x,y)}{\varepsilon}}$ is a positive definite universal kernel. Then S_ε is symmetric positive definite, smooth, convex in each of its input variables and metrizes the weak-$*$ convergence.*

Here 'positive definite universal kernel' is understood in the framework of Reproducing Kernel Hilbert Spaces (RKHS) [22, 23], this assumption is satisfied in the usual case where the cost c is the Euclidean distance on $\mathcal{X} \subset \mathbb{R}^d$. With these statistical and numerical advantages along with its smoothness, the Sinkhorn divergence is emerging as an object of interest in itself rather than only as an approximation of Optimal Transport.

1.4 Gradient flows in the geometry of the Sinkhorn divergence

Since JKO flows in the Wasserstein geometry discretize important PDEs, but the Monge-Kantorovitch distance can be difficult to compute numerically, a natural idea is to approximate it with Entropic OT. This idea has been studied in a few articles. The first [24] studies the scheme

$$\mu_{k+1}^\tau = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathcal{X})} F(\mu) + \tfrac{1}{2\tau} \mathrm{OT}_\varepsilon(\mu, \mu_k^\tau) \tag{1.14}$$

by reformulating it as a proximal stepping, with respect to the KL divergence, of a sum of functionals (one corresponding to F and another enforcing the constraints). The author takes advantage of this structure to introduce a computational scheme utilizing Dykstra's algorithm for Bregman divergences [25], in order to efficiently perform each JKO step. The resulting numerical scheme enjoys low computational cost and is shown to be versatile as it can deal with attraction (i.e. minimizing distance to a target distribution), congestion (i.e. capped density) and multiple interacting measures. The main drawback is the bias induced by OT_ε, which when looking at the limit $\tau \to 0$ for a fixed $\varepsilon > 0$ could induce discontinuities in time (since $\mathrm{OT}_\varepsilon(\mu, \mu) > 0$ generically). This leads to the need of taking $\varepsilon \to 0$ as well during the analysis, which was studied by the follow-up article [26]. This work gives a self contained proof that OT_ε converges to the Wasserstein distance as $\varepsilon \to 0$ in the sense of Γ convergence, and that the corresponding gradient flows also converge to the solution of the Fokker-Planck equation when both ε and the time step τ jointly vanish at a suitable rate, namely $\varepsilon |\log \varepsilon| = O(\tau^2)$ [26, Theorem 3.16]. The authors highlight that one must at least have $\varepsilon = o(\tau)$ to hope for convergence [26, Proposition 3.6], and an ongoing work [27] shows it is sufficient. The latter work also shows that when taking $\varepsilon = \theta\tau$ for some $\theta > 0$, an entropic term is added to the limiting equation. The results of [26] were later extended to a slightly more general class of PDEs in [28], which includes nonlinear kinetic Fokker-Planck and Kolmogorov-type diffusion equations.

A notable pitfall is that the condition $\varepsilon = o(\tau)$ (or even $\varepsilon = \mathcal{O}(\tau)$) can be challenging to ask from a computational standpoint, since to approximate the continuous solution of a PDE we ideally want τ as small as possible, while a vanishing ε makes the convergence of Sinkhorn's algorithm much slower, even with state-of-the-art ε-scaling methods (the convergence rate is around $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$, see [29, Proposition 18 and Section 5.2]). It could therefore be beneficial to consider the limit $\tau \to 0$ for a **fixed** value of ε, and to use the debiased Sinkhorn divergence instead of OT_ε to remedy the discontinuity problem mentioned in [24]. We have seen Section 1.3 that the Sinkhorn divergence has desirable properties lost when ε vanishes, which further motivates keeping $\varepsilon > 0$ fixed. This leads to the following **Sinkhorn-JKO (SJKO)** scheme which is the starting point of this thesis:

$$\mu_{k+1}^\tau = \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathcal{X})} F(\mu) + \tfrac{1}{2\tau} S_\varepsilon(\mu, \mu_k^\tau). \tag{1.15}$$

As of writing, we have been unable to find literature studying this de-biased scheme and its limit $\tau \to 0$ for fixed $\varepsilon > 0$. This is in part due to the fact that this problem requires a finer understanding of the local behavior

of the Sinkhorn divergence and the geometry it induces, which was only recently studied in the work [30] where the authors exhibit a Riemannian structure on the space of probability measures which locally agrees with the Sinkhorn divergence. To do so, they also embed the space of probability measures into a subset of the sphere in a RKHS through a change of variables. We will further detail the results of this paper in Section 2 as it will lay the foundations necessary for our analysis.

1.5 Contributions and outline

In this thesis, we focus on the case of a potential energy as it is the most well-understood in the Wasserstein case. The structure of this report is as follows, with an emphasis on the distinction between literature review and novel results.

Review of Entropic OT and its geometry:

- In Section 2, we first present basic results about the dual of the Entropic OT problem and most importantly the recent results of [30], as a thorough understanding of this work is needed to proceed with our analysis.

Novel results:

- A reformulation of the metric tensor defined by [30] after the change of variables is also provided in Section 2, Lemma 2.12 (it is the only original result of that section).
- In Section 3, we derive (informally at first) the equation describing the GMM corresponding to the Sinkhorn divergence, then properly compute its equivalent after the change of variables studied in [30], and further rewrite it to exhibit the structure of a constrained rotational motion on the RKHS sphere. As an example, the Sinkhorn flow of a single Dirac mass is shown to correspond to the Wasserstein case (i.e. the classical gradient flow of the location of the mass) when the potential is smooth and convex (Proposition 3.14).
- Section 4 is dedicated to the analysis of the main properties of the flow. Existence, uniqueness, and contractivity of the flow (Theorem 4.1) are proven Section 4.1 using a discretization of the space. The long time behavior of the solution is also investigated Section 4.2, where convergence to the minimum of the energy is proven for a broad class of potentials including non-convex ones (Theorem 4.8).
- In Section 5, the validity of the limit of the SJKO scheme as the time step goes to 0 is proven when the ambient space $\mathcal{X}$ is a finite set of points (Theorem 5.1). We provide generalizations of some results of [30, Section 3] as required by the proof. We also discuss the reason why the general case is more involved in Remark 5.8.
- Finally, we present in Section 6 a simple gradient descent scheme to numerically compute SJKO steps and use it to illustrate the theory on simple examples. The rotational structure of the motion is observed on the three point space, the Sinkhorn flow of a Dirac mass is compared to its classical flow, and Sinkhorn flows for both convex and non convex potentials are computed in the Eulerian and Lagrangian discretizations to exemplify different behaviors and compare with Wasserstein flows.

2 Preliminaries on Entropic OT and its Riemannian geometry

2.1 Dual problem

The Entropic OT problem (1.12) has the following dual formulation [16, Remark 4.24]

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \sup_{f,g\in\mathcal{C}(\mathcal{X})} \langle \mu, f\rangle + \langle \nu, g\rangle - \varepsilon \left\langle \mu\otimes\nu, \exp\left(\tfrac{1}{\varepsilon}\left(f\oplus g - c\right)\right) - 1\right\rangle \tag{2.1}$$

with the notation $f\oplus g : (x,y)\mapsto f(x)+g(y)$. This problem has maximizers that we call **Schrödinger potentials** $f_{\mu,\nu}, g_{\mu,\nu}$, which are solutions of the Schrödinger system

$$\begin{cases} f_{\mu,\nu} = T_\varepsilon(g_{\mu,\nu},\nu) \\ g_{\mu,\nu} = T_\varepsilon(f_{\mu,\nu},\mu) \end{cases} \tag{2.2}$$

where we define the **Sinkhorn mapping**

$$T_\varepsilon : \left| \begin{array}{rcl} \mathcal{C}(\mathcal{X})\times\mathcal{P}(\mathcal{X}) & \longrightarrow & \mathcal{C}(\mathcal{X}) \\ f,\mu & \longmapsto & -\varepsilon\log\displaystyle\int_{\mathcal{X}} \exp\left(\tfrac{1}{\varepsilon}\left(f(x)-c(x,\cdot)\right)\right) d\mu(x). \end{array} \right. \tag{2.3}$$

These potentials are unique up to constant shifts i.e. the set $\{(f_{\mu,\nu}+\lambda, g_{\mu,\nu}-\lambda), \lambda\in\mathbb{R}\}$ describes all solutions of (2.2). Moreover, at optimality the loss reads

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \langle \mu, f_{\mu,\nu}\rangle + \langle \nu, g_{\mu,\nu}\rangle \tag{2.4}$$

meaning that the Schrödinger potentials $(f_{\mu,\nu}, g_{\mu,\nu})$ correspond to the gradients of OT_ε with respect to the input measures (μ,ν) (see [21, Proposition 2]). The optimal transport plan in (1.12) is additionally given by

$$\pi = \exp\left(\tfrac{1}{\varepsilon}\left(f_{\mu,\nu}\oplus g_{\mu,\nu} - c\right)\right)(\mu\otimes\nu). \tag{2.5}$$

The symmetry of the optimization problem (2.1) and the uniqueness up to constant shifts of the optimal potentials allows one to choose $f_{\mu,\nu} = g_{\nu,\mu}$ and in particular we can take $f_{\mu,\mu} = g_{\mu,\mu}$ which we shall denote f_μ for short. The Sinkhorn divergence defined by (1.13) is differentiable with gradients given by $(f_{\mu,\nu}-f_\mu, f_{\nu,\mu}-f_\nu)$.

2.2 Riemannian Geometry induced by the Sinkhorn divergence

Despite all the advantages of the Sinkhorn divergence, it lacks the properties to define a metric on $\mathcal{P}(\mathcal{X})$, as $\sqrt{S_\varepsilon}$ does not satisfy the triangle inequality for $\varepsilon > 0$ [30, Section 7.1], contrary to $\sqrt{S_0} = \mathbb{W}_2$. This has motivated the recent work [30] to define a Riemannian structure on $\mathcal{P}(\mathcal{X})$ which keeps the geometric faithfulness and smoothness of the Sinkhorn divergence. This is done by computing its Hessian and using it as metric tensor, which makes the new metric locally agree with the Sinkhorn divergence. As this theory will be the foundation necessary to the analysis developed in the present thesis, we summarize the main notations and results that we will use.

The Hessian is first computed for vertical perturbations, defined as follows.

Definition 2.1 [30, Definition 3.1]. *The set of balanced measures is denoted* $\mathcal{M}_0(\mathcal{X}) \coloneqq \{\sigma\in\mathcal{M}(\mathcal{X}) \mid \langle\sigma,1\rangle = 0\}$. *Let* $\mu\in\mathcal{P}(\mathcal{X})$, $\dot\mu\in\mathcal{M}_0(\mathcal{X})$. *The curve*

$$t\mapsto \mu_t \coloneqq \mu + t\dot\mu \tag{2.6}$$

is a ***vertical perturbation*** *of* μ *if* $\forall t \in (-\tau, \tau), \mu_t \in \mathcal{P}(\mathcal{X})$ *for some* $\tau > 0$.

Some operators need to be introduced as they appear in the expression of the Hessian.

Definition 2.2 [30, Definition 3.2]. *The self transport kernel of* $\mu \in \mathcal{P}(\mathcal{X})$ *is defined as*

$$k_\mu \coloneqq \exp\left(\tfrac{1}{\varepsilon}\left(f_\mu \oplus f_\mu - c\right)\right) \tag{2.7}$$

and the operators $H_\mu : \mathcal{M}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ *and* $K_\mu : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ *are defined as*

$$\forall \nu \in \mathcal{M}(\mathcal{X}),\ H_\mu[\nu] : y \mapsto \int_{\mathcal{X}} k_\mu(x, y) d\nu(x) \tag{2.8}$$

$$\forall \phi \in \mathcal{C}(\mathcal{X}),\ K_\mu[\phi] \coloneqq H_\mu[\phi\mu] : y \mapsto \int_{\mathcal{X}} k_\mu(x, y)\phi(x) d\mu(x). \tag{2.9}$$

The set of balanced measures $\mathcal{M}_0(\mathcal{X})$ corresponding to derivatives of vertical perturbations is the dual of $\mathcal{C}(\mathcal{X})/\mathbb{R}$, the space of functions up to constant shifts (i.e. $\mathcal{C}(\mathcal{X})$ quotiented by the relation $f \sim g \iff \exists \lambda \in \mathbb{R}, f = g + \lambda$). One can always map a function of $\mathcal{C}(\mathcal{X})$ to its equivalence class in $\mathcal{C}(\mathcal{X})/\mathbb{R}$ to work in that space.

The expansion of the Sinkhorn divergence is then given by the following theorem.

Theorem 2.3 [30, Theorem 3.3]. *Let* $(\mu_t)_t$ *be a vertical perturbation given by* (2.6). *Then*

$$\frac{1}{t^2} S_\varepsilon(\mu, \mu_t) \xrightarrow[t\to 0]{} \frac{\varepsilon}{2} \left\langle \dot{\mu}, (\mathrm{Id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \right\rangle. \tag{2.10}$$

In (2.10), the operator $(\mathrm{Id} - K_\mu^2)^{-1}$ is well defined as a bounded operator on $\mathcal{C}(\mathcal{X})/\mathbb{R}$ [30, Theorem 3.8]. This is because the optimality condition (2.2) for the self transport of μ can be rewritten as $K_\mu[1] = 1$, giving $(\mathrm{Id} - K_\mu^2)[1] = 0$ and obstructing the invertibility on the whole space $\mathcal{C}(\mathcal{X})$. The duality product in (2.10) is therefore between $\mathcal{M}_0(\mathcal{X})$ and $\mathcal{C}(\mathcal{X})/\mathbb{R}$.

Following the expression of the Hessian, the authors of [30] define the metric tensor as follows.

Definition 2.4 [30, Definition 4.1]. *For* $\mu \in \mathcal{P}(\mathcal{X})$, $\dot{\mu}_1, \dot{\mu}_2 \in \mathcal{M}_0(\mathcal{X})$, *define the metric tensor at* μ *as*

$$\mathbf{g}_\mu(\dot{\mu}_1, \dot{\mu}_2) \coloneqq \langle \dot{\mu}_1, G_\mu[\dot{\mu}_2] \rangle \tag{2.11}$$

where

$$G_\mu \coloneqq \tfrac{\varepsilon}{2}(\mathrm{Id} - K_\mu^2)^{-1} H_\mu : \mathcal{M}_0(\mathcal{X}) \to \mathcal{C}(\mathcal{X})/\mathbb{R}. \tag{2.12}$$

The fact that the local behavior of the Sinkhorn divergence is given by this metric tensor will make the SJKO flows correspond to gradient flows in the manifold $(\mathcal{P}(\mathcal{X}), \mathbf{g})$. The proof of Theorem 2.3 involves computing the time derivatives of the Schrödinger potentials along curves, and this expression will appear when deriving the limit of the JKO flow.

Proposition 2.5 [30, Proposition 3.11]. *Let* μ_t *be given by* (2.6). *Then the potentials* $f_{t,s} \coloneqq f_{\mu_t,\mu_s}$ *are continuously differentiable in* $\mathcal{C}(\mathcal{X})/\mathbb{R}$ *with respect to* (t, s) *in a neighborhood of* $(0, 0)$, *and we have*

$$\left.\frac{\partial f_{0,s}}{\partial s}\right|_{s=0} = -2G_\mu[\dot{\mu}]. \tag{2.13}$$

Next, the article analyzes the metric tensor (2.11) and extends it to the (Cauchy) completion of $\mathcal{M}_0(\mathcal{X})$, which coïncides with a subset of a RKHS. Denoting by $\mathcal{H}_\mu$ the RKHS corresponding to the positive definite (PD)

kernel k_μ, one can see from $\forall \sigma \in \mathcal{M}(\mathcal{X})$, $H_\mu[\sigma] : y \mapsto \langle \sigma, k_\mu(\cdot, y)\rangle$ and the reproducing kernel property that H_μ corresponds exactly to the isometry given by the Riesz representation theorem, i.e. it verifies for any σ

$$\forall \phi \in \mathcal{H}_\mu, \langle \sigma, \phi\rangle = \langle H_\mu[\sigma], \phi\rangle_{\mathcal{H}_\mu}. \tag{2.14}$$

Therefore, it is naturally defined on the space $\mathcal{H}_\mu^*$ of continuous linear functionals on $\mathcal{H}_\mu$, and takes values in $\mathcal{H}_\mu$. Note that $\mathcal{H}_\mu^* \supset \mathcal{M}(\mathcal{X})$ since $\mathcal{H}_\mu \subset \mathcal{C}(\mathcal{X})$. This also means that K_μ is valued in $\mathcal{H}_\mu$. Since there is conservation of mass and thus perturbations $\dot{\mu}$ have zero total mass, we need to instead consider the space

$$\mathcal{H}_{\mu,0}^* \coloneqq \left\{\sigma \in \mathcal{H}_\mu^*, \langle \sigma, 1\rangle = 0\right\} \tag{2.15}$$

which is the dual of $\mathcal{H}_\mu/\mathbb{R}$. H_μ can then be restricted to an operator $\mathcal{H}_{\mu,0}^* \to \mathcal{H}_\mu/\mathbb{R}$ which still corresponds to the Riesz isometry in the Hilbert space $\mathcal{H}_\mu/\mathbb{R}$. The following theorem relates $\mathcal{H}_{\mu,0}^*$ and $\mathcal{M}_0(\mathcal{X})$, making sense of the former as the tangent space of $\mathcal{P}(\mathcal{X})$ endowed with the metric tensor $\mathbf{g}_\mu$.

Theorem 2.6 [30, Theorem 4.5]. *The metric tensor $\mathbf{g}_\mu$ is a positive definite quadratic form on $\mathcal{M}_0(\mathcal{X})$, and the completion of that space with respect to this form is $\mathcal{H}_{\mu,0}^*$. There furthermore exists a constant C such that*

$$\forall \dot{\mu} \in \mathcal{H}_{\mu,0}^*,\ \tfrac{\varepsilon}{2}\|\dot{\mu}\|_{\mathcal{H}_{\mu,0}^*}^2 \le \mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) \le \tfrac{\varepsilon}{2} C \|\dot{\mu}\|_{\mathcal{H}_{\mu,0}^*}^2.$$

The tangent space $\mathcal{H}_{\mu,0}^*$ (isometric to $\mathcal{H}_\mu/\mathbb{R}$) being a subset of $\mathcal{H}_\mu$ which is unfortunately dependant on μ, the authors of [30] perform a change of variables to show that it is isometric to $\mathcal{H}_c$, the RKHS of kernel $k_c \coloneqq \exp\left(-\frac{c}{\varepsilon}\right)$. This change of variables will allow us to simplify the equation of the gradient flow and utilize the structure of $\mathcal{H}_c$. In the following we denote

$$a_\mu \coloneqq \exp\left(\tfrac{1}{\varepsilon} f_\mu\right), \tag{2.16}$$

$$b_\mu \coloneqq \exp\left(-\tfrac{1}{\varepsilon} f_\mu\right) = a_\mu^{-1}. \tag{2.17}$$

Proposition 2.7 [30, Proposition 4.7]. *The map*

$$\iota_\mu : \left| \begin{array}{ccc} \mathcal{H}_c & \longrightarrow & \mathcal{H}_\mu \\ \phi & \longmapsto & a_\mu \phi \end{array} \right.$$

is an isometry.

This gives a change of variables on the tangent space, and the article next considers the corresponding change of variables on the space of probability measures by considering the maps

$$A : \left| \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \longrightarrow & \mathcal{M}(\mathcal{X}) \\ \mu & \longmapsto & a_\mu \mu \end{array} \right. \tag{2.18}$$

and

$$B : \left| \begin{array}{ccc} \mathcal{P}(\mathcal{X}) & \longrightarrow & \mathcal{H}_c \\ \mu & \longmapsto & b_\mu \end{array} \right. \tag{2.19}$$

where we can see that $B(\mu) = H_c[a_\mu \mu]$ from the optimality condition (2.2). These mappings are appropriate changes of variables in the sense of the following theorem, where $\mathcal{M}_+(\mathcal{X})$ denotes the set of nonnegative Radon measures on $\mathcal{X}$.

Theorem 2.8 [30, Theorem 4.8].

(a) The map A is a weak-$*$ homeomorphism onto its image $\mathcal{A} \coloneqq \left\{\alpha \in \mathcal{M}_+(\mathcal{X}) \;\middle|\; \|H_c[\alpha]\|_{\mathcal{H}_c} = 1\right\}$, and $\mathcal{A}$ is weak-$*$ compact.

(b) *The map B is a weak-$*$-to-weak homeomorphism onto its image $\mathcal{B} \coloneqq \left\{b \in H_c\left[\mathcal{M}_+(\mathcal{X})\right] \,\middle|\, \|b\|_{\mathcal{H}_c} = 1\right\}$. Weak and norm convergence agree on the set $\mathcal{B}$ which is weakly and norm compact.*

These results will allow us to work on $\mathcal{B}$ rather than $\mathcal{P}(\mathcal{X})$, and obtain a more interpretable equation (Section 3.2).

To compute the pushforward of the metric tensor $\mathbf{g}$ by this change of variables, the article reviews its differential properties. We briefly introduce the notion of differentiable curve in Topological Vector Spaces (TVSs) used in the paper (in the following, the TVSs in question will be Banach or Hilbert spaces, with their norm, weak or weak-$*$ topologies).

Definition 2.9 [30, Definition 4.10]. *Let $(\mathbb{X}, \mathcal{T})$ be a TVS and I a subinterval of $\mathbb{R}$. A path $(x_s)_{s\in I}$ valued in $\mathbb{X}$ is said to be $\mathcal{T}$-**differentiable** at $t \in I$ if $\frac{1}{u}(x_{t+u} - x_t)$ converges in $\mathcal{T}$ to a limit $\dot{x}_t$ as $u \to 0$. If $\mathbb{X}$ is normed and $\mathcal{T}$ is the norm topology, we simply call the curve **differentiable**.*

The relationship between the derivatives of the curve on $\mathcal{B}$ and that on $\mathcal{P}(\mathcal{X})$ is explicit, as given by the following lemma. We will use this expression to relate the differential equation on $\mathcal{P}(\mathcal{X})$ to that on $\mathcal{B}$.

Lemma 2.10 [30, Lemma 4.12]. *Let $(\mu_t)_t$ be a path in $\mathcal{P}(\mathcal{X})$ such that the curve $(b_t)_t \coloneqq (B(\mu_t))_t$ is continuous and weakly differentiable in $\mathcal{H}_c$ at $t = 0$. Then $(\mu_t)_t$ is also weakly differentiable in $\mathcal{H}_{\mu_0}$ at $t = 0$ and the derivatives verify*

$$H_\mu[\dot{\mu}] = (\mathrm{Id} + K_\mu)[a_\mu \dot{b}] \tag{2.20}$$

where $\mu \coloneqq \mu_0$, $\dot{\mu} \coloneqq \dot{\mu}_0, \dot{b} \coloneqq \dot{b}_0$.

The pushforward of the metric tensor is then given by the following theorem, where for $b \in \mathcal{B}$ we denote $b^\perp \coloneqq \left\{h \in \mathcal{H}_c \,\middle|\, \langle b, h\rangle_{\mathcal{H}_c} = 0\right\}$.

Theorem 2.11 [30, Theorem 4.11]. *Let $(\mu_s)_s$ be a path valued in $\mathcal{P}(\mathcal{X})$ such that the path $(b_s)_s = (B(\mu_s))_s$ is continuous and weakly differentiable in $\mathcal{H}_c$ at t. Then $\left\langle \dot{b}_t, b_t \right\rangle_{\mathcal{H}_c} = 0$, $(\mu_s)_s$ is weakly differentiable at t in $\mathcal{H}^*_{\mu_t}$ and*

$$\mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t) = \tilde{\mathbf{g}}_{\mu_t}(\dot{b}_t, \dot{b}_t)$$

where for $\mu \in \mathcal{P}(\mathcal{X})$, $a \coloneqq a_\mu$, $b \coloneqq B(\mu)$, we define for $\dot{b} \in b^\perp$

$$\tilde{\mathbf{g}}_\mu\left(\dot{b}, \dot{b}\right) \coloneqq \tfrac{\varepsilon}{2}\left(\left\langle \dot{b}, \dot{b}\right\rangle_{\mathcal{H}_c} + 2\left\langle a\dot{b}, (\mathrm{Id} - K_\mu)^{-1}\left[a\dot{b}\right]\right\rangle_{L^2_\mu(\mathcal{X})}\right), \tag{2.21}$$

$L^2_\mu(\mathcal{X})$ being the Hilbert space of functions $\mathcal{X} \to \mathbb{R}$ square integrable against μ.

We provide in this thesis the expression of the operator $\tilde{G}_\mu$ corresponding to G_μ (defined in (2.12)) after the change of variables, as it was absent from the original article. It will help simplify computations when manipulating the metric tensor.

Lemma 2.12. *The metric tensor on $\mathcal{H}_c$ defined by (2.21) can be rewritten as*

$$\tilde{\mathbf{g}}_\mu\left(\dot{b}, \dot{b}\right) = \left\langle \dot{b}, \tilde{G}_\mu\left[\dot{b}\right]\right\rangle_{\mathcal{H}_c} \tag{2.22}$$

where the operator $\tilde{G}_\mu : b^\perp \to b^\perp$ is defined by

$$\tilde{G}_\mu\left[\dot{b}\right] \coloneqq \tfrac{\varepsilon}{2} b\,(\mathrm{Id} + K_\mu)\,(\mathrm{Id} - K_\mu)^{-1}\left[a\dot{b}\right]. \tag{2.23}$$

Proof. One can remark that $\forall \phi, \psi \in \mathcal{H}_\mu, \langle \phi, \psi\rangle_{L^2_\mu(\mathcal{X})} = \langle K_\mu \phi, \psi\rangle_{\mathcal{H}_\mu}$, and thus

$$\tilde{\mathbf{g}}_\mu\left(\dot{b}, \dot{b}\right) = \tfrac{\varepsilon}{2}\left(\left\langle \dot{b}, \dot{b}\right\rangle_{\mathcal{H}_c} + 2\left\langle a\dot{b}, K_\mu\,(\mathrm{Id} - K_\mu)^{-1}\left[a\dot{b}\right]\right\rangle_{\mathcal{H}_\mu}\right)$$

which after using the isometry $\mathcal{H}_\mu \ni \phi \mapsto b\phi \in \mathcal{H}_c$ gives

$$\begin{aligned}\tilde{\mathbf{g}}_\mu\left(\dot{b},\dot{b}\right) &= \tfrac{\varepsilon}{2}\left(\left\langle \dot{b},\dot{b}\right\rangle_{\mathcal{H}_c} + 2\left\langle \dot{b}, bK_\mu\left(\mathrm{Id}-K_\mu\right)^{-1}\left[a\dot{b}\right]\right\rangle_{\mathcal{H}_c}\right)\\ &= \tfrac{\varepsilon}{2}\left\langle \dot{b}, ba\dot{b} + 2bK_\mu\left(\mathrm{Id}-K_\mu\right)^{-1}\left[a\dot{b}\right]\right\rangle_{\mathcal{H}_c}\\ &= \tfrac{\varepsilon}{2}\left\langle \dot{b}, b\left(\mathrm{Id}+2K_\mu\left(\mathrm{Id}-K_\mu\right)^{-1}\right)\left[a\dot{b}\right]\right\rangle_{\mathcal{H}_c}\\ &= \left\langle \dot{b}, \tfrac{\varepsilon}{2}b\left(\mathrm{Id}+K_\mu\right)\left(\mathrm{Id}-K_\mu\right)^{-1}\left[a\dot{b}\right]\right\rangle_{\mathcal{H}_c}.\end{aligned}$$

To see that $\tilde{G}_\mu$ is valued in $b^\perp$, observe that using the same computation

$$\left\langle b, \tilde{G}_\mu\left[\dot{b}\right]\right\rangle_{\mathcal{H}_c} = \tfrac{\varepsilon}{2}\left(\left\langle \dot{b},b\right\rangle_{\mathcal{H}_c} + 2\left\langle a\dot{b}, \left(\mathrm{Id}-K_\mu\right)^{-1}\left[ab\right]\right\rangle_{L^2_\mu(\mathcal{X})}\right)$$

where the first term in the sum on the right hand side is null for $\dot{b}\in b^\perp$ and the second as well since $ab=1$ is a constant and thus null in $\mathcal{H}_\mu/\mathbb{R}$. □

In the article, the authors also get a version of the bound in Theorem 2.6 after the change of variables, given below. These bounds will allow us to work within the Hilbertian structure of $\mathcal{H}_c$ rather than the more complex Riemannian structure of $(\mathcal{B},\tilde{\mathbf{g}})$.

Proposition 2.13 [30, Proposition 4.14]. *Let $\mu\in\mathcal{P}(\mathcal{X}), \dot{b}\in\mathcal{H}_c$ with $\dot{b}\in(B(\mu))^\perp$. Then there exists a constant C such that*

$$\tfrac{\varepsilon}{2}\left\|\dot{b}\right\|^2_{\mathcal{H}_c} \le \tilde{\mathbf{g}}_\mu\left(\dot{b},\dot{b}\right) \le \tfrac{\varepsilon}{2}C\left\|\dot{b}\right\|^2_{\mathcal{H}_c}. \tag{2.24}$$

Using the metric tensor, the article defines the corresponding Riemannian distance by minimizing lengths of paths. We write it here as a distance on $\mathcal{B}$ since it is the space we will work in, but it is in direct correspondence with a distance on $\mathcal{P}(\mathcal{X})$ through the change of variables.

Definition 2.14 [30, Definition 5.1]. *For $b^0,b^1\in\mathcal{B}$, define the set of admissible paths $P\left(b^0,b^1\right)$ as the collection of curves $(b_t)_t$ valued in $\mathcal{B}$, which belong to the Sobolev space $\mathscr{H}^1([0,1],\mathcal{H}_c)$ of square integrable, differentiable curves $[0,1]\to\mathcal{H}_c$ with square integrable derivative, and have end points $b_0=b^0, b_1=b^1$. The Riemannian distance d_S is then defined by*

$$\mathsf{d}_S\left(b^0,b^1\right) \coloneqq \left(\inf_{\substack{(b_t)_t\in P(b^0,b^1)\\ \mu_t\coloneqq B^{-1}(b_t)}}\int_0^1 \tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t,\dot{b}_t\right)dt\right)^{\frac{1}{2}}. \tag{2.25}$$

We conclude this section with the following result, guaranteeing existence of geodesics for this distance as well as an estimate with regards to the norm distance on $\mathcal{H}_c$, that we will use when proving the limit of the SJKO scheme Section 5 by using a geodesic interpolation between points of the sequence.

Theorem 2.15 [30, Theorems 5.2 and 5.3]. *The function d_S is a geodesic distance (i.e. there exists a minimizer for the right hand side of (2.25)), and there exists a constant C such that*

$$\sqrt{\tfrac{\varepsilon}{2}}\left\|b^0-b^1\right\|_{\mathcal{H}_c} \le \mathsf{d}_S\left(b^0,b^1\right) \le C\sqrt{\tfrac{\varepsilon}{2}}\left\|b^0-b^1\right\|_{\mathcal{H}_c}. \tag{2.26}$$

3 The differential equation and its structure

3.1 Derivation of the equation and change of variables

In this section, we start by informally deriving the limit of the SJKO scheme, the rigorous proof (in the case of a discrete space) being postponed to Section 5. We obtain a differential equation that we reformulate after the change of variables $\mu \to B(\mu)$, which allows us to define precisely what we consider a solution of the flow, whose existence is proven later in Section 4.

We start from the SJKO scheme

$$\mu_{k+1}^\tau \in \operatorname*{arg\,min}_{\mu \in \mathcal{P}(\mathcal{X})} \langle \mu, V \rangle + \tfrac{1}{2\tau} S_\varepsilon(\mu, \mu_k^\tau) \tag{3.1}$$

and derive the first order optimality conditions, which give

$$\exists p \in \mathcal{C}(\mathcal{X}), \begin{cases} \frac{1}{2\tau}\left(f_{\mu_{k+1}^\tau, \mu_k^\tau} - f_{\mu_{k+1}^\tau}\right) + V + p = 0 \\ p \leq 0 \\ \left\langle \mu_{k+1}^\tau, p \right\rangle = 0 \\ \mu_{k+1}^\tau \in \mathcal{P}(\mathcal{X}) \end{cases} \tag{3.2}$$

where the first equality is understood in $\mathcal{C}(\mathcal{X})/\mathbb{R}$. We can see a difference quotient of Schrödinger potentials appear, which is why the Hessian will be involved in the limit: due to the fact that for appropriate curves $(\mu_t)_t$, denoting $f_{t,s} \coloneqq f_{\mu_t, \mu_s}$, we have $-\frac{1}{2} \frac{\partial f_{t,s}}{\partial s}\Big|_{s=t} = G_{\mu_t}[\dot{\mu}_t]$ from Proposition 2.5, (3.2) will intuitively correspond at the limit to the existence of $p_t \in \mathcal{C}(\mathcal{X})$ such that

$$\begin{cases} G_{\mu_t}[\dot{\mu}_t] + V + p_t = 0 & (3.3) \\ p_t \leq 0 & (3.4) \\ \langle \mu_t, p_t \rangle = 0 & (3.5) \\ \mu_t \in \mathcal{P}(\mathcal{X}). & (3.6) \end{cases}$$

This can also be seen as the equation of the flow that would arise if we replaced S_ε in the SJKO scheme by its local expansion in terms of the metric tensor.

We will omit the dependency in t to ease the notations for now. We next consider the conditions (3.3–3.6) as our object of study, and rigorously rewrite them after the change of variables $b \coloneqq B(\mu) = H_c[a\mu]$ with $a \coloneqq a_\mu$ (we intentionally forget the dependency in μ in the notation for convenience). Firstly, one can compute $G_\mu[\dot{\mu}]$ thanks to the fact that $H_\mu[\dot{\mu}] = (\mathrm{Id} + K_\mu)\left[a\dot{b}\right]$ (by Lemma 2.10) and $G_\mu[\dot{\mu}] = \frac{\varepsilon}{2}(\mathrm{Id} - K_\mu)^{-1}(\mathrm{Id} + K_\mu)^{-1} H_\mu[\dot{\mu}]$, yielding

$$G_\mu[\dot{\mu}] = \tfrac{\varepsilon}{2}(\mathrm{Id} - K_\mu)^{-1}\left[a\dot{b}\right]. \tag{3.7}$$

To make the parallel with the Riemannian gradient flow on $\mathcal{H}_c$, we aim to make the metric tensor $\tilde{G}_\mu$ from Lemma 2.12 appear, so we compose (3.3) by the invertible operator $I + K_\mu$ [30, Theorem 3.8] and multiply by $b > 0$ to obtain

$$G_\mu[\dot{\mu}] + V + p = 0 \iff \tilde{G}_\mu[\dot{b}] + \tfrac{2}{\varepsilon} b\,(V + K_\mu[V] + (I + K_\mu)[p]) = 0, \tag{3.8}$$

where we can compute

$$\begin{aligned} \forall x \in \mathcal{X},\ bK_\mu[V](x) &= \int \exp\left(\tfrac{1}{\varepsilon}(f_\mu - c(x,\cdot))\right) V d\mu \\ &= H_c[Va\mu](x) \\ &= H_c\left[VH_c^{-1}[b]\right](x). \end{aligned} \tag{3.9}$$

Remark 3.1. *It must be highlighted that a priori, multiplying $b \in \mathcal{H}_c$ by $V \in \mathcal{C}(\mathcal{X})$ only yields an element of $\mathcal{C}(\mathcal{X})$ and not necessarily $\mathcal{H}_c$.*

In the following, we drop the square brackets to ease the notations (by seeing V as an operator $\mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ corresponding to pointwise multiplication). We also simplify the notation of $H_c V H_c^{-1}$ by analogy with the adjoint operator as motivated by the following lemma.

Lemma 3.2. *Let $V^* \coloneqq H_c V H_c^{-1} : H_c[\mathcal{M}(\mathcal{X})] \to \mathcal{H}_c$. It is a linear continuous operator, and for $g, h \in H_c[\mathcal{M}(\mathcal{X})]$, denoting abusively $\langle g, Vh\rangle_{\mathcal{H}_c} \coloneqq \langle H_c^{-1} g, Vh\rangle$ we have*

$$\langle g, Vh\rangle_{\mathcal{H}_c} = \langle V^* g, h\rangle_{\mathcal{H}_c}. \tag{3.10}$$

Proof. Observe that V^* is well defined, as multiplying a signed measure by a continuous function yields a signed measure again. The linearity of V^* is evident and its continuity follows from that of H_c^{-1} on $H_c[\mathcal{M}(\mathcal{X})]$ (see Lemma A.3), weak-$*$-to-weak-$*$ continuity of multiplying measures by V, and weak-$*$-to-norm continuity of H_c [30, Lemma B.2]. We obtain the 'adjointedness' by computing

$$\begin{aligned}\langle g, Vh\rangle_{\mathcal{H}_c} &= \langle H_c^{-1} g, Vh\rangle \\ &= \langle V H_c^{-1} g, h\rangle \\ &= \langle H_c V H_c^{-1} g, h\rangle_{\mathcal{H}_c}.\end{aligned}$$

□

In order to rewrite the term involving p in (3.8), notice that conditions (3.4–3.5) can be written as

$$\left\{\begin{array}{l} p \le 0 \\ \langle \mu, p\rangle = 0 \end{array}\right. \iff \left\{\begin{array}{l} p \le 0 \\ p \overset{\mu}{=} 0 \end{array}\right. \tag{3.11}$$

$$\iff \left\{\begin{array}{l} p \le 0 \\ \langle a\mu, p\rangle = 0 \end{array}\right. \tag{3.12}$$

$$\iff \left\{\begin{array}{l} p \le 0 \\ \langle H_c^{-1} b, p\rangle = 0 \end{array}\right. \tag{3.13}$$

where $\overset{\mu}{=}$ denotes equality μ-a.e. We give a name to the conditions (3.13) motivated by their correspondence with a Lagrange multiplier.

Definition 3.3. *Define the cone $\mathcal{K} \coloneqq H_c[\mathcal{M}_+(\mathcal{X})]$. For $b \in \mathcal{K}$, the set of **pressure vectors** at b is defined as*

$$\mathfrak{P}b \coloneqq \left\{ p \in \mathcal{C}(\mathcal{X}) \,\middle|\, p \le 0 \text{ and } \langle H_c^{-1} b, p\rangle = 0 \right\}. \tag{3.14}$$

From (3.11) it is evident that $\forall p \in \mathfrak{P}b, K_\mu[p] = 0$, and to further simplify (3.8) we state the following 'conic' property of the set $\mathfrak{P}b$.

Lemma 3.4. *For $b \in \mathcal{K}$ and any continuous function $g > 0$,*

$$p \in \mathfrak{P}b \iff gp \in \mathfrak{P}b. \tag{3.15}$$

Proof. The result is obtained by the same reasoning as (3.11–3.12), substituting μ by $H_c^{-1} b$ and a by g. □

We have therefore rephrased the flow as

$$\exists p \in \mathfrak{P}b,\ \tilde{G}_\mu \dot{b} + (V + V^*)b + p = 0. \tag{3.16}$$

Remark 3.5. *The potential energy, linear in the variable μ, is in fact quadratic in the variable b:*

$$\begin{aligned}\langle \mu, V\rangle &= \langle a\mu, Vb\rangle \\ &= \left\langle H_c^{-1}b, Vb\right\rangle.\end{aligned}$$

Writing this last term as $\langle b, Vb\rangle_{\mathcal{H}_c}$ with the same abuse as in Lemma 3.2, we can see intuitively that the gradient of this energy for $\langle\cdot,\cdot\rangle_{\mathcal{H}_c}$ would correspond to $Vb + V^*b$, meaning the flow (3.16) indeed corresponds to what we would expect to get when deriving the gradient flow of this energy in the space $\mathcal{H}_c$ endowed with $\tilde{G}$. In the following, we shall denote indiscriminately $E(\mu) = \langle \mu, V\rangle$ and $E(b) = \left\langle H_c^{-1}b, Vb\right\rangle$.

Remark 3.6. *We can see that not all components of (3.16) are in $\mathcal{H}_c$ since $Vb \in \mathcal{C}(\mathcal{X})$ highlighted in Remark 3.1, and p is also in $\mathcal{C}(\mathcal{X})$ generically. We shall see Section 4.1 that this subtlety disappears in the case where $\mathcal{X}$ is finite, allowing us to work entirely within $\mathcal{H}_c$ and utilize its Hilbert space structure, before generalizing the results to any compact space.*

We can now define rigorously the notion of solution to our equation.

Definition 3.7. *Let $V \in \mathcal{C}(\mathcal{X})$ and $b^0 \in \mathcal{B}$. We say that a curve $(b_t)_t$ valued in $\mathcal{B}$ is a **Sinkhorn potential flow** of V starting at b^0 if it belongs to the Sobolev space $\mathscr{H}^1([0,+\infty), \mathcal{H}_c)$ of square integrable, differentiable curves $[0,+\infty) \to \mathcal{H}_c$ with square integrable derivative, and verifies for Lebesgue almost every t*

$$\exists p_t \in \mathfrak{P}b_t,\ \tilde{G}_{\mu_t}\dot{b}_t + (V + V^*)b_t + p_t = 0 \tag{3.17}$$

as an equality between functions of $\mathcal{C}(\mathcal{X})$, where $\mu_t \coloneqq B^{-1}(b_t)$.

The existence and uniqueness being given later in Theorem 4.1, we now state the link of such flows with the flow in the variable μ shown above as the following proposition.

Proposition 3.8. *For a curve $(\mu_t)_t$ admissible in the sense that $(b_t)_t = (B(\mu_t))_t$ is in $\mathscr{H}^1([0,+\infty), \mathcal{H}_c)$, it verifies for Lebesgue-a.e. t the conditions (3.3–3.6) if and only if b is a Sinkhorn flow of the potential V.*

Thanks to this result, we can analyze the equation in the variable b knowing that it is a one to one correspondence with the limiting equation of the JKO flow. In the following section, we further rephrase equation (3.17) to exhibit more structure.

3.2 The structure of the flow

First, we highlight the structure of pressure vectors as outward normal to the cone $\mathcal{K}$, otherwise seen as the subgradient of the indicator of this cone which will be useful when considering the finite space case Section 4.1.

Proposition 3.9. *Denote $\iota_{\mathcal{K}}$ the convex indicator of $\mathcal{K}$ defined by*

$$\iota_{\mathcal{K}} : b \mapsto \begin{cases} 0 & \text{if } b \in \mathcal{K} \\ +\infty & \text{otherwise.}\end{cases}$$

Then, its subgradient $\partial\iota_{\mathcal{K}}$ (understood for the dot product $\langle\cdot,\cdot\rangle_{\mathcal{H}_c}$) with domain $\mathcal{K}$ verifies

$$\forall b \in \mathcal{K},\ \partial\iota_{\mathcal{K}}(b) = \mathcal{H}_c \cap \mathfrak{P}b. \tag{3.18}$$

Proof. For $b \in \mathcal{K}$ (i.e. $\iota_{\mathcal{K}}(b) = 0$), from the definition of subdifferential

$$\partial\iota_{\mathcal{K}}(b) = \left\{ p \in \mathcal{H}_c \,\middle|\, \forall \bar{b} \in \mathcal{H}_c, \left\langle p, \bar{b} - b\right\rangle_{\mathcal{H}_c} \leq \iota_{\mathcal{K}}(\bar{b}) \right\}$$

and the inequality is evidently verified for $\overline{b} \notin \mathcal{K}$, yielding

$$\begin{aligned}\partial \iota_{\mathcal{K}}(b) &= \left\{ p \in \mathcal{H}_c \,\middle|\, \forall \overline{b} \in \mathcal{K}, \langle p, \overline{b} - b \rangle_{\mathcal{H}_c} \leq 0 \right\} \\ &= \left\{ p \in \mathcal{H}_c \,\middle|\, \forall \overline{b} \in \mathcal{K}, \left\langle H_c^{-1}\overline{b}, p \right\rangle \leq \left\langle H_c^{-1} b, p \right\rangle \right\}.\end{aligned}$$

For p to verify this last condition, taking successively $\overline{b} = 2b$ and $\overline{b} = \frac{1}{2}b$ one must have $\left\langle H_c^{-1} b, p \right\rangle = 0$, and then taking $\overline{b} = H_c \delta_x$ with varying $x \in \mathcal{X}$ gives $p \leq 0$. Conversely, if $\left\langle H_c^{-1} b, p \right\rangle = 0$ and $p \leq 0$ it is easy to see that p verifies the condition. □

To better study (3.16), we would like to isolate $\dot{b}$ and thus compose by the inverse of the metric tensor. Denoting cl the closure, we can see that $\tilde{G}_\mu^{-1} : h \mapsto \frac{2}{\varepsilon} b(I - K_\mu)\left(I + K_\mu\right)^{-1}[ah]$ is in fact well defined as an operator on $\mathcal{C}(\mathcal{X})$, which is still invertible on $\mathrm{cl}_{\mathcal{C}(\mathcal{X})}\left(b^\perp\right) = \left\{ h \in \mathcal{C}(\mathcal{X}) \,\middle|\, \left\langle H_c^{-1} b, h \right\rangle = 0 \right\}$ thanks to [30, Theorem 3.8]. We can therefore state the following results.

Proposition 3.10. *For $b \in \mathcal{B}$, there holds*

$$\tilde{G}_\mu^{-1}(V + V^*)b = \tfrac{2}{\varepsilon}\left(V - V^*\right)b, \tag{3.19}$$

$$\mathfrak{P}b \subset \ker\left(\tilde{G}_\mu - \tfrac{\varepsilon}{2}\mathrm{Id}\right), \tag{3.20}$$

$$\tilde{G}_\mu \mathfrak{P} b = \mathfrak{P} b. \tag{3.21}$$

Proof. Recall that from (3.9) $V^* b = b K_\mu V$, and thus

$$\begin{aligned}\tilde{G}_\mu^{-1}(V + V^*)b &= \tfrac{2}{\varepsilon} b(I - K_\mu)\left(I + K_\mu\right)^{-1} ab\left(I + K_\mu\right)V \\ &= \tfrac{2}{\varepsilon} b(I - K_\mu)V \\ &= \tfrac{2}{\varepsilon}(V - V^*)b\end{aligned}$$

whence (3.19). Now let $p \in \partial \iota_{\mathcal{K}}(b)$, for any $g \in \mathcal{H}_c$ we have from Lemma 2.12

$$\begin{aligned}\left\langle \tilde{G}_\mu p, g \right\rangle_{\mathcal{H}_c} &= \tfrac{\varepsilon}{2}\left(\langle p, g \rangle_{\mathcal{H}_c} + 2\left\langle ap, \left(\mathrm{Id} - K_\mu\right)^{-1} ag \right\rangle_{L^2_\mu(\mathcal{X})} \right) \\ &= \tfrac{\varepsilon}{2}\langle p, g \rangle_{\mathcal{H}_c}\end{aligned}$$

since $p \overset{\mu}{=} 0$ from (3.11), whence $\tilde{G}_\mu p = \frac{\varepsilon}{2} p$. From Proposition 3.9, the density of $\mathcal{H}_c$ in $\mathcal{C}(\mathcal{X})$ and the continuity of the operator $\tilde{G}_\mu$ we obtain (3.20). Using the conic property given by Lemma 3.4 and the invertibility of $\tilde{G}_\mu$ on $\mathfrak{P}b \subset \mathrm{cl}_{\mathcal{C}(\mathcal{X})}\left(b^\perp\right)$ (from Definition 3.3) we deduce (3.21). □

We now turn briefly our attention to the operator appearing on the left hand side of (3.19) and highlight its properties before rewriting the flow.

Proposition 3.11. *The operator $\mathbf{W} \coloneqq \frac{2}{\varepsilon}\left(V - V^*\right) : H_c\left[\mathcal{M}(\mathcal{X})\right] \to \mathcal{C}(\mathcal{X})$ is continuous and verifies*

$$\forall g, h \in H_c\left[\mathcal{M}(\mathcal{X})\right], \left\langle H_c^{-1} g, \mathbf{W} h \right\rangle = -\left\langle H_c^{-1} h, \mathbf{W} g \right\rangle. \tag{3.22}$$

Proof. Direct from Lemma 3.2. □

We can now conclude this section by rewriting the Sinkhorn potential flow (3.17).

Proposition 3.12. *For a curve $(b_t)_t \in \mathscr{H}^1([0, +\infty), \mathcal{H}_c)$ valued in $\mathcal{B}$, the Sinkhorn potential flow (3.17) is equivalent the differential inclusion*

$$\dot{b}_t + \mathbf{W} b_t + \mathfrak{P} b_t \ni 0 \tag{3.23}$$

understood for a.e. t.

PROOF. Since for any b, $\mathfrak{P}b \subset \mathrm{cl}_{\mathcal{C}(\mathcal{X})}\left(b^{\perp}\right)$ and $\mathbf{W}b \in \mathrm{cl}_{\mathcal{C}(\mathcal{X})}\left(b^{\perp}\right)$ from Proposition 3.11, using the invertibility on $\tilde{G}_{\mu}$ on that set with Proposition 3.10 yields the equivalence. □

Remark 3.13. *Proposition 3.11 highlights a 'skew symmetry' property of $\mathbf{W}$, and thus the ODE without pressure constraints $\dot{b} + \mathbf{W}b$ will intuitively correspond to rotational motion (due to the Stone theorem [31, Theorem 1.10.8]). We can thus see the two forces at play in the differential inclusion* (3.23)*: this rotational motion induced by the gradient of the energy, and the pressure constraining the flow within the admissible set $\mathcal{K}$ corresponding to nonnegative measures. We will observe this behavior numerically in Section 6.*

In the remainder of this report, we will indiscriminately use the three expressions of the flow (3.3–3.6), (3.17) and (3.23) since Proposition 3.8 and Proposition 3.12 guarantee their equivalence.

3.3 A particular case: motion of a Dirac measure

An interesting case to understand the behavior of the flow is that of a Dirac mass i.e. a single particle. In the Wasserstein case and for a smooth potential, the flow corresponds to the continuity equation for a single particle i.e. the classical gradient flow of the potential V. We obtain a similar result in the case of a Sinkhorn flow.

Proposition 3.14. *Let $\mathcal{X}$ be a compact convex subset of $\mathbb{R}^d$, and $c(x,y) \coloneqq \|x-y\|_2^2$ be the square Euclidean distance on that space. Consider a particle following a smooth trajectory, i.e. an absolutely continuous curve $(x_t)_t \subset \mathcal{X}$, let $\mu_t \coloneqq \delta_{x_t}$ and $b_t \coloneqq B(\mu_t)$ the corresponding flow on $\mathcal{B}$. Then for any t holds*

$$\dot{b}_t + \mathbf{W}b_t + \mathfrak{P}b_t \ni 0 \iff \dot{x}_t \in -\partial V(x_t). \tag{3.24}$$

In particular, for V convex and $b^0 = B(\delta_{x_0})$, the Sinkhorn potential flow $(b_t)_t$ of V starting at b^0 can be written $b_t = B(\delta_{x_t})$ with x_t the unique subdifferential flow of V starting at x_0.

PROOF. Observe that for any $x \in \mathcal{X}$, f_{δ_x} verifies from the Schrödinger system (2.2)

$$\forall y, f_{\delta_x}(y) = \|x-y\|_2^2 - f_{\delta_x}(x)$$

and in particular for $y = x$ one has $f_{\delta_x}(x) = 0$, yielding

$$\forall y, f_{\delta_x}(y) = \|x-y\|_2^2$$

and thus we obtain

$$b_t(y) = \exp\left(-\tfrac{1}{\varepsilon}\|x_t - y\|_2^2\right).$$

Consequently, one can compute

$$\forall y, \dot{b}_t(y) = \tfrac{2}{\varepsilon}\langle \dot{x}_t, y - x_t\rangle_{\mathbb{R}^d} b_t(y).$$

Additionally, since $H_c^{-1}b_t$ is supported on $\{x_t\}$,

$$VH_c^{-1}b_t = V(x_t)H_c^{-1}b_t$$

and thus

$$\begin{aligned} V^*b_t &= H_cVH_c^{-1}b_t \\ &= V(x_t)b_t \end{aligned} \tag{3.25}$$

We can now compute $\dot{b}_t + \mathbf{W}b_t$ and check the necessary and sufficient condition for it to be an element of $-\mathfrak{P}b_t$. We obtain

$$\dot{b}_t(y) + \mathbf{W}b_t(y) = \tfrac{2}{\varepsilon}\left(V(y) - V(x_t) + \langle \dot{x}_t, y - x_t\rangle_{\mathbb{R}^d}\right) b_t(y).$$

The left side is evidently 0 for $y = x_t$, and it is nonnegative for all y if and only if $-\dot{x}_t$ is a subgradient of V, giving (3.24). The rest of the statement is an immediate consequence. □

Remark 3.15. *A simple corallary of Proposition 3.14 is that for V convex, the flow of a particle will converge to the minimum of V on $\mathcal{X}$ as it is well-known for classical subgradient flows. However, in the non-convex case, the subgradient flow of V may be undefined, yet we will see that the Sinkhorn potential flow still exists (Theorem 4.1), so the two notions will differ. This leaves the possibility that, unlike the classical subgradient flow or the Wasserstein gradient flow which get stuck in local minima, the Sinkhorn flow could still converge to a global minimum. This is addressed Section 4.2.*

4 Well posedness and properties

4.1 Existence, uniqueness, contractivity

This section is dedicated to the proof of the following theorem giving the well-posedness of the Sinkhorn flow and its contractivity.

Theorem 4.1. *For any $V \in \mathcal{C}(\mathcal{X})$ and $b^0 \in \mathcal{B}$, there exists a unique Sinkhorn potential flow $(b_t)_t$ of V starting at b^0 which additionally verifies*

(a) The norm $\left\|\dot{b}_t\right\|_{\mathcal{H}_c}$ decreases.

Moreover,

(b) The flow is contractive i.e. if $\left(b_t^1\right)_t$, $\left(b_t^2\right)_t$ are two Sinkhorn potential flows of V with possibly different starting points, then $t \mapsto \left\|b_t^1 - b_t^2\right\|_{\mathcal{H}_c}$ decreases.

To prove Theorem 4.1, we first consider the case where the space is finite i.e. $\mathcal{X} = \{x_1, \dots, x_n\}$ for some n, in order to be able to utilize the Hilbert space structure of $\mathcal{H}_c$ and apply well-studied results of the theory of differential inclusions in Hilbert spaces despite the difficulty mentioned in Remark 3.6. This greatly simplifies the situation as it means that $\mathcal{C}(\mathcal{X}) \simeq \mathbb{R}^n$, and since $\mathcal{H}_c$ is a dense linear subspace of the former we have $\mathcal{H}_c = \mathcal{C}(\mathcal{X})$ and the Hilbertian norm topology and supremum norm topology coincide (as do all norm topologies on finite dimensional spaces). The whole equation can therefore be understood in $\mathcal{H}_c$, and the notation in Lemma 3.2 is no longer abusive meaning that V^* becomes the true adjoint of V. Additionally, from Proposition 3.9 we have in this case $\mathfrak{P} = \partial \iota_{\mathcal{K}}$ which will prove useful. We will be able to obtain in this simplified case the following results, which we state for a flow on $\mathcal{K}$ more generally than $\mathcal{B}$ since $\mathbf{W}$ becomes well defined and continuous on the whole space (as it essentially becomes a matrix).

Theorem 4.2. *Assume that $\mathcal{X}$ is a finite set of points. Then for any $b^0 \in \mathcal{K}$, the differential inclusion problem*

$$\begin{cases} \dot{b}_t + \mathbf{W} b_t + \mathfrak{P} b_t \ni 0 \\ b_0 = b^0 \end{cases} \tag{4.1}$$

has a unique absolutely continuous solution on the time interval $[0, \infty)$. It additionally verifies the following:

(a) $\forall t, b_t \in \mathcal{K}$.

(b) $\forall t, \|b_t\|_{\mathcal{H}_c} = \left\|b^0\right\|_{\mathcal{H}_c}$.

(c) The curve b has a right-hand derivative verifying $\dot{b}_{t^+} = -\mathbf{W} b_t - p_t$ with $p_t = \underset{p \in \mathfrak{P} b_t}{\arg\min} \|\mathbf{W} b_t + p\|_{\mathcal{H}_c}$.

(d) The norm $\left\|\dot{b}_t\right\|_{\mathcal{H}_c}$ decreases.

To prove Theorem 4.2, we will use the Hille-Yosida theorem for multivalued operators, a result utilizing the notion of maximal monotonicity of an operator which we briefly recall.

Definition 4.3 [32, Section 1.1]. *Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space and $\mathbf{M} : \mathcal{H} \rightrightarrows \mathcal{H}$ a multivalued operator i.e. a set-valued map with nonempty domain $\operatorname{dom}(\mathbf{M}) \coloneqq \{x \in \mathcal{H}, \mathbf{M}x \neq \varnothing\}$. We denote $[x, y] \in \mathbf{M}$ to say $y \in \mathbf{M}x$. $\mathbf{M}$ is said to be **monotone** if*

$$\forall [x, y], [x', y'] \in \mathbf{M}, \left\langle x' - x, y' - y \right\rangle \geq 0.$$

A monotone operator is further called ***maximal*** *if no other monotone operator has a strictly greater graph (in the sense of inclusion).*

Lemma 4.4. *Under the assumption of Theorem 4.2,* $\mathbf{W} : \mathcal{H}_c \to \mathcal{H}_c$ *is maximal monotone.*

Proof. The monotonicity of $\mathbf{W}$ is a simple consequence of Proposition 3.11 which translates to actual skew-symmetry for $\langle \cdot, \cdot \rangle_{\mathcal{H}_c}$ in our case. Its maximality follows from its linearity and continuity, by applying [33, Proposition 2.7]. □

We now have the tools to prove Theorem 4.2.

Proof of Theorem 4.2. The maximal monotonicity of $\mathbf{W}$ is given by Lemma 4.4, and that of $\mathfrak{P} = \partial \iota_{\mathcal{K}}$ is well-known, see e.g. [34, Theorem 2.15]. We verify the condition of [35, Theorem 1 (b)] to ensure the sum is also maximal monotone: for any $b \in \mathcal{K}$ then of course $b \in \mathrm{cl}\left(\mathrm{dom}\left(\mathbf{W}\right)\right) = \mathcal{H}_c$ and $b \in \mathrm{cl}\left(\mathrm{dom}\left(\iota_{\mathcal{K}}\right)\right) = \mathcal{K}$. $\mathbf{W}$ is additionally bounded since linear continuous, so the referenced theorem applies giving that $\mathbf{W} + \mathfrak{P}$ is maximal monotone. The existence and uniqueness of a solution is given by [32, Theorem 10] (originally found in [36, Lemma 3.2]). Assertions (a), (c) and (d) correspond to [32, Theorem 10, 2. and 3.], and finally observe that for any $p \in \mathfrak{P}b$,

$$\langle \mathbf{W}b + p, b \rangle_{\mathcal{H}_c} = 0$$

from Proposition 3.11 (3.22) and the definition of $\mathfrak{P}$ (Definition 3.3), and therefore the solution verifies

$$\begin{aligned} \frac{d}{dt}\left(\tfrac{1}{2}\left\|b\right\|_{\mathcal{H}_c}^2\right) &= \left\langle \dot{b}, b \right\rangle_{\mathcal{H}_c} \\ &= 0 \end{aligned}$$

whence the fact (b). □

Next, we state a property regarding the dissipation of energy over the flow which is expected for gradient flows on finite-dimensional manifolds, and deduce properties that will be necessary when proving the generalization of Theorem 4.2.

Proposition 4.5. *Let* $(b_t)_t$ *follow the flow on a finite space given by Theorem 4.2, and denote the energy* E *following Remark 3.5. Then we have for almost every* t

$$\frac{d}{dt} E(b_t) = -\tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t, \dot{b}_t\right). \tag{4.2}$$

In particular:

(a) $t \mapsto E(b_t)$ *decreases*

(b) $\displaystyle\int_0^{+\infty} \left\|\dot{b}\right\|_{\mathcal{H}_c}^2 dt \leq \tfrac{2}{\varepsilon}\left(\sup_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu) - \inf_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu)\right)$.

Proof. The time derivative of the energy is given by

$$\frac{d}{dt}\left\langle b_t, V b_t \right\rangle_{\mathcal{H}_c} = \left\langle \dot{b}_t, (V + V^*) b_t \right\rangle_{\mathcal{H}_c}.$$

Then due to Theorem 4.2 (c) and the fact that b has an a.e. derivative which is equal to the right derivative where defined,

$$\left\langle \dot{b}_t, (V + V^*) b_t \right\rangle_{\mathcal{H}_c} = -\left\langle \mathbf{W} b_t, (V + V^*) b_t \right\rangle_{\mathcal{H}_c} - \left\langle (V + V^*) b_t, p_t \right\rangle_{\mathcal{H}_c} \tag{4.3}$$

where $p_t = \underset{p \in \mathfrak{P} b_t}{\arg\min} \left\|p + \mathbf{W} b_t\right\|_{\mathcal{H}_c}$. Dropping the dependency in t and denoting $g \coloneqq \mathbf{W} b$ for convenience, this writes $p = \mathrm{proj}\left(-g | \mathfrak{P} b\right)$ (where proj is the Hilbert projection on a closed convex set) which equivalently means

$$\forall \overline{p} \in \mathfrak{P} b, \left\langle g + p, p - \overline{p} \right\rangle_{\mathcal{H}_c} \leq 0$$

From Definition 3.3 it is clear that we can take $\overline{p} = 0$, so that $\langle g + p, p\rangle_{\mathcal{H}_c} \leq 0$. From Lemma 3.4 we can also take $\overline{p} = 2p$ to get $-\langle g + p, p\rangle_{\mathcal{H}_c} \leq 0$ and thus $\langle p, p\rangle_{\mathcal{H}_c} = -\langle g, p\rangle_{\mathcal{H}_c}$. Now since $\tilde{G}_\mu p = \frac{\varepsilon}{2} p$ from Proposition 3.10 (3.20) and $\tilde{G}_\mu$ is self-adjoint, using Proposition 3.10 (3.19) to write $(V + V^*)b = \tilde{G}_\mu g$ in (4.3) we get

$$\begin{aligned}
\frac{d}{dt}\langle b_t, V b_t\rangle_{\mathcal{H}_c} &= -\left(\left\langle g, \tilde{G}_\mu g\right\rangle_{\mathcal{H}_c} + \left\langle \tilde{G}_\mu g, p\right\rangle_{\mathcal{H}_c}\right) \\
&= -\left(\left\langle g, \tilde{G}_\mu g\right\rangle_{\mathcal{H}_c} + 2\left\langle \tilde{G}_\mu g, p\right\rangle_{\mathcal{H}_c} - \left\langle \tilde{G}_\mu g, p\right\rangle_{\mathcal{H}_c}\right) \\
&= -\left(\left\langle g, \tilde{G}_\mu g\right\rangle_{\mathcal{H}_c} + 2\left\langle \tilde{G}_\mu g, p\right\rangle_{\mathcal{H}_c} - \tfrac{\varepsilon}{2}\langle g, p\rangle_{\mathcal{H}_c}\right) \\
&= -\left(\left\langle g, \tilde{G}_\mu g\right\rangle_{\mathcal{H}_c} + 2\left\langle \tilde{G}_\mu g, p\right\rangle_{\mathcal{H}_c} + \tfrac{\varepsilon}{2}\langle p, p\rangle_{\mathcal{H}_c}\right) \\
&= -\left(\left\langle g, \tilde{G}_\mu g\right\rangle_{\mathcal{H}_c} + 2\left\langle \tilde{G}_\mu g, p\right\rangle_{\mathcal{H}_c} + \left\langle p, \tilde{G}_\mu p\right\rangle_{\mathcal{H}_c}\right) \\
&= -\left\langle g + p, \tilde{G}_\mu (g + p)\right\rangle_{\mathcal{H}_c} \\
&= -\tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t, \dot{b}_t\right).
\end{aligned}$$

The fact (a) follows from the fact that $\tilde{\mathbf{g}}_\mu$ is nonnegative and thus the derivative is nonpositive. For (b), first notice that the right hand terms are well defined since E is a continuous functional on the (weak-$*$) compact space $\mathcal{P}(\mathcal{X})$. $(E(b_t))_t$ is decreasing lower bounded and thus convergent to some E_∞, yielding

$$\begin{aligned}
\int_0^{+\infty} \tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t, \dot{b}_t\right) dt &= E(b^0) - E_\infty \\
&\leq \sup_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu) - \inf_{\mu \in \mathcal{P}(\mathcal{X})} E(\mu)
\end{aligned}$$

and the fact that $\frac{\varepsilon}{2}\left\|\dot{b}_t\right\|_{\mathcal{H}_c}^2 \leq \tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t, \dot{b}_t\right)$ from Proposition 2.13 gives the result. □

Now that we have established results in the case of a finite space, we aim to generalize them to any compact (and thus separable) space and prove Theorem 4.1. To obtain existence, we will discretize the space $\mathcal{X}$ and 'approximate' it with some $\mathcal{X}_n$ containing n points as mentioned at the beginning of this section, but doing so will yield solutions that are functions on $\mathcal{X}_n$, which we need to extend to $\mathcal{X}$ so that the sequence is contained in a fixed space where we can take the limit. This is done using the following lemma.

Lemma 4.6. *Let $\mathcal{X}_n = \{x_1, \ldots, x_n\} \subset \mathcal{X}$. For a PD kernel k on $\mathcal{X}$, denote $\mathcal{H}_k$ the corresponding RKHS and $\mathcal{H}_k^n$ the RKHS for the restriction k^n of k to $\mathcal{X}_n$. Then there is an isometric embedding $\mathcal{I}^n : \mathcal{H}_k^n \to \mathcal{H}_k$ verifying for any $h^n \in \mathcal{H}_k^n$*

$$\forall i \in \{1, \ldots, n\}, \mathcal{I}^n[h^n](x_i) = h^n(x_i). \tag{4.4}$$

Proof. Since we have by construction of the RKHS

$$\mathcal{H}_k^n = \operatorname{span}\left(k^n(x_1, \cdot), \ldots, k^n(x_n, \cdot)\right),$$

We define for $h^n = \sum_{i=1}^n h_i^n k^n(x_i, \cdot)$ its embedding

$$\mathcal{I}^n[h^n] \coloneqq \sum_{i=1}^n h_i^n k(x_i, \cdot). \tag{4.5}$$

It is evidently linear, and we have by the reproducing kernel property that

$$\begin{aligned}\|\mathcal{I}^n[h^n]\|^2_{\mathcal{H}_k} &= \sum_{i,j=1}^n h_i^n h_j^n k(x_i,x_j)\\ &= \sum_{i,j=1}^n h_i^n h_j^n k^n(x_i,x_j)\\ &= \|h^n\|^2_{\mathcal{H}_k^n}\end{aligned}$$

which makes $\mathcal{I}^n$ an isometry. The interpolation property (4.4) is clear from the definition of the embedding (4.5). □

We now state the compactness necessary to obtain a limit as $n \to +\infty$.

Lemma 4.7. *Let $\mathcal{X}$ be a compact metric space, take $\{x_i\}_{i=1}^\infty$ a countable set dense in $\mathcal{X}$, and denote for all $n \geq 1$ $\mathcal{X}_n \coloneqq \{x_i\}_{i=1}^n$. Let $T > 0$ be fixed, $(\underline{b}_t^n)_{t\in[0,T]}$ be the unique Sinkhorn potential flow on $[0,T]$ with ambient space $\mathcal{X}_n$ guaranteed by Theorem 4.2, and denote $b_t^n \coloneqq \mathcal{I}^n\left[\underline{b}_t^n\right]$ the corresponding extension to $\mathcal{X}$ as in Lemma 4.6. Then the sequence $(b^n)_n$ has a subsequence converging uniformly in $\mathcal{C}([0,T],\mathcal{H}_c)$ to some b and $\left(\dot{b}^n\right)_n$ has a subsequence converging weakly in $L^2([0,T],\mathcal{H}_c)$ to the derivative $\dot{b}$ of b, which additionally verifies*

$$\int_0^T \left\|\dot{b}_t\right\|^2_{\mathcal{H}_c} dt \leq \tfrac{2}{\varepsilon}\left(\sup_{\mathcal{P}(\mathcal{X})} E - \inf_{\mathcal{P}(\mathcal{X})} E\right). \tag{4.6}$$

Proof. Observe that from the linearity and continuity of $\mathcal{I}^n$, one has $\dot{b}_t^n = \mathcal{I}^n\left[\dot{\underline{b}}_t^n\right]$. From Proposition 4.5 (b), we obtain that

$$\begin{aligned}\int_0^T \left\|\dot{b}_t^n\right\|^2_{\mathcal{H}_c} dt &= \int_0^T \left\|\dot{\underline{b}}_t^n\right\|^2_{\mathcal{H}_c^n} dt\\ &\leq \tfrac{2}{\varepsilon}\left(\sup_{\mathcal{P}(\mathcal{X})} E - \inf_{\mathcal{P}(\mathcal{X})} E\right)\end{aligned}$$

and thus the sequence $(\dot{b}^n)_n$ is bounded in $L^2([0,T];\mathcal{H}_c)$ which is a Hilbert space with the usual inner product. As bounded sets in reflexive Banach spaces are weak-$*$ compact (from the Banach-Alaoglu theorem, see e.g. [37, Theorem 4.2]), we can extract a weak-$*$ converging subsequence with limit denoted $\dot{b}$. The statement (4.6) is consequence of the lower semi continuity of the norm for the weak-$*$ topology. Now considering that b^n is in the Sobolev space $\mathscr{H}^1([0,T];\mathcal{H}_c)$ due to Lemma 4.6 and Bochner's theorem [38, Theorem 2.5], we use Morrey's inequality (Lemma B.4) and the boundedness of $\left\|\dot{b}^n\right\|_{L^2([0,T];\mathcal{H}_c)}$, giving existence of a constant C such that

$$\forall n, \forall t,s, \|b_t^n - b_s^n\|_{\mathcal{H}_c} \leq C\,|t-s|^{\frac{1}{2}}.$$

Since for all t, $(b_t^n)_n \subset \mathcal{B}$ which is norm compact thanks to Theorem 2.8 (b), we have the required assumptions for the Arzelà-Ascoli theorem [39, Lemma 1] which yields a uniformly converging subsequence with limit b. The limit $\dot{b}$ of the derivatives has coherent notation as it is indeed the derivative of b: uniform convergence implies $L^2([0,T],\mathcal{H}_c)$ convergence (by the dominated convergence theorem), and thus implies distributional convergence. □

We now have the necessary tools to prove Theorem 4.1. The methodology to prove existence will be similar to [1, Proposition 2.3], but the discretization is in space rather than time.

PROOF OF THEOREM 4.1. We first prove the existence. Consider the converging subsequences given by Lemma 4.7 (which we identify with the sequences to simplify notation). Define

$$p_t^n \coloneqq -\left(\dot{b}_t^n + \mathbf{W} b_t^n\right)$$

and

$$p_t \coloneqq -\left(\dot{b}_t + \mathbf{W} b_t\right).$$

Showing $p_t \in \mathfrak{P} b_t$ for almost every t will yield the existence. From the construction of b^n it holds that

$$\forall i \in \{1, \dots, n\}, p_t^n(x_i) \leq 0$$

for almost every t, and thus for an arbitrary continuous nonnegative (scalar) curve $(\lambda_t)_t \in \mathcal{C}([0,T]; \mathbb{R})$ and $i \in \{1, \dots, n\}$,

$$\int_0^T \lambda_t p_t^n(x_i) dt \leq 0. \tag{4.7}$$

Take $x \in \mathcal{X}$. By density, there is a sequence of indices $(i_n)_n$ such that $x_{i_n} \to x$, and we can additionally take $i_n \leq n$ for all n. We abusively write n instead of i_n to simplify notations. Then holds

$$\begin{aligned}\int_0^T \lambda_t p_t^n(x_n) dt &= -\int_0^T \lambda_t (\mathbf{W} b_t^n)(x_n) dt - \int_0^T \lambda_t \dot{b}_t^n(x_n) dt \\ &= -\int_0^T \lambda_t \left\langle \delta_{x_n}, \mathbf{W} b_t^n \right\rangle dt - \int_0^T \lambda_t \left\langle \dot{b}_t^n, k_c(x_n, \cdot) \right\rangle_{\mathcal{H}_c} dt.\end{aligned}$$

The first term can be passed to the limit since easily δ_{x_n} converges weak-$*$ to δ_x and for all t $\mathbf{W} b_t^n \to \mathbf{W} b_t$ strongly in $\mathcal{C}(\mathcal{X})$, so the duality pairing converges [40, Proposition 3.13 (iv)], and the dominated convergence theorem gives convergence of the integral. Similarly, for the second term, the convergence of $\dot{b}^n$ is weak in $L^2([0,T], \mathcal{H}_c)$ and that of $(\lambda_t k_c(x_n, \cdot))_t$ is pointwise in time by Lemma A.2 and thus strong in the same L^2 space. This allows one to pass the duality pairing to the limit. We therefore get

$$\begin{aligned}\int_0^T \lambda_t p_t^n(x_n) dt &\to \int_0^T \lambda_t \left(-\mathbf{W} b_t(x) - \dot{b}_t(x)\right) \\ &= \int_0^T \lambda_t p_t(x)\end{aligned}$$

which combined with (4.7) yields

$$\int_0^T \lambda_t p_t(x) dt \leq 0 \tag{4.8}$$

and the arbitrariness of λ and x give that $p_t \leq 0$ for almost every t. We also have for a.e. t

$$\begin{aligned}\left\langle H_c^{-1} b_t, p_t \right\rangle &= -\left\langle H_c^{-1} b_t, \mathbf{W} b_t \right\rangle - \left\langle b_t, \dot{b}_t \right\rangle_{\mathcal{H}_c} \\ &= 0\end{aligned}$$

since Proposition 3.11 gives that the first term is null, and the fact that the norm of b_t remains constant makes the second term null as well. We thus have $p_t \in \mathfrak{P} b_t$ for a.e. t, concluding the proof of existence on $[0,T]$. The uniqueness which we next prove combined with the arbitrariness of T will allow one to extend the solution to $[0, +\infty)$, still in $\mathscr{H}^1([0,+\infty), \mathcal{H}_c)$ since the bound (4.6) in Lemma 4.7 is independent from T. We now prove the contractivity (b), which implies uniqueness when the initial point is fixed. Denoting p^1, p^2 pressure curves corresponding to the solutions b^1, b^2 respectively, the chain rule gives

$$\begin{aligned}\frac{d}{dt} \left\| b_t^1 - b_t^2 \right\|_{\mathcal{H}_c}^2 &= \left\langle \dot{b}_t^1 - \dot{b}_t^2, b_t^1 - b_t^2 \right\rangle_{\mathcal{H}_c} \\ &= -\left\langle H_c^{-1} \left[b_t^1 - b_t^2\right], \mathbf{W}(b_t^1 - b_t^2) + p_t^1 - p_t^2 \right\rangle\end{aligned}$$

and Proposition 3.11 gives that the term in $\mathbf{W}$ vanishes. We rewrite the remaining term using that $\left\langle H_c^{-1} b_t^i, p_t^i \right\rangle = 0$ ($i \in \{1,2\}$), yielding

$$\begin{aligned} -\left\langle H_c^{-1}\left[b_t^1 - b_t^2\right], p_t^1 - p_t^2 \right\rangle &= \left\langle H_c^{-1} b_t^1, p_2 \right\rangle + \left\langle H_c^{-1} b_t^2, p_1 \right\rangle \\ &\leq 0 \end{aligned}$$

and thus $\frac{d}{dt}\left\|b_t^1 - b_t^2\right\|_{\mathcal{H}_c}^2 \leq 0$ which gives the contractivity. Now if b follows a Sinkhorn flow, evidently $(b_{t+h})_t$ also does for any $h > 0$, and as a result for $s > t$

$$\|b_{s+h} - b_s\|_{\mathcal{H}_c} \leq \|b_{t+h} - b_t\|_{\mathcal{H}_c}$$

which after dividing by h and making $h \to 0$ gives $\left\|\dot{b}_s\right\|_{\mathcal{H}_c} \leq \left\|\dot{b}_t\right\|_{\mathcal{H}_c}$ and thus (a) holds. □

4.2 Long time behavior

We investigate in this section the behavior of the flow as $t \to +\infty$, and show that it converges to the minimizer of the energy under mild conditions as in the following theorem.

Theorem 4.8. *Assume V has a unique minimizer $x^\star$ on $\mathcal{X}$. Let $(b_t)_t$ be the Sinkhorn potential flow of V starting at some $b^0 \in \mathcal{B}$. Then,*

$$b_t \xrightarrow[t\to\infty]{} b_{\min} := B(\delta_{x^\star}) \tag{4.9}$$

in $\mathcal{B}$ and thus

$$\mu_t \xrightarrow[t\to\infty]{}^{*} \delta_{x^\star} \tag{4.10}$$

where $\rightharpoonup^*$ denotes weak-$*$ convergence.

The above theorem is the direct consequence of Lemma 4.11 and Lemma 4.12 below. We first give the behavior of the derivative of the flow as $t \to \infty$, easily deduced from Theorem 4.1.

Lemma 4.9. *Let $(b_t)_t$ be a Sinkhorn potential flow. Then its derivative $\dot{b}_t$ converges to 0 in $\mathcal{H}_c$ as $t \to \infty$.*

Proof. The statement is an immediate consequence of $\dot{b} \in L^2([0,+\infty), \mathcal{H}_c)$ and Theorem 4.1 (a). □

Next, we state the following closedness of $\mathfrak{P}$ needed to take limits.

Lemma 4.10. *The graph of $\mathfrak{P}$ is closed in $\mathcal{B} \times \mathcal{C}(\mathcal{X})$.*

Proof. Consider a sequence $b_n \to b$ in $\mathcal{B}$ and $p_n \to p$ in $\mathcal{C}(\mathcal{X})$ where $p_n \in \mathfrak{P} b_n$ for all n. It is evident that $p \leq 0$ by uniform (and thus pointwise) limit, and by Lemma A.3 the term $H_c^{-1} b_n$ converges weakly-$*$ to $H_c^{-1} b$ in $\mathcal{M}(\mathcal{X})$, which paired with a strongly convergent sequence yields $0 = \left\langle H_c^{-1} b_n, p_n \right\rangle \to \left\langle H_c^{-1} b, p \right\rangle$ by [40, Proposition 3.13 (iv)], whence $p \in \mathfrak{P} b$. □

We can now show the two lemmas proving Theorem 4.8.

Lemma 4.11. *Let $(b_t)_t$ follow the Sinkhorn potential flow of V starting at $b^0 \in \mathcal{B}$. Then this curve has at least one accumulation point and all of its accumulation points $\bar{b}$ are critical in the sense that*

$$0 \in \mathbf{W}\bar{b} + \mathfrak{P}\bar{b}. \tag{4.11}$$

Proof. $\mathcal{B}$ is (norm) compact thanks to Theorem 2.8 (b) which gives existence of a converging subsequence. Taking such a sequence denoted as $b_{t_n} \xrightarrow[n\to\infty]{} \bar{b}$ for $(t_n)_n$ increasing and going to $+\infty$, Lemma 4.9 ensures that

$$\mathbf{W} b_{t_n} + p_{t_n} \xrightarrow[t\to\infty]{} 0 \quad \text{in } \mathcal{H}_c \tag{4.12}$$

and since $\mathbf{W}$ is continuous from $\mathcal{B}$ to $\mathcal{C}(\mathcal{X})$ (Proposition 3.11), we have $\mathbf{W}b_{t_n} \xrightarrow[n\to\infty]{} \mathbf{W}\overline{b}$ in $\mathcal{C}(\mathcal{X})$ and thus p_{t_n} converges to $-\mathbf{W}\overline{b}$ in $\mathcal{C}(\mathcal{X})$. The graph of $\mathfrak{P}$ is closed by Lemma 4.10, meaning $-\mathbf{W}\overline{b} \in \mathfrak{P}\overline{b}$ and therefore $\overline{b}$ is a critical point. □

Lemma 4.12. *Assume V has a unique minimizer $x^\star$. Then the only critical point in the sense of* (4.11) *is* $b_{\min} \coloneqq B(\delta_{x^\star})$.

Proof. Thanks to the computation (3.25) we have $-\mathbf{W}b_{\min} = \frac{2}{\varepsilon}\left(V(x^\star) - V\right) b_{\min}$, which integrates to 0 against $\delta_{x^\star}$ and is nonpositive, i.e. $-\mathbf{W}b_{\min} \in \mathfrak{P}b_{\min}$ and $b_{\min}$ is critical. We now proceed by contraposition and take $\overline{b} \neq b_{\min}$ to show it is not a critical point. Write $\overline{\mu} = B^{-1}\left(\overline{b}\right)$, and consider the curve

$$\mu_t \coloneqq (1-t)\overline{\mu} + t\delta_{x^\star} \tag{4.13}$$

and its associated curve $(b_t)_t$. The former is a vertical perturbation (of derivative $\dot{\mu} = \delta_{x^\star} - \overline{\mu} \in \mathcal{M}(\mathcal{X})$) and thus admissible in the sense that $\dot{b}$ is well defined (as consequence of [30, Proposition 3.11 and Lemma B.2.]). We first compute the derivative of the energy over time in the variable μ_t: using Remark 3.5 and differentiating we have

$$\begin{aligned} \frac{d}{dt}E(b_t) &= \langle \dot{\mu}, V\rangle \\ &= V(x^\star) - \langle \bar{\mu}, V\rangle \\ &< 0 \end{aligned} \tag{4.14}$$

since $\overline{b} \neq b_{\min}$ and thus $\overline{\mu} \neq \delta_{x^\star}$. We can also express this derivative in the variable b as follows. Using the fact that $H_{\mu_t}\dot{\mu} = (I + K_{\mu_t})\left[a_t\dot{b}_t\right]$ from Lemma 2.10 and $H_{\mu_t}\dot{\mu} = a_t H_c\left[a_t\dot{\mu}\right]$ we get

$$H_c\left[a_t\dot{\mu}\right] = \dot{b}_t + H_c\left[a_t^2\dot{b}_t\mu_t\right] \tag{4.15}$$

and thus

$$\begin{aligned} \dot{\mu} &= b_t H_c^{-1}\dot{b}_t + a_t\dot{b}_t\mu_t \\ &= b_t H_c^{-1}\dot{b}_t + \dot{b}_t H_c^{-1} b_t. \end{aligned}$$

This yields

$$\begin{aligned} \frac{d}{dt}E(b_t) &= \left\langle b_t H_c^{-1}\dot{b}_t + \dot{b}_t H_c^{-1} b_t, V\right\rangle \\ &= \left\langle H_c^{-1}\dot{b}_t, Vb_t\right\rangle + \left\langle H_c^{-1} b_t, V\dot{b}_t\right\rangle \\ &= \left\langle H_c^{-1}\dot{b}_t, (V+V^*)b_t\right\rangle \end{aligned}$$

from Lemma 3.2, and combined with (4.14) we get

$$\left\langle H_c^{-1}\dot{b}_t, (V+V^*)b_t\right\rangle < 0. \tag{4.16}$$

Now observe that for any t and any $p_t \in \mathfrak{P}b_t$, $\left\langle H_c^{-1}b_t, p_t\right\rangle = 0$ and for any s, $\left\langle H_c^{-1}b_{t+s}, p_t\right\rangle \leq 0$. As a result,

$$\left\langle H_c^{-1}\left[\frac{b_{t+s} - b_t}{s}\right], p_t\right\rangle \leq 0.$$

We will now use the continuity of H_c^{-1} from Lemma A.3 to take the limit $s \to 0$, which requires to check that everything stays in $H_c[\mathcal{M}(\mathcal{X})]$. Clearly for any s, $\frac{b_{t+s}-b_t}{s} \in H_c[\mathcal{M}(\mathcal{X})]$, and $\frac{b_{t+s}-b_t}{s}$ converges in $\mathcal{H}_c$ to $\dot{b}_t$ as $s \to 0$, where from (4.15)

$$H_c^{-1}\dot{b}_t = a_t\left(\dot{\mu} - a_t\dot{b}_t\mu\right)$$

which is an element of $\mathcal{M}(\mathcal{X})$ due to $\dot{\mu} \in \mathcal{M}(\mathcal{X})$. We can thus pass to the limit and obtain $\left\langle H_c^{-1}\dot{b}_t, p_t \right\rangle \leq 0$. Using (4.16) we get

$$\left\langle H_c^{-1}\dot{b}_t, (V+V^*)b_t + p_t \right\rangle \leq \left\langle H_c^{-1}\dot{b}_t, (V+V^*)b_t \right\rangle < 0, \tag{4.17}$$

and since this holds for any $p_t \in \mathfrak{P}b_t$, it follows that $0 \notin (V+V^*)b_t + \mathfrak{P}b_t$. Taking $t=0$ then using Proposition 3.10 to compose by the inverse of the metric tensor $\tilde{G}_\mu$ gives the result. □

Remark 4.13. *We see from the proof that the main reason why we can obtain convergence to a global minimizer even for non-convex potentials in the Sinkhorn case but not in the Wasserstein case is that vertical perturbations are admissible for the former but not for the latter (all absolutely continuous curves for the Wasserstein distance are 'horizontal perturbations' verifying a continuity equation [7, Theorem 5.14]). Vertical perturbations intuitively allow for a sort of 'tunneling' effect making mass able to pass through potential barriers.*

5 Proof of the convergence of the SJKO scheme in the case of a finite space

We prove in this section the following theorem showing the validity of the limit as $\tau \to 0$ that we have derived informally Section 3.1, when the space $\mathcal{X}$ is made of n points.

Theorem 5.1. *Let $\mathcal{X} = \{x_1, \dots, x_n\}$ be finite, and for $\tau > 0$, define $(\mu_k^\tau)_k$ the sequence given by the SJKO scheme (3.1) with initialization $\mu_0 \in \mathcal{P}(\mathcal{X})$. Let $b_k^\tau \coloneqq B(\mu_k^\tau)$, $\left(\overline{b}_t^\tau\right)_t$ its piecewise constant interpolation i.e.*

$$\overline{b}_t^\tau = b_k^\tau \text{ for } t \in [k\tau, (k+1)\tau),$$

and $(b_t^\tau)_t$ its piecewise geodesic interpolation i.e. $(b_t^\tau)_{t\in[k\tau,(k+1)\tau]}$ is a constant-speed geodesic between b_k^τ and b_{k+1}^τ for the Riemannian metric d_S (as given by Theorem 2.15). Then on $[0,T]$ for arbitrary $T > 0$, up to a subsequence, $\overline{b}^\tau$ and b^τ both converge uniformly as $\tau \to 0$ to the Sinkhorn potential flow of V starting at $B(\mu_0)$.

The reason why the general case is much more involved will be made apparent Remark 5.8.

In order to prove Theorem 5.1, we will at each step write the difference $f_{\mu_{k+1}^\tau,\mu_k^\tau} - f_{\mu_{k+1}^\tau}$ appearing in (3.2) as the integral of its derivative along a curve interpolating between μ_k^τ and μ_{k+1}^τ, which involves the derivative of the Schrödinger potentials outside of the diagonal (in contrary to Proposition 2.5) and thus makes a generalization of the operators K_μ and H_μ appear. We first define these operators and adapt some results of [30, Section 3].

5.1 The operators $H_{\mu,\nu}$ and $K_{\mu,\nu}$

The following definitions and results hold in the general case where $\mathcal{X}$ need not be finite.

Definition 5.2. *Define the kernel $k_{\mu,\nu}$ by*

$$\forall x, y \in \mathcal{X},\ k_{\mu,\nu}(x,y) \coloneqq \exp\left(\tfrac{1}{\varepsilon}\left(f_{\mu,\nu}(x) + f_{\nu,\mu}(y) - c(x,y)\right)\right).$$

We define the operators $H_{\mu,\nu}$ and $K_{\mu,\nu}$ by

$$\begin{aligned} &H_{\mu,\nu} : \mathcal{M}(\mathcal{X}) \to \mathcal{C}(\mathcal{X}), \quad H_{\mu,\nu}[\sigma](x) \coloneqq \langle \sigma, k_{\mu,\nu}(x,\cdot)\rangle \\ &K_{\mu,\nu} : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X}), \quad K_{\mu,\nu}\phi \coloneqq H_{\mu,\nu}[\phi\nu]. \end{aligned}$$

The notation is chosen with this convention because $H_{\mu,\nu}$ is actually valued in the space $\mathcal{H}_{\mu,\nu} \coloneqq \exp\left(\frac{f_{\mu,\nu}}{\varepsilon}\right)\mathcal{H}_c$. As when $\mu = \nu$, the Schrödinger system (2.2) gives $K_{\mu,\nu}1 = 1$. We now state some results extending those of [30, Section 3].

Proposition 5.3. *$H_{\mu,\nu} : \mathcal{M}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ and $K_{\mu,\nu} : \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ are compact.*

Proof. [30, Proposition 3.6] proves the result when $\mu = \nu$, and the proof still works when $\mu \neq \nu$. □

Proposition 5.4. *Let $q \coloneqq 1 - \exp\left(-\frac{4\|c\|_\infty}{\varepsilon}\right) \in (0,1)$. Then:*

(a) For any $\phi \in \mathcal{C}(\mathcal{X})$, $\|K_{\mu,\nu}\phi\|_{\mathcal{C}(\mathcal{X})/\mathbb{R}} \leq q\,\|\phi\|_{\mathcal{C}(\mathcal{X})/\mathbb{R}}$.

(b) The operator $I - K_{\mu,\nu}K_{\nu,\mu}$ is invertible on $\mathcal{C}(\mathcal{X})/\mathbb{R}$ and $I + K_{\mu,\nu}$ is invertible on $\mathcal{C}(\mathcal{X})$.

(c) For $\phi, \psi \in \mathcal{C}(\mathcal{X})$, $\langle \psi, K_{\mu,\nu}\phi\rangle_{L^2_\mu(\mathcal{X})} = \langle K_{\nu,\mu}\psi, \phi\rangle_{L^2_\nu(\mathcal{X})}$.

PROOF. Items (a) and (b) are proven the same way as [30, Proposition 3.7, Theorem 3.8], and (c) is an easy consequence of the fact that $k_{\mu,\nu}(x,y) = k_{\nu,\mu}(y,x)$. □

Corollary 5.5. *The eigenvalues of $K_{\mu,\nu}K_{\nu,\mu}$ on $\mathcal{C}(\mathcal{X})$ belong to $[0,q^2]$, and that of $(I - K_{\mu,\nu}K_{\nu,\mu})^{-1}$ on $\mathcal{C}(\mathcal{X})/\mathbb{R}$ are contained in $\left[1, \frac{1}{1-q^2}\right]$.*

PROOF. The positivity of the eigenvalues of $K_{\mu,\nu}K_{\nu,\mu}$ is a consequence of Proposition 5.4 (c), and the bound of q^2 comes from Proposition 5.4 (a). The rest of the statement is easily deduced from the fact that $\lambda \neq 1$ is an eigenvalue of $K_{\mu,\nu}K_{\nu,\mu}$ if and only if $\frac{1}{1-\lambda}$ is an eigenvalue of $(I - K_{\mu,\nu}K_{\nu,\mu})^{-1}$. □

With these results, we can study the derivative of the Schrödinger potentials and generalize Proposition 2.5.

5.2 The derivative of a Schrödinger potential

Proposition 5.6. *For $(\mu_t)_t$ a curve valued in $\mathcal{P}(\mathcal{X})$ and weakly differentiable in $\mathcal{M}(\mathcal{X})$, denote $f_{t,s} \coloneqq f_{\mu_t,\mu_s}$, $K_{t,s} \coloneqq K_{\mu_t,\mu_s}$ and $H_{t,s} \coloneqq H_{\mu_t,\mu_s}$. It holds*

$$\frac{\partial f_{t,s}}{\partial s} = -\varepsilon \left(\mathrm{Id} - K_{t,s}K_{s,t}\right)^{-1} H_{t,s}\dot{\mu}_s. \tag{5.1}$$

PROOF. We will follow the same method as the proof of [30, Lemma 3.10]. The existence of derivatives is obtained by applying the implicit function theorem [41, Theorem 10.2.1] to the function

$$\tilde{T} : f, s \mapsto f - T_\varepsilon(T_\varepsilon(f, \mu_t), \mu_s) \tag{5.2}$$

where t is fixed. The optimal conditions given by the Schrödinger system (2.2) characterize $f_{t,s}$ as the unique solution (in $\mathcal{C}(\mathcal{X})/\mathbb{R}$) of $\tilde{T}(f,s) = 0$, and the mapping $\tilde{T}$ is continuously differentiable since it is the case of T_ε, which has partial derivatives computed as

$$D_1 T_\varepsilon(f,\mu)[g](x) = -\frac{\left\langle \mu, g \exp\left(\frac{1}{\varepsilon}f\right) k_c(x,\cdot)\right\rangle}{\left\langle \mu, \exp\left(\frac{1}{\varepsilon}f\right) k_c(x,\cdot)\right\rangle} \tag{5.3}$$

$$D_2 T_\varepsilon(f,\mu)[\dot{\mu}](x) = -\varepsilon\frac{\left\langle \dot{\mu}, \exp\left(\frac{1}{\varepsilon}f\right) k_c(x,\cdot)\right\rangle}{\left\langle \mu, \exp\left(\frac{1}{\varepsilon}f\right) k_c(x,\cdot)\right\rangle}. \tag{5.4}$$

We can thus compute using the chain rule and $f_{s,t} = T_\varepsilon(f_{t,s}, \mu_t)$ the partial derivative

$$D_1\tilde{T}(f_{t,s}, s) = \mathrm{Id} - D_1 T_\varepsilon(f_{s,t}, \mu_s) D_1 T_\varepsilon(f_{t,s}, \mu_t)$$

which is required to be invertible to apply the implicit function theorem. Using that $K_{\mu,\nu}1 = 1$ in (5.3) yields

$$\begin{aligned} D_1 T_\varepsilon(f_{s,t}, \mu_s)[g](x) &= -\frac{\left\langle \mu_s, g \exp\left(\frac{1}{\varepsilon}\left(f_{t,s}(x) + f_{s,t}\right)\right) k_c(x,\cdot)\right\rangle}{\left\langle \mu_s, \exp\left(\frac{1}{\varepsilon}\left(f_{t,s}(x) + f_{s,t}\right)\right) k_c(x,\cdot)\right\rangle} \\ &= -(K_{t,s}g)(x) \end{aligned}$$

and similarly

$$D_1 T_\varepsilon(f_{t,s}, \mu_t)[g] = -K_{s,t}g. \tag{5.5}$$

Thus we get

$$D_1\tilde{T}(f_{t,s}, s) = \mathrm{Id} - K_{t,s}K_{s,t} \tag{5.6}$$

which is invertible on $\mathcal{C}(\mathcal{X})/\mathbb{R}$ by Proposition 5.4. The implicit function theorem applies, yielding the existence of the partial derivative $\frac{\partial f_{t,s}}{\partial s}$ and its expression given by

$$\frac{\partial f_{t,s}}{\partial s} = -\left(D_1\tilde{T}(f_{t,s}, s)\right)^{-1}\left(\frac{\partial \tilde{T}(f_{t,s}, s)}{\partial s}\right) \tag{5.7}$$

with

$$\frac{\partial \tilde{T}(f_{t,s}, s)}{\partial s} = -D_2 T_\varepsilon(f_{s,t}, \mu_s)\left[\dot{\mu}_s\right] \tag{5.8}$$

and the right hand term is computed using that $K_{\mu,\nu}1 = 1$ in (5.4) as before, yielding

$$\frac{\partial \tilde{T}(f_{t,s}, s)}{\partial s} = \varepsilon H_{t,s}\left[\dot{\mu}_s\right]. \tag{5.9}$$

Substituting (5.6) and (5.9) into (5.7) yields the desired expression (5.1). □

Notice that Proposition 5.6 only works when $\dot{\mu} \in \mathcal{M}(\mathcal{X})$, which is part of the reason we must next consider a finite space. We will further explain this difficulty in Remark 5.8.

5.3 Limit $\tau \to 0$

We can now move to the proof of Theorem 5.1, where we identify $\mathcal{M}(\mathcal{X}) = \mathcal{C}(\mathcal{X}) = \mathcal{H}_c = \mathbb{R}^n$, operators on these spaces are identified with $n \times n$ matrices, and we will omit the topology considered when dealing with limits in those spaces since all weak and norm topologies become equivalent [40, Proposition 3.6].

We quickly recall the following estimate from [42] necessary to relate the Sinkhorn divergence and the norm distance after the change of variables.

Lemma 5.7. *Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $b_\mu \coloneqq B(\mu), b_\nu \coloneqq B(\nu)$. Then it holds*

$$\|b_\mu - b_\nu\|_{\mathcal{H}_c}^2 \leq \tfrac{2}{\varepsilon} S_\varepsilon(\mu, \nu). \tag{5.10}$$

Proof. We have $\left\| e^{\frac{f_\mu}{\varepsilon}}\mu - e^{\frac{f_\nu}{\varepsilon}}\nu \right\|_{\mathcal{H}_c^*}^2 \leq \frac{2}{\varepsilon} S_\varepsilon(\mu, \nu)$ from [42, Proposition 16], and using the isometry H_c gives the result. □

Proof of Theorem 5.1. Using the suboptimality of μ_k^τ in (3.1) gives

$$E(\mu_{k+1}^\tau) + \frac{S_\varepsilon(\mu_k^\tau, \mu_{k+1}^\tau)}{2\tau} \leq E(\mu_k^\tau)$$

and for any ℓ, we sum over $k \leq \ell$ to get from the boundedness of E

$$\sum_{k=0}^{\ell} \frac{S_\varepsilon(\mu_k^\tau, \mu_{k+1}^\tau)}{2\tau} \leq E(\mu_0^\tau) - E(\mu_{\ell+1}^\tau) \leq C$$

where $C > 0$ is a constant. We will abusively denote C various constants and not bother with their exact expressions for simplicity. Since $\mathsf{d}_S(b_{k+1}^\tau, b_k^\tau) \leq C \left\| b_{k+1}^\tau - b_k^\tau \right\|_{\mathcal{H}_c}$ from Theorem 2.15 and $\left\| b_{k+1}^\tau - b_k^\tau \right\|_{\mathcal{H}_c}^2 \leq \frac{2}{\varepsilon} S_\varepsilon(\mu_k^\tau, \mu_{k+1}^\tau)$ by Lemma 5.7, we get

$$\sum_{k=0}^{\ell} \frac{\mathsf{d}_S(b_{k+1}^\tau, b_k^\tau)^2}{2\tau} \leq C. \tag{5.11}$$

Additionally, for $t \in [k\tau, (k+1)\tau]$ and with $\mu_t^\tau \coloneqq B^{-1}(b_t^\tau)$,

$$\begin{aligned} \frac{\mathsf{d}_S(b_{k+1}^\tau, b_k^\tau)^2}{2\tau} &= \tfrac{\tau}{2}\left(\frac{\mathsf{d}_S(b_{k+1}^\tau, b_k^\tau)}{\tau}\right)^2 \\ &= \tfrac{\tau}{2}\tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t^\tau, \dot{b}_t^\tau\right) \end{aligned}$$

since $(b_t^\tau)_{t\in[k\tau,(k+1)\tau]}$ is a constant speed geodesic. Thus

$$\frac{\mathsf{d}_S(b_{k+1}^\tau, b_k^\tau)^2}{2\tau} = \tfrac{1}{2}\int_{k\tau}^{(k+1)\tau} \tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t^\tau, \dot{b}_t^\tau\right) dt$$

which combined with (5.11) gives

$$\int_0^T \tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t^\tau, \dot{b}_t^\tau\right) dt \leq C.$$

From the bound of Proposition 2.13, we deduce

$$\begin{aligned}\int_0^T \left\|\dot{b}_t^\tau\right\|_{\mathcal{H}_c}^2 dt &\leq \tfrac{2}{\varepsilon}\int_0^T \tilde{\mathbf{g}}_{\mu_t}\left(\dot{b}_t^\tau, \dot{b}_t^\tau\right) dt \\ &\leq C.\end{aligned}$$

Thus $(b^\tau)_\tau$ is a bounded sequence in $\mathscr{H}^1([0,T],\mathcal{H}_c)$ and $\left(\dot{b}^\tau\right)_\tau$ is bounded in $L^2([0,T],\mathcal{H}_c)$. Morrey's inequality (Lemma B.4) also gives

$$\left\|b_t^\tau - b_s^\tau\right\|_{\mathcal{H}_c} \leq C\left|t-s\right|^{\frac{1}{2}}$$

and since $\overline{b}^\tau$ and b^τ agree at $k\tau$ for any k,

$$\left\|b_t^\tau - \overline{b}_t^\tau\right\|_{\mathcal{H}_c} \leq C\tau^{\frac{1}{2}}.$$

The Ascoli-Arzelà theorem applies, giving a uniformly converging subsequence of $(b^\tau)_\tau$ to a limit b valued in $\mathcal{B}$, and $\overline{b}^\tau$ converges to the same limit due to the previous estimate. Banach-Alaoglu gives a subsequence of $\dot{b}^\tau$ converging weakly in $L^2([0,T],\mathcal{H}_c)$, to a limit curve $\dot{b}$ which is the derivative of b by distributional convergence. We now aim to show that the limit follows the expected equation. Recall the optimality conditions

$$\tfrac{1}{2\tau}\left(f_{\mu_{k+1}^\tau,\mu_k^\tau} - f_{\mu_{k+1}^\tau}\right) + V + p_{k+1}^\tau = 0 \tag{5.12}$$

for some $p_{k+1}^\tau \in \mathfrak{P}b_{k+1}^\tau$. Writing $f_{t,s}^\tau \coloneqq f_{\mu_t^\tau,\mu_s^\tau}$, we write the difference of Schrödinger potentials as the integral of its derivative i.e.

$$\tfrac{1}{2\tau}\left(f_{(k+1)\tau,k\tau}^\tau - f_{(k+1)\tau,(k+1)\tau}^\tau\right) = -\tfrac{1}{2\tau}\int_{k\tau}^{(k+1)\tau}\frac{\partial f_{(k+1)\tau,s}^\tau}{\partial s}ds.$$

Since there is no ambiguity between the functional spaces at hand, Proposition 5.6 applies which combined with $H_s\dot{\mu}_s^\tau = (\mathrm{Id}+K_s)\left[a_s^\tau \dot{b}_s^\tau\right]$ (Lemma 2.10) yields

$$-\tfrac{1}{2}\frac{\partial f_{(k+1)\tau,s}^\tau}{\partial s} = J_{(k+1)\tau,s}\dot{b}_s^\tau \tag{5.13}$$

where

$$J_{t,s}^\tau \coloneqq \tfrac{\varepsilon}{2}\left(\mathrm{Id} - K_{t,s}K_{s,t}\right)^\dagger H_{t,s}H_s^{-1}(I+K_s)\mathrm{diag}\left(a_s^\tau\right) \tag{5.14}$$

with $a_s^\tau \coloneqq (b_s^\tau)^{-1}$, $\mathrm{diag}\left(a_s^\tau\right)$ the corresponding diagonal matrix, and $M^\dagger$ denoting the Moore-Penrose pseudoinverse of a matrix $M \in \mathbb{R}^{n\times n}$ (which appears due to the fact that $\mathrm{Id} - K_{t,s}K_{s,t}$ is only invertible in $\mathcal{C}(\mathcal{X})/\mathbb{R}$). Every matrix involved in the above expression is continuous with respect to t,s since μ^τ is continuous and the potentials vary continuously with respect to the measures [21, Proposition 13 (Appendix B)]. With the continuity of the inverse, and that of the pseudoinverse for bounded sequences [43, Proposition 2.3] (the bound coming from Corollary 5.5), we have that the curve defined by

$$\widehat{J}_t^\tau \coloneqq J_{\left(\lfloor\frac{t}{\tau}\rfloor+1\right)\tau,t}^\tau$$

converges pointwise as $\tau \to 0$ to J_t, where

$$J_t \coloneqq \tfrac{\varepsilon}{2}\left(\mathrm{Id} - K_t^2\right)^\dagger (\mathrm{Id} + K_t)\mathrm{diag}\,(a_t)$$

with $a_t \coloneqq (b_t)^{-1}$. Notice that as seen in (3.7),

$$J_t \dot{b}_t = G_{\mu_t}\dot{\mu}_t \tag{5.15}$$

where $(\mu_t)_t \coloneqq (B^{-1}(b_t))_t$. Denoting the piecewise interpolation

$$\begin{aligned} g_t^\tau &\coloneqq \tfrac{1}{2\tau}\left(f^\tau_{(k+1)\tau,k\tau} - f^\tau_{(k+1)\tau,(k+1)\tau}\right) \quad \text{for } t \in [k\tau,(k+1)\tau) \\ &= \tfrac{1}{\tau}\int_{k\tau}^{(k+1)\tau} \widehat{J}_s^\tau \dot{b}_s^\tau ds, \end{aligned}$$

we now show that g^τ converges weakly in $L^2([0,T],\mathbb{R}^n)$ to $\left(J_t\dot{b}_t\right)_t$. Take a curve $(\phi_t)_t \in \mathcal{C}([0,T],\mathbb{R}^n)$ and compute

$$\begin{aligned} \int_0^T \langle \phi_t, g_t^\tau\rangle\, dt &= \sum_{k=0}^K \int_{k\tau}^{(k+1)\tau} \left\langle \phi_t, \tfrac{1}{\tau}\int_{k\tau}^{(k+1)\tau} \widehat{J}_s^\tau \dot{b}_s^\tau ds \right\rangle dt \\ &= \sum_{k=0}^K \tfrac{1}{\tau}\int_{k\tau}^{(k+1)\tau}\int_{k\tau}^{(k+1)\tau} \left\langle \phi_t, \widehat{J}_s^\tau \dot{b}_s^\tau \right\rangle dtds \\ &= \int_0^T \left\langle h_s^\tau, \dot{b}_s^\tau\right\rangle ds \end{aligned}$$

where

$$h_s^\tau = \tfrac{1}{\tau}\left(\widehat{J}_s^\tau\right)^* \int_{k\tau}^{(k+1)\tau} \phi_t dt \text{ for } s \in [k\tau,(k+1)\tau)$$

and M^* is the transpose of a matrix M. The sequence $(h^\tau)_\tau$ is uniformly bounded and converges pointwise to $t \mapsto J_t^*\phi_t$ using the convergence of $t \mapsto \widehat{J}_t^\tau$ and the fundamental theorem of calculus. The Lebesgue dominated convergence theorem gives strong convergence of the sequence in $L^2([0,T],\mathbb{R}^n)$ and thus the duality pairing with $\dot{b}^\tau$ converges from [40, Proposition 3.5 (iv)]. We have convergence of $\langle \phi, g^\tau\rangle_{L^2([0,T],\mathbb{R}^n)}$ to $\langle \phi, G_\mu\dot{\mu}\rangle_{L^2([0,T],\mathbb{R}^n)}$ for all continuous curves ϕ, and using the density of such curves in $L^2([0,T],\mathbb{R}^n)$ with the boundedness of the sequence, [44, Lemma 4.8-7] gives the weak limit. Writing

$$p_t^\tau \coloneqq p_{k+1}^\tau \text{ for } t \in [k\tau,(k+1)\tau)$$

i.e. $p_t^\tau = -V - g_t^\tau$, we have easily the convergence of the curves p^τ to $p = (-V - G_{\mu_t}\dot{\mu}_t)_t$ weakly in $L^2([0,T],\mathbb{R}^n)$. Thus, showing that $p_t \le 0$ and $\langle \mu_t, p_t\rangle = 0$ for a.e. t will give the desired result. Since $p_t^\tau \le 0$, we obtain for any curve λ valued in $\mathbb{R}_+$ and for any $x \in \mathcal{X}$ that

$$\begin{aligned} 0 \ge \int \lambda_t p_t^\tau(x) dt &= \int \langle \lambda_t \delta_x, p_t^\tau\rangle\, dt \\ &\to \int \langle \lambda_t\delta_x, p_t\rangle\, dt \end{aligned}$$

giving $p_t \le 0$ for a.e. t. We have from the continuity of B^{-1} given by Theorem 2.8 (b) that $\overline{\mu}^\tau$ converges to μ pointwise, and therefore the dominated convergence theorem yields that the convergence is also strong in $L^2([0,T],\mathbb{R}^n)$. As a result, since $\forall t, \langle \overline{\mu}_t^\tau, p_t^\tau\rangle = 0$ then

$$\begin{aligned} \int \lambda_t \langle \mu_t^\tau, p_t^\tau\rangle\, dt &= 0 \\ &\to \int \lambda_t \langle \mu_t, p_t\rangle \end{aligned}$$

which combined with $p_t \leq 0$ gives $\langle \mu_t, p_t \rangle = 0$ for a.e. t. □

Remark 5.8 (Obstructions and hopes in the general case). *When moving from a finite to a continuous space, the difficulty comes from the lack of homogeneity between the tangent spaces. Indeed, we have proven Proposition 5.6 when the derivative $(\dot{\mu}_t)_t$ is always valued in $\mathcal{M}(\mathcal{X})$ which is the case when all spaces are simply $\mathbb{R}^n$. However, generally speaking, we would have $\dot{\mu}_t \in \mathcal{H}^*_{\mu_t,0}$ when building the curve $(\mu_t)_t$ from the geodesic interpolation as we have done, and the operator $H_{\mu,\nu}$ from Definition 5.2 is at most extended to $\mathcal{H}^*_{\nu,\mu}$ which will not usually contain $\mathcal{H}^*_\mu$. The expression on the right hand side of* (5.1) *is thus ill defined for the more general class of curves. Still, the expected limit is known to be well defined from Theorem 4.1, and moreover it is retrieved as the limit $n \to \infty$ of the discretized space case. The limit results that we have are summarized by the diagram Figure 5.1 below.*

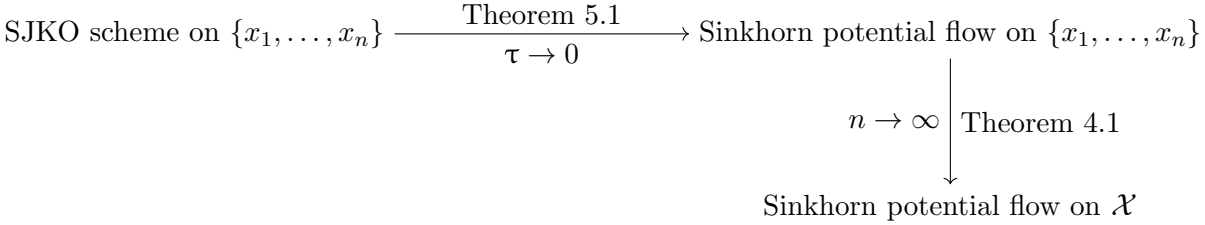


Figure 5.1: Summary of limit results

As a result, one could hope to show the diagram commutes under appropriate regularity assumptions to show the desired limit $\tau \to 0$ in the general case.

6 Numerics

In this section, we illustrate numerically the theory developed in the previous sections. The full code, parameters used and animations are available online [1] (some code was adapted from the Geomloss package [21], and we use the library PyTorch [45] for its automatic differentiation capabilities and GPU compatibility). We first detail the algorithm that was used before moving to the results.

6.1 Numerical scheme

We focus on how to compute a single SJKO step starting from $\mu \in \mathcal{P}(\mathcal{X})$, i.e. minimize the objective

$$\nu \mapsto 2\tau \langle \nu, V \rangle + S_\varepsilon(\nu, \mu) \tag{6.1}$$

over $\nu \in \mathcal{P}(\mathcal{X})$ (which has the same minimizer as the original objective in (3.1) but avoids the numerical instabilities of dividing by τ). To compute OT_ε appearing in the Sinkhorn divergence, the basic method is the Sinkhorn algorithm given by iterating alternatively the two conditions in the Schrödinger system (2.2), which converges to the solution of (2.1) [16, Theorem 4.2]. For the self transport of a measure, a more simple fixed-point algorithm can be utilized [21, Section 3.1]. The fact that τ will be small will make the convergence at each step faster since the values of the Schrödinger potentials at the previous step will already be close to optimal and thus make a good initialization. The gradients of the Sinkhorn divergence are easily deduced from the dual potentials computed by these algorithms [21, Section 3.2], which will allow us to perform each SJKO step through a first order method such as gradient descent. When considering an Eulerian discretization (i.e. fixing positions $\{x_1, \dots, x_n\}$ and representing a measure by its weights at each of these points), the objective (6.1) is convex in the weights (thanks to Theorem 1.3) with Lipschitz gradient [46], guaranteeing convergence of such methods to the optimum if the gradients are computed accurately enough. The drawback is the need to enforce the positivity and sum constraints on the weights, which slows down convergence. When considering a Lagrangian discretization (i.e. the weights are fixed to $\frac{1}{n}$ and the positions $\{x_1, \dots, x_n\}$ are variable), the convexity is lost but the need to enforce constraints disappears. We summarize in Pseudocode 1 below the resulting algorithm.

Input: measure μ, potential V
Initialize $\nu \leftarrow \mu$
Repeat until convergence (see (6.2)):
1. Compute $f_{\mu,\nu}$ and f_ν using Sinkhorn's algorithm [21, Section 3.1]
2. If the discretization is Eulerian:
   (a) Compute the gradient of (6.1) as $2\tau V + f_{\mu,\nu} - f_\nu$
   (b) Perform a gradient descent step and project it onto the probability simplex using [47] to update ν
3. If the discretization is Lagrangian:
   (a) Compute the gradients of (6.1) with respect to the positions using automatic differentiation [21, Section 3.2]
   (b) Perform a gradient step to update ν

Output: ν approximating the minimizer of (6.1)

Pseudocode 1: Algorithm to compute a single JKO step

The convergence criterion that we used is based on the optimality conditions (3.2), which we wrote before in $\mathcal{C}(X)/\mathbb{R}$ but the missing constant can be computed: denoting $g(\nu) \coloneqq 2\tau V + f_{\mu,\nu} - f_\nu$ for simplicity, the

[1] `https://github.com/mhardion/sjko_numerics`

conditions for ν to optimize (6.1) on $\mathcal{P}(\mathcal{X})$ write

$$\exists p \in \mathcal{C}(\mathcal{X}), \exists \lambda \in \mathbb{R}, \begin{cases} g(\nu) + p + \lambda = 0 \\ p \leq 0 \\ \langle \nu, p \rangle = 0 \\ \nu \in \mathcal{M}_+(\mathcal{X}) \\ \langle \nu, 1 \rangle = 1 \end{cases}$$

where by integrating the first line against ν we get $\lambda = -\langle \nu, g(\nu) \rangle$ and thus rewriting $p = \langle \nu, g(\nu) \rangle - g(\nu)$ (which immediately implies $\langle \nu, p \rangle = 0$) we obtain the conditions

$$\begin{cases} g(\nu) - \langle \nu, g(\nu) \rangle \geq 0 & (6.2) \\ \nu \in \mathcal{P}(\mathcal{X}). & (6.3) \end{cases}$$

The constraint (6.3) is enforced throughout the algorithm (by projection in Eulerian discretization, and by fixing the weights in Lagrangian discretization), so (6.2) can become our convergence criterion: we stop when the inequality is valid up to a small tolerance. This inequality is only checked on the positions $\{x_1, \ldots, x_n\}$ since those values are directly given by the algorithm. This is not a problem in the Eulerian case where those points describe the whole considered space, but it can lead to suboptimality in the Lagrangian case where the possible locations are still within a continuous space.

The next sections show the results of the simulations, in which c will be chosen as the square Euclidean distance.

6.2 The three-point space: embedding and rotational motion

When the space is made of three points (x_1, x_2, x_3), it is possible to visualize the RKHS sphere and the embedding of the flow on it (since the Schrödinger potentials are computed at each step), to illustrate the correspondence of the limiting PDE with constrained rotational motion (Remark 3.13). We do so via the change of basis given by $H_c^{-\frac{1}{2}}$ which makes $\mathcal{H}_c$ correspond to Euclidean space. We can then easily compute the axis of rotation of the unconstrained motion $\dot{b} + \mathbf{W}b$ to compare theory and numerics. The result is shown Figure 6.1, where we observe the rotation and the constraints as expected.

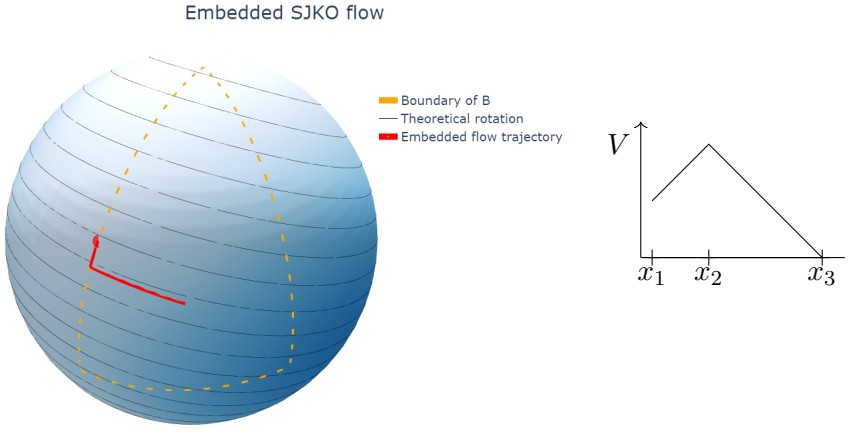


Figure 6.1: SJKO flow after embedding on the RKHS sphere (left), for the potential and three point space on the right. The boundary of the admissible set $\mathcal{B}$ is indicated with the dashed orange line. The initial point ($t = 0$) of the trajectory (red) is at the center and the endpoint ($t = 5$) is the circular marker. The heatmap on the sphere illustrates the quadratic form $E : b \mapsto \langle b, Vb \rangle_{\mathcal{H}_c}$ that the flow minimizes. The black lines indicate the rotation given by the theory. They are followed quite closely by the flow, and we can see the tendency of the latter to minimize the energy.

6.3 Flow of a single Dirac mass: SJKO versus classical gradient flow

We next illustrate the behavior of the flow of a single Dirac mass studied in Section 3.3, to observe its correspondence with the classical gradient flow when the potential is convex. We take a quadratic potential and compute the flow to compare it with the theoretical evolution. The flow is illustrated Figure 6.2 and the distance to the expected motion is given Figure 6.3. The latter stays low enough to be attributed to the numerical approximation at each step and decreases with τ, which stays consistent with Proposition 3.14.

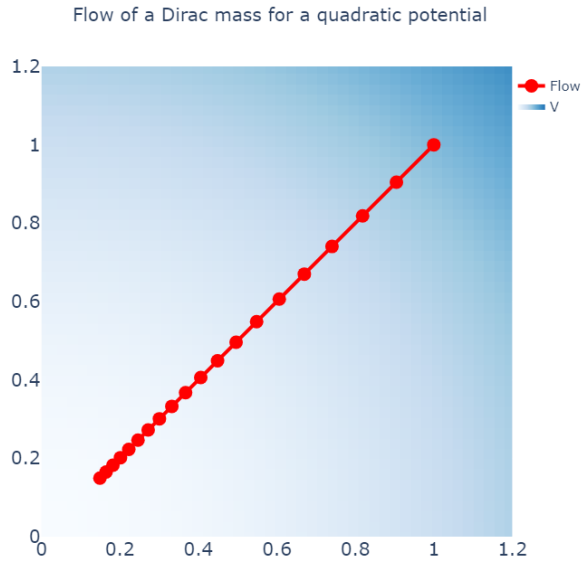
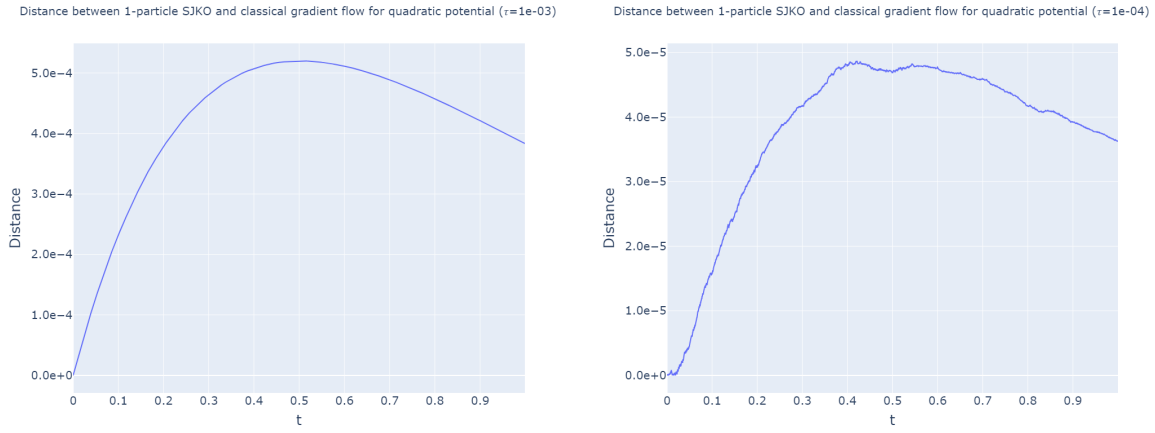


Figure 6.2: Flow of a single Dirac mass for $V = \|\cdot\|^2$. Here $\tau = 10^{-3}$ but the markers correspond to intervals of width 5×10^{-2} to better visualize the evolution of the velocity.



$\tau = 10^{-3}$ $\tau = 10^{-4}$

Figure 6.3: Distance between the SJKO gradient flow for a quadratic potential and the theoretical evolution given by the classical gradient flow ($x_t = e^{-2t}x_0$), with different values of τ. The distance stays of magnitude smaller than τ which tends to corroborate the expected result.

6.4 Convex and non convex potentials

We now show experiments on convex and non-convex potentials, in order to observe both horizontal and vertical perturbations. The latter can only happen in an Eulerian discretization since they can only be continuous by progressively modifying weights, but those are fixed in the Lagrangian case. This shows a tradeoff in the choice of discretization: Eulerian allows vertical perturbations but the optimization procedure is slower, and vice-versa for Lagrangian. In Figure 6.4 we show the evolution for a quadratic potential and relatively small ε in the Eulerian case. Figure 6.5 shows in Lagrangian discretization (most suited to the Wasserstein case) the difference between Sinkhorn and Wasserstein on the same potential. Figure 6.6 illustrates in Eulerian discretization the case of a non convex potential, showing that for large enough ε we observe vertical perturbations which allow convergence towards the minimum despite potential barriers (as proven Theorem 4.8). In contrast Figure 6.7 shows that a Lagrangian discretization does not allow this behavior.

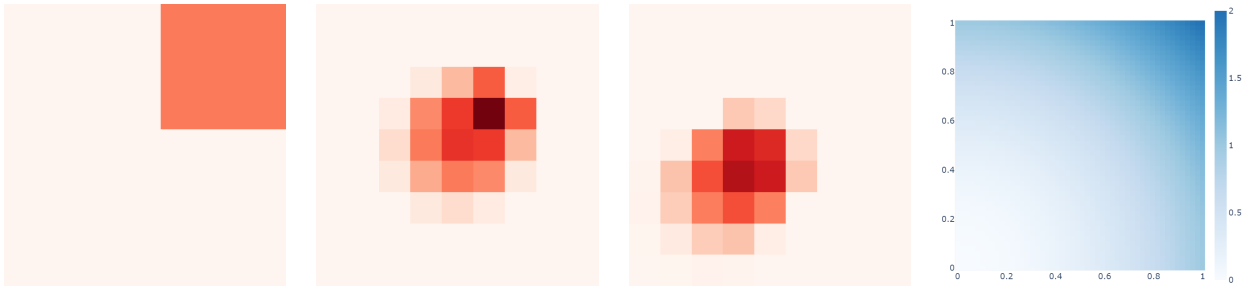


Figure 6.4: Eulerian flow for a quadratic potential (right). Darker red values indicate more mass. The regularization parameter is chosen as $\varepsilon = 0.2$ on a domain $\mathcal{X} = [0, 1]^2$ to be rather close to the Wasserstein case. We indeed observe a behavior closer to a continuity equation, where mass moves horizontally rather than vertically.

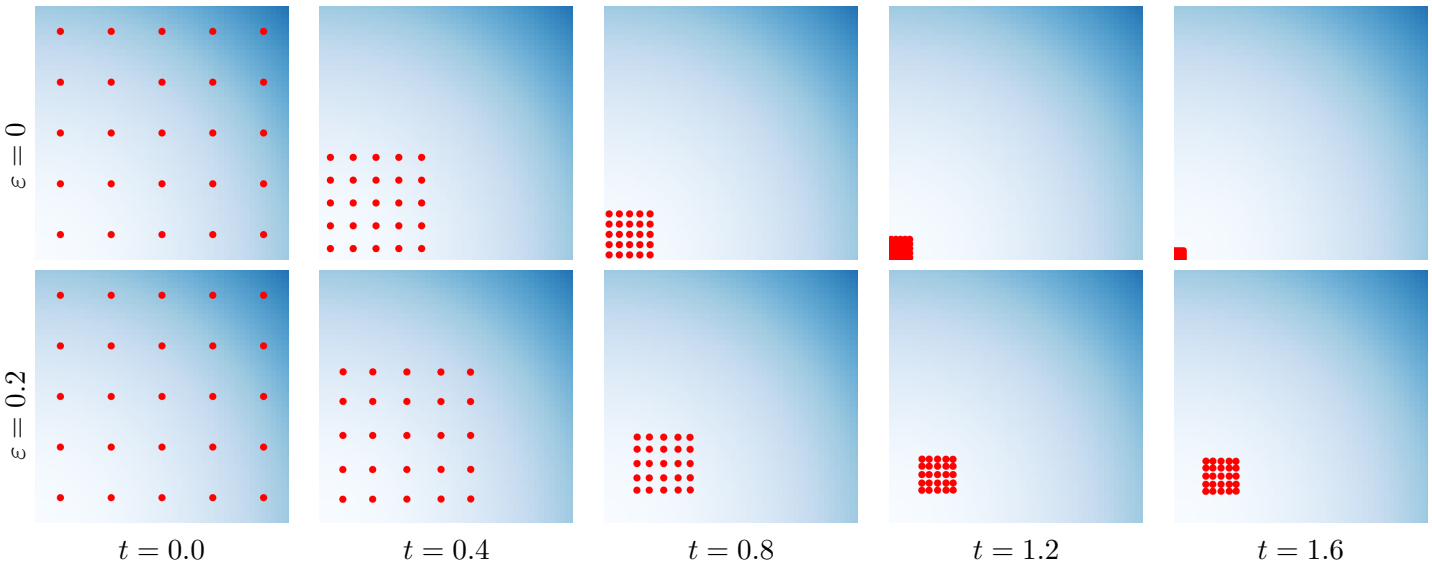


Figure 6.5: Lagrangian flow in with the same parameters as Figure 6.4, Wasserstein (top) against $S_\varepsilon, \varepsilon = 0.2$ (bottom). The positions of the masses are the red dots, and the blue heatmap illustrates the potential V. We observe in the Sinkhorn case interaction between the particles, which tend to aggregate together before moving towards the minimum, as opposed to the Wasserstein case where each particle follows the classical gradient flow independently.

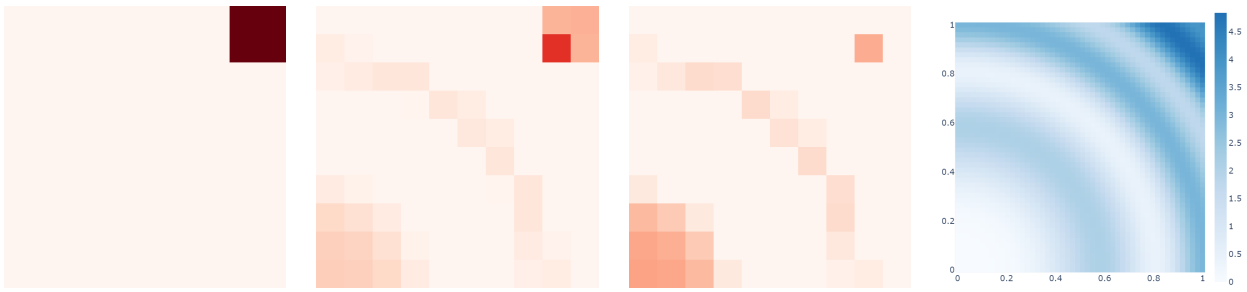


Figure 6.6: Eulerian flow for a non convex potential (shown right). Here $\varepsilon = 5$ which is large with respect to the diameter of $\mathcal{X} = [0, 1]^2$ to get close to the MMD behavior of the Sinkhorn divergence, and we observe a vertical displacement of the mass rather than horizontal, which transfers the mass towards the minimum despite the potential barrier.

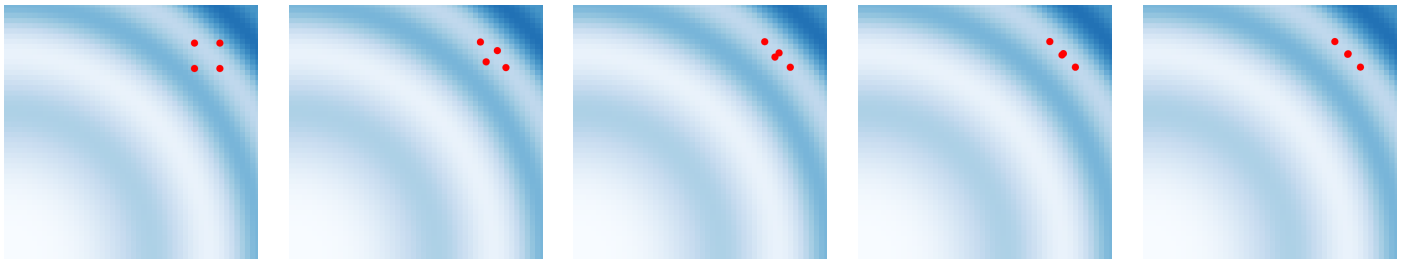


Figure 6.7: Lagrangian flow for the same potential and value of ε as Figure 6.6. We observe that the particles get stuck in a local minimum, since vertical perturbations are not allowed by the discretization scheme as mentioned above.

7 Conclusion and outlook

In this thesis, we have derived the differential equation corresponding to the gradient flow of a potential energy in the geometry induced by the Sinkhorn divergence, and studied its main properties including existence, uniqueness, contractivity and convergence in time to the minimum of the energy for possibly nonconvex potentials. Simple numerical observations illustrate the rotational structure of the motion, the presence of vertical perturbations allowing the long time convergence and interaction between particles as opposed to the Wasserstein case. A few directions of further research arise quite naturally, which we now discuss.

First, the recovery of the expected equation as the time step vanishes in the SJKO scheme was only proven in the case of a finite space, and future work could hope to prove a more general result as mentioned in Remark 5.8.

Next, the asymptotic behavior studied Section 4.2 was purely qualitative, but perhaps deriving estimates on the dissipation (by looking to generalize Proposition 4.5 to continuous spaces) and using a Grönwall lemma or the like could yield more quantitative information with regards to convergence speed.

Finally, now that we have a good basis to understand the case of a potential energy, one could imagine adding an entropic term to the energy and derive the flow of the resulting free energy in the Sinkhorn geometry, to study the difference with the Fokker-Planck equation recovered in the Wasserstein geometry. The analysis would be more involved, as nonlinearity would arise from the entropic term and the same change of variables used in our work may not yield an equation as interpretable as the one we obtained in the potential energy case.

Appendix

A. Short lemmas about Reproducing Kernel Hilbert Spaces

Lemma A.1. *Let $(\mathcal{H}_k, \langle\cdot,\cdot\rangle_{\mathcal{H}_k})$ be the RKHS for a continuous PD kernel k on $\mathcal{X}$ and $\|\cdot\|_{\mathcal{H}_k}$ the corresponding norm. Then for any $h \in \mathcal{H}_k \subset \mathcal{C}(\mathcal{X})$,*

$$\|h\|_\infty \leq \sup_{x\in X} k(x,x)\, \|h\|_{\mathcal{H}_k}\,. \tag{A.1}$$

Proof. For $x \in \mathcal{X}$, using the reproducing kernel property and Cauchy Schwarz yields

$$\begin{aligned} h(x) &= \langle k(x,\cdot), h\rangle_{\mathcal{H}_k} \\ &\leq \|k(x,\cdot)\|_{\mathcal{H}_k}\, \|h\|_{\mathcal{H}_k} \\ &= k(x,x)\, \|h\|_{\mathcal{H}_k} \end{aligned}$$

whence the result. □

Lemma A.2. *With the notations of Lemma A.1, the mapping $x \mapsto k(x,\cdot)$ from $\mathcal{X}$ to $\mathcal{H}_c$ (with its norm topology) is continuous.*

Proof. Taking $x_n \to x$ in $\mathcal{X}$, we have

$$\begin{aligned} \|k(x_n,\cdot) - k(x,\cdot)\|^2_{\mathcal{H}_k} &= \|k(x_n,\cdot)\|^2_{\mathcal{H}_k} + \|k(x_n,\cdot)\|^2_{\mathcal{H}_k} - 2\,\langle k(x_n,\cdot), k(x,\cdot)\rangle_{\mathcal{H}_k} \\ &= k(x_n,x_n) + k(x,x) - 2k(x_n,x) \end{aligned}$$

which converges to 0 from the continuity of k. □

Lemma A.3. *With the notations of Lemma A.1, define the Riesz isometry*

$$H_k^{-1} : \left| \begin{array}{ccc} \mathcal{H}_k & \longrightarrow & \mathcal{H}_k^* \\ h & \longmapsto & \langle h,\cdot\rangle_{\mathcal{H}_k} \end{array} \right. . \tag{A.2}$$

If the kernel k is universal, then the operator $H_k^{-1} : H_k\left[\mathcal{M}(\mathcal{X})\right] \to \mathcal{M}(\mathcal{X})$ is norm-to-weak-$*$ continuous.

Proof. The Riesz-Fréchet theorem [40, Theorem 5.5] gives that $H_k^{-1} : \mathcal{H}_k \to \mathcal{H}_k^*$ is an isometry and thus norm-to-norm continuous. Since k is universal, convergence in $\mathcal{H}_k^*$ implies weak-$*$ convergence of measures when restricting to $\mathcal{M}(\mathcal{X})$ (proven on $\mathcal{P}(\mathcal{X})$ in [48, Theorem 23], the proof still works in $\mathcal{M}(\mathcal{X})$). This yields the result. □

B. Morrey's inequality in the Sobolev space of Hilbert space valued curves

Lemma B.4. *Let $\mathcal{H}$ be a Hilbert space. For any $(h_t)_t \in \mathscr{H}^1([0,T];\mathcal{H})$, it holds that*

$$\forall t,s \in [0,T], \|h_t - h_s\|_{\mathcal{H}} \leq \left\|\dot{h}\right\|_{L^2([0,T];\mathcal{H}_c)} |t-s|^{\frac{1}{2}}\,. \tag{A.3}$$

PROOF. Using the fact that curves in $\mathscr{H}^1([0,T];\mathcal{H})$ can be written as the integral of their derivative [30, Theorem C.2] and Cauchy-Schwarz (in $L^2([0,T];\mathbb{R})$), we can write for all $t,s \in [0,T]$

$$\begin{aligned}\|h_t - h_s\|_{\mathcal{H}} &\leq \int_0^T \left\|\dot{h}_t\right\|_{\mathcal{H}} dt \\ &\leq \left\|\dot{h}\right\|_{L^2([0,T];\mathcal{H})} |t-s|^{\frac{1}{2}}\end{aligned}$$

i.e. (A.3). □

Bibliography

[1] F. Santambrogio. "{Euclidean, Metric, and Wasserstein} Gradient Flows: an overview". In: *Bulletin of Mathematical Sciences* 7 (Mar. 2017). DOI: 10.1007/s13373-017-0101-1.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. 2nd ed. Lectures in Mathematics. ETH Zürich. Basel: Birkhäuser Basel, 2008, pp. IX, 334. DOI: 10.1007/978-3-7643-8722-8.

[3] E. D. Giorgi. "New problems on minimizing movements". In: *Boundary Value Problems for PDE and Applications*. Paris: Masson, 1993, pp. 81–98.

[4] W. Rudin. *Real and complex analysis, 3rd ed.* USA: McGraw-Hill, Inc., 1987.

[5] C. Villani. *Optimal Transport: Old and New*. 1st ed. Vol. 338. Grundlehren der mathematischen Wissenschaften. Springer Berlin, Heidelberg, 2009, pp. XXII, 976. DOI: 10.1007/978-3-540-71050-9.

[6] R. Jordan, D. Kinderlehrer, and F. Otto. "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17. DOI: 10.1137/S0036141096303359.

[7] F. Santambrogio. *Optimal Transport for Applied Mathematicians. Calculus of Variations, PDEs, and Modeling*. 1st ed. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser Cham, 2015, pp. XXVII, 353. DOI: 10.1007/978-3-319-20828-2.

[8] J. Pedlosky. *Geophysical Fluid Dynamics*. 2nd ed. Springer Book Archive. New York, NY: Springer-Verlag New York Inc., 1987, pp. XIV, 710. DOI: 10.1007/978-1-4612-4650-3.

[9] E. M. Purcell and D. J. Morin. *Electricity and magnetism*. 3rd. Cambridge University Press, 2013, p. 4.

[10] D. W. Stroock and S. R. S. Varadhan. *Multidimensional Diffusion Processes*. 1st ed. Classics in Mathematics. Berlin, Heidelberg: Springer Berlin, Heidelberg, 2006, pp. XII, 338. DOI: 10.1007/3-540-28999-2.

[11] F. Otto. "The geometry of dissipative evolution equations: the porous medium equation". In: *Communications in Partial Differential Equations* 26.1-2 (2001), pp. 101–174. DOI: 10.1081/PDE-100002243.

[12] A. Blanchet, V. Calvez, and J. A. Carrillo. "Convergence of the Mass-Transport Steepest Descent Scheme for the Subcritical Patlak–Keller–Segel Model". In: *SIAM Journal on Numerical Analysis* 46.2 (2008), pp. 691–721. DOI: 10.1137/070683337.

[13] B. Maury, A. Roudneff-Chupin, and F. Santambrogio. "A macroscopic crowd motion model of gradient flow type". In: *M3AS* 20.10 (2010), pp. 1787–1821. URL: http://cvgmt.sns.it/paper/269/.

[14] C. Bunne, L. Papaxanthos, A. Krause, and M. Cuturi. "Proximal Optimal Transport Modeling of Population Dynamics". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 6511–6528. URL: https://proceedings.mlr.press/v151/bunne22a.html.

[15] A. T. Lin, W. Li, S. Osher, and G. Montufar. *Wasserstein Proximal of GANs*. 2021. URL: https://arxiv.org/abs/2102.06862.

[16] G. Peyré and M. Cuturi. *Computational Optimal Transport*. 2020.

[17] M. Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

[18] C. Léonard. "From the Schrödinger problem to the Monge-Kantorovich problem". In: *Journal of Functional Analysis* 262 (2012), pp. 1879–1920. URL: https://hal.science/hal-00534910.

[19] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. *Sample Complexity of Sinkhorn divergences*. 2019.

[20] A. Genevay, G. Peyré, and M. Cuturi. "Learning Generative Models with Sinkhorn Divergences". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Vol. 84. Proceedings of Machine Learning Research. PMLR, Apr. 2018, pp. 1608–1617. URL: https://proceedings.mlr.press/v84/genevay18a.html.

[21] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, Apr. 2019, pp. 2681–2690. URL: https://proceedings.mlr.press/v89/feydy19a.html.

[22] C. A. Micchelli, Y. Xu, and H. Zhang. "Universal Kernels". In: *Journal of Machine Learning Research* 7.95 (2006), pp. 2651–2667. URL: http://jmlr.org/papers/v7/micchelli06a.html.

[23] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2011. URL: https://books.google.it/books?id=bX3TBwAAQBAJ.

[24] G. Peyré. "Entropic Approximation of Wasserstein Gradient Flows". In: *SIAM Journal on Imaging Sciences* 8.4 (2015), pp. 2323–2351. DOI: 10.1137/15M1010087.

[25] R. L. Dykstra. "An Iterative Procedure for Obtaining I-Projections onto the Intersection of Convex Sets". In: *The Annals of Probability* 13.3 (1985), pp. 975–984. URL: http://www.jstor.org/stable/2243723 (visited on 08/20/2024).

[26] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. "Convergence of Entropic Schemes for Optimal Transport and Gradient Flows". In: *SIAM Journal on Mathematical Analysis* 49 (Dec. 2015). DOI: 10.1137/15M1050264.

[27] A. Baradat. "Using Sinkhorn in JKO adds diffusion in the limiting PDE". In: *Applications of Optimal Transport, Mathematisches Forschungsinstitut Oberwolfach, Report No. 2406/2024*. Mathematisches Forschungsinstitut Oberwolfach. Feb. 2024, pp. 14–16. URL: https://publications.mfo.de/handle/mfo/4152.

[28] D. Adams, M. H. Duong, and G. dos Reis. "Entropic Regularization of NonGradient Systems". In: *SIAM Journal on Mathematical Analysis* 54.4 (2022), pp. 4495–4535. DOI: 10.1137/21M1414668.

[29] B. Schmitzer. *Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems*. 2019. URL: https://arxiv.org/abs/1610.06519.

[30] H. Lavenant, J. Luckhardt, G. Mordant, B. Schmitzer, and L. Tamanini. *The Riemannian geometry of Sinkhorn divergences*. 2024. URL: https://arxiv.org/abs/2405.04987.

[31] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations*. 1st ed. Vol. 44. Applied Mathematical Sciences. New York, NY: Springer-Verlag New York, Inc., 1983, pp. X, 282. DOI: 10.1007/978-1-4612-5561-1.

[32] J. Peypouquet and S. Sorin. *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*. 2009.

[33] H. Bauschke and J. (Borwein. "Continuous Linear Monotone Operators on Banach Spaces". In: *CiteSeerX* (Sept. 1995).

[34] R. R. Phelps. "Lectures on maximal monotone operators". In: *Extracta Mathematicae* 12.3 (1997), pp. 193–230.

[35] R. T. Rockafellar. "On the Maximality of Sums of Nonlinear Monotone Operators". In: *Transactions of the American Mathematical Society* 149.1 (1970), pp. 75–88. URL: http://www.jstor.org/stable/1995660 (visited on 05/24/2024).

[36] M. G. Crandall and A. Pazy. "Semi-groups of nonlinear contractions and dissipative sets". In: *Journal of Functional Analysis* 3.3 (1969), pp. 376–418. DOI: https://doi.org/10.1016/0022-1236(69)90032-9.

[37] J. B. Conway. *A Course in Functional Analysis*. 2nd ed. Vol. 96. Graduate Texts in Mathematics. New York, NY: Springer, 2007, pp. XVI, 400. DOI: 10.1007/978-1-4757-4383-8.

[38] M. Kreuter. "Sobolev Spaces of Vector-Valued Functions". MA thesis. Ulm University, 2015.

[39] J. Simon. "Compact sets in the space $L^p(0,T;B)$". In: *Annali di Matematica Pura ed Applicata* 146 (Jan. 1986), pp. 65–96. DOI: 10.1007/BF01762360.

[40] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. 1st ed. Universitext. Springer New York, NY, 2011, pp. XIV, 600. DOI: 10.1007/978-0-387-70914-7.

[41] J. Dieudonné. *Foundations of Modern Analysis*. Pure and Applied Mathematics. Academic Press New York and London, 1960.

[42] T. Séjourné, J. Feydy, F.-X. Vialard, A. Trouvé, and G. Peyré. *Sinkhorn Divergences for Unbalanced Optimal Transport*. 2023. URL: https://arxiv.org/abs/1910.12958.

[43] S. Izumino. "Convergence of generalized inverses and spline projectors". In: *Journal of Approximation Theory* 38.3 (1983), pp. 269–278. DOI: https://doi.org/10.1016/0021-9045(83)90133-8.

[44] E. Kreyszig. *Introductory functional analysis with applications*. Vol. 17. John Wiley & Sons, 1991.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. URL: https://arxiv.org/abs/1912.01703.

[46] G. Carlier, L. Chizat, and M. Laborde. "Lipschitz Continuity of the Schrödinger Map in Entropic Optimal Transport". Oct. 2022. URL: https://hal.science/hal-03793562.

[47] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. "Efficient projections onto the ℓ_1-ball for learning in high dimensions". In: *ICML '08: Proceedings of the 25th international conference on Machine learning*. New York, NY, USA, 2008, pp. 272–279. URL: http://doi.acm.org/10.1145/1390156.1390191.

[48] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. "Hilbert Space Embeddings and Metrics on Probability Measures". In: *Journal of Machine Learning Research* 11.50 (2010), pp. 1517–1561. URL: http://jmlr.org/papers/v11/sriperumbudur10a.html.