Ann. Inst. H. Poincaré C
Anal. Non Linéaire 43 (2026), 1079–1144
DOI 10.4171/AIHPC/165

The Riemannian geometry of Sinkhorn divergences

Hugo Lavenant, Jonas Luckhardt, Gilles Mordant, Bernhard Schmitzer, and Luca Tamanini

Abstract. We propose a new metric between probability measures on a compact metric space that mirrors the Riemannian-manifold-like structure of quadratic optimal transport but includes entropic regularization. Its metric tensor is given by the Hessian of the Sinkhorn divergence, a debiased variant of entropic optimal transport. We precisely identify the tangent space it induces, which turns out to be related to a reproducing kernel Hilbert space (RKHS). As usual in Riemannian geometry, the distance is built by looking for shortest paths. We prove that our distance is geodesic, metrizes the weak-$*$ topology, and is equivalent to an RKHS norm. Still it retains the geometric flavor of optimal transport: as a paradigmatic example, translations are geodesics for the quadratic cost on $\mathbb{R}^d$. We also show two negative results on the Sinkhorn divergence that may be of independent interest: that it is not jointly convex, and that its square root is not a distance because it fails to satisfy the triangle inequality.

Contents

Mathematics Subject Classification 2020: 49Q22 (primary); 51F99 (secondary).
Keywords: Riemannian metric, entropic optimal transport, Sinkhorn divergence, reproducing kernel Hilbert spaces.

1. Introduction

Optimal transport is a mathematical optimization problem which allows one to lift a metric d on a space X to the space $\mathcal{P}(X)$ of probability measures over X, interpreted as the average distance that infinitesimal units of mass have to travel when efficiently transforming one measure into the other. When the base space X is a Riemannian manifold, then the space of probability measures can be formally interpreted as a Riemannian manifold as well.

In recent years there has been an increasing interest in entropic regularization of optimal transport, which adds some stochasticity to the coupling of probability measures. This variant is easier to compute numerically, has better statistical properties, and is more regular as a function of its input measures and ground cost. Its debiased version defines a smooth positive definite loss function on $\mathcal{P}(X)$. However, it does not define a distance, or even a Riemannian geometry, on $\mathcal{P}(X)$.

> In this work we aim at combining both worlds: having a Riemannian metric on $\mathcal{P}(X)$ which lifts the geometry of X in a similar way to what optimal transport does, while retaining the smoothness of the Sinkhorn divergence.

To that end, we consider the second-order expansion of the Sinkhorn divergence on the diagonal: this yields a positive definite quadratic form on a meaningful tangent space to the space of probability measures, in other words a Riemannian metric tensor. This metric tensor induces a Riemannian distance on $\mathcal{P}(X)$ via minimization over lengths of curves, which is aware of the geometry of X while being "smoother" than the optimal transport geometry in the sense that the metric tensor is jointly continuous in its inputs. Investigation of the additional properties of our distance, in particular regarding computational and statistical aspects, is postponed to future works. In the rest of this introduction we recall what is meant by the "geometry of optimal transport" before summarizing the main results of the present work. We emphasize that geodesics for our distance are different from Schrödinger bridges: a more detailed discussion is given at the end of the introduction.

Optimal transport and its geometry. On a (compact) space X, the value of the optimal transport problem with ground cost $c\colon X \times X \to [0,\infty)$ between two probability measures $\mu, \nu \in \mathcal{P}(X)$ will be denoted by $\mathrm{OT}_0(\mu,\nu)$. It is defined as [4, 68, 74]

$$\mathrm{OT}_0(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint_{X\times X} c(x,y)\,\mathrm{d}\pi(x,y), \tag{1.1}$$

where $\Pi(\mu,\nu) = \{\pi \in \mathcal{P}(X\times X) \mid (\mathrm{P}_1)_\#\pi = \mu,\ (\mathrm{P}_2)_\#\pi = \nu\}$ is the set of couplings having marginals μ, ν, and $(P_i)_\#\pi$ is the push-forward of the measure π by the projection P_i on the i-th factor. If $c = d^p$ is a power of a distance d on X for some $p \geq 1$, then $(\mu,\nu) \mapsto \mathrm{OT}_0(\mu,\nu)^{\frac{1}{p}}$ defines a distance on $\mathcal{P}(X)$, the so-called p-Wasserstein or Monge–Kantorovich distance [4, Sect. 7].

To explain the *geometry* it induces, we restrict to X being a convex subset of a Euclidean space $\mathbb{R}^d$ and $c(x, y) = \|x - y\|^2$, though everything could be extended to Riemannian manifolds. Historically this geometric structure was introduced in [59, Sect. 4] as the quotient of the space of maps by measure-preserving maps with applications to dissipative evolution equations, and can be traced back to hydrodynamics [5, 12]. In this presentation we will rather focus on the "metric tensor" that optimal transport defines. If $(\mu_t)_t$ is a curve of probability measures obtained by following the flow of a potential velocity field $v = \nabla\psi$, that is, $\mu_t = \Psi(t, \cdot)_\# \mu_0$ is the push-forward of μ_0 by the flow $\Psi(t, \cdot)$ of the ODE $\dot{x}_t = \nabla\psi(x_t)$, then by [4, Prop. 8.5.6],

$$\lim_{t \to 0} \frac{\mathrm{OT}_0(\mu_0, \mu_t)}{t^2} = \int_X |\nabla\psi(x)|^2 \,\mathrm{d}\mu_0(x). \tag{1.2}$$

That is, on small distances, the squared Wasserstein distance is quadratic in $v = \nabla\psi$ which parametrizes how the curve moves on $\mathcal{P}(X)$. As the curve $(\mu_t)_t$ satisfies the continuity equation $\dot{\mu}_t = -\operatorname{div}(\mu_t \nabla\psi)$, we can think of the first-order distribution $\dot{\mu}_t$ as the tangent vector whose squared length is given by the right-hand side of (1.2). That is, for $\mu \in \mathcal{P}(X)$ the formal metric tensor $\mathbf{g}^0$ is

$$\mathbf{g}^0_\mu(\dot{\mu}, \dot{\mu}) = \int_X |\nabla\psi(x)|^2 \,\mathrm{d}\mu(x), \quad \dot{\mu} = -\operatorname{div}(\mu\nabla\psi). \tag{1.3}$$

That $\mathcal{P}(X)$ has such a Riemannian structure means that the squared distance OT_0 can be recovered by minimizing the energy of paths joining μ to ν:

$$\mathrm{OT}_0(\mu, \nu) = \inf_{(\mu_t)_{t\in[0,1]}} \int_0^1 \mathbf{g}^0_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t) \,\mathrm{d}t, \tag{1.4}$$

where the infimum is taken over sufficiently regular curves $(\mu_t)_{t\in[0,1]}$ of measures joining μ to ν: this is the dynamical formulation of optimal transport, the so-called Benamou–Brenier formula [6], [68, Sect. 6.1].

In summary, for $\mu \in \mathcal{P}(X)$ a "point" in the manifold, the tangent vectors at μ are given by $\dot{\mu} = -\operatorname{div}(\mu\nabla\psi)$ for $\psi \colon X \to \mathbb{R}$, and $\mathbf{g}^0_\mu(\dot{\mu}, \dot{\mu})$ defined in (1.3) is the squared length of the tangent vector $\dot{\mu}$. The proper tangent space is actually the completion of $\{-\operatorname{div}(\mu\nabla\psi) \mid \psi \in \mathcal{C}^\infty(X)\}$ with respect to $\mathbf{g}^0_\mu$. It is canonically identified with vector fields $v \in L^2(X, \mu; \mathbb{R}^d)$ which can be approximated in $L^2(X, \mu; \mathbb{R}^d)$ by smooth gradients. Even though this does not define an infinite-dimensional Riemannian manifold in the sense of [41], a precise meaning can still be given thanks to the tools of analysis on non-smooth spaces [4, Sect. 8].

Once viewing the problem through this lens, we can import many concepts of Riemannian geometry into $\mathcal{P}(X)$. Geodesics are defined as the minimizers of the right-hand side of (1.4) and correspond to McCann's displacement interpolation [51]. For two probability measures $\mu, \nu \in \mathcal{P}(X)$, the Riemannian logarithmic map $\log_\mu(\nu)$ is the initial velocity of the geodesic from μ to ν. Conversely, given a tangent vector $\dot{\mu} = -\operatorname{div}(\mu\nabla\psi)$, the exponential map is $\exp_\mu(\dot\mu) = (\mathrm{id} + \nabla\psi)_\#\mu$, but the extension to the whole tangent space and the injectivity radius can be tricky to characterize. With logarithmic and exponential maps, one can for instance define a linearized version of optimal transport [69, 75] and subsequently perform principal component analysis in tangent space [11]. Several extensions of optimal transport rely on its Riemannian structure, such as barycenters [3], gradient flows [4], harmonic maps [43], parallel transport [33], to cite a few examples. As the Riemannian interpretation is mostly formal, giving a rigorous meaning to these concepts usually requires some (hard) work on analysis on non-smooth spaces.

The Sinkhorn divergence. For a regularization parameter $\varepsilon > 0$, entropic optimal transport [24, 61] is given by the minimization problem

$$\mathrm{OT}_\varepsilon(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \Big\{ \iint_{X\times X} c \,\mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi \mid \mu\otimes\nu) \Big\}, \tag{1.5}$$

where KL is the Kullback–Leibler divergence

$$\mathrm{KL}(\pi \mid \rho) := \begin{cases} \displaystyle\int_X h\Big(\frac{\mathrm{d}\pi}{\mathrm{d}\rho}\Big)\,\mathrm{d}\rho & \text{if } \pi \ll \rho,\ \rho \geq 0, \\ +\infty & \text{otherwise,} \end{cases}$$

with $h(s) := s\log(s) - s + 1$ for $s > 0$, $h(0) := 1$, $h(s) := +\infty$ for $s < 0$. As $\varepsilon \to 0$, OT_ε converges to OT_0 [17, 44, 54]. Compared to the original transport problem it can be computed faster [61] and the statistical complexity suffers much less from the curse of dimensionality [30, 52]. Moreover, as a function of μ and ν, the cost OT_ε is more regular and its gradients can be computed efficiently [31].

However, even if $c = d^2$ for d a distance on X, the function $\sqrt{\mathrm{OT}_\varepsilon}$ is not a distance, as generically $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$ and $\operatorname{argmin}_{\nu\in\mathcal{P}(X)} \mathrm{OT}_\varepsilon(\mu,\nu) \neq \mu$. For this reason [31] introduced the Sinkhorn divergence, or debiased entropic optimal transport:

$$S_\varepsilon(\mu,\nu) := \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu). \tag{1.6}$$

By construction it satisfies $S_\varepsilon(\mu,\mu) = 0$, in fact more generally $S_\varepsilon(\mu,\nu) \geq 0$ with equality if and only if $\mu = \nu$, and it also converges to OT_0 as $\varepsilon \to 0$. In [28] the authors prove the following theorem which justifies its use in machine learning as a loss function.

Theorem 1.1 ([28, Thm. 1]). *When $k_c = \exp(-c/\varepsilon)$ is a positive definite, universal, jointly Lipschitz continuous kernel over a compact space X, then S_ε defines a symmetric, positive definite, and smooth loss function that is convex in each of its input variables. It also metrizes the convergence in law.*

For the definition of positive definiteness and universality of the kernel see Section 4.1. The assumptions on the kernel are satisfied in particular for the quadratic cost on a Euclidean domain. Nevertheless, $\sqrt{S_\varepsilon}$ is not a distance, because it does not satisfy the triangle inequality. Although this is well known in the community, we have not been able to find an explicit statement of this property and consequently provide an elementary proof in Section 7.1 of this article.

Outline of the contributions. As mentioned earlier we seek to build a Riemannian distance out of the loss function S_ε. We restrict to X being a compact metric space with a continuous cost function $c \in \mathcal{C}(X \times X)$. Similarly to [28], we assume that it induces a positive definite universal kernel $k_c := \exp(-c/\varepsilon)$ on $X \times X$ (see [53, 72]).

Throughout this article we consider some fixed $\varepsilon > 0$ and we will briefly comment on the limits $\varepsilon \to 0$ and $\varepsilon \to \infty$ in Section 4.4.

The Hessian of the Sinkhorn divergence. To construct a Riemannian structure based on the Sinkhorn divergence, we follow an approach from information geometry. For each point $\mu \in \mathcal{P}(X)$ we build a metric tensor, that is, a positive definite quadratic form $\mathbf{g}_\mu$ on some appropriate tangent space to $\mathcal{P}(X)$ at μ, by considering the Hessian of $S_\varepsilon(\mu, \nu)$ on the diagonal $\mu = \nu$. Applying this strategy to the Kullback–Leibler divergence in place of the Sinkhorn divergence, as in the pioneering works of Hotelling [38] and Rao [63], leads to the Fisher information metric. In Theorem 3.4 we derive the formula for the Hessian of S_ε along paths $(\mu_t)_t$ in $\mathcal{P}(X)$ following a $\mathcal{C}^m$-perturbation, meaning that $\dot{\mu}_t$ exists and is continuous as a linear form on $\mathcal{C}^m(X)$ (see Definition 3.1). The case of a "horizontal" perturbation $\mu_t = \Psi(t, \cdot)_\# \mu_0$ corresponds to $m = 1$, while a "vertical perturbation" $\mu_t = \mu_0 + t\nu$ with $\nu \in \mathcal{M}_0(X)$ corresponds to $m = 0$. It requires the cost function to be of class $\mathcal{C}^m$, with the setting $m = 0$ being the one which encompasses continuous cost functions on compact metric space. Writing $(f_{\mu_t,\mu_s}, g_{\mu_t,\mu_s})$ for the entropic dual variables between μ_t and μ_s (i.e. the entropic analogues of Kantorovich potentials, see Section 2), the formula reads

$$\lim_{t\to 0} \frac{S_\varepsilon(\mu, \mu_t)}{t^2} = \frac{1}{2}\Big\langle \dot{\mu}, \frac{\partial f_{\mu,\mu_s}}{\partial s}\Big|_{s=0} \Big\rangle = \frac{\varepsilon}{2}\langle \dot{\mu}, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}]\rangle. \tag{1.7}$$

Here, K_μ and H_μ, introduced in Definition 3.3, are integral operators induced by the self-transport kernel

$$k_\mu := \exp\Big(\frac{1}{\varepsilon}(f_{\mu,\mu} \oplus f_{\mu,\mu} - c)\Big).$$

The result is not completely new and can be related to works studying the central limit theorem for the Sinkhorn divergence [36,37], where the "perturbations" μ_t are estimations of the measure μ by i.i.d. samples.

The metric tensor and its canonical tangent space. We *define* the metric tensor $\mathbf{g}_\mu$ as the right-hand side of (1.7), that is, at least for a signed measure $\dot{\mu}$ which is "balanced" in the sense $\langle \dot{\mu}, \mathbb{1}_X\rangle = 0$,

$$\mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) = \frac{\varepsilon}{2}\langle \dot{\mu}, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}]\rangle.$$

We then want to understand for which class of "tangent vectors" $\dot{\mu}$ this can be extended. This leads us to the most technical part of the paper in Section 4. Whereas we have already seen that distributions of order m should be allowed if the cost function is $\mathcal{C}^m$, it turns out that the right regularity condition of the perturbation is expressed in terms of the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mu$ induced by the self-transport kernel k_μ. Specifically, $\mathbf{g}_\mu$ defines a positive definite quadratic form on the *dual* $\mathcal{H}_\mu^*$ of the RKHS $\mathcal{H}_\mu$, restricted to functionals $\dot{\mu}$ such that $\langle \dot{\mu}, \mathbb{1}_X \rangle = 0$. On this subspace of $\mathcal{H}_\mu^*$ the norm induced by the metric tensor $\mathbf{g}_\mu$ is equivalent to the operator norm. For a smooth cost function c this space contains all $\mathcal{C}^m$-perturbations, including the "vertical" and "horizontal" ones. In contrast, the tangent space for the optimal transport metric tensor (1.3) only allows for horizontal perturbations.

A useful change of variables. In order to build a distance from the metric tensor $\mathbf{g}_\mu$ and to prove the existence of geodesics, we need some uniform coercivity. To that end we exploit the change of variables $\mu \leftrightarrow \beta := \exp(-f_{\mu,\mu}/\varepsilon)$ which maps $\mathcal{P}(X)$ into a subset of the unit sphere of the reproducing kernel Hilbert spaces $\mathcal{H}_c$ induced by $k_c := \exp(-c/\varepsilon)$ (Theorem 4.8). This change of variables is not completely new and can be seen implicitly in [28, Prop. 3] when the authors derive lower bounds on the Sinkhorn divergence. For a perturbation $\dot{\beta}$ of the variable $\beta = \exp(-f_{\mu,\mu}/\varepsilon)$ and its associated perturbation $\dot{\mu}$ of the measure μ, Theorem 4.12 expresses $\mathbf{g}_\mu$ as

$$\mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) = \tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta}) := \frac{\varepsilon}{2}\left(\|\dot{\beta}\|_{\mathcal{H}_c}^2 + 2\left\langle \exp(f_{\mu,\mu}/\varepsilon) \cdot \dot{\beta}, (\mathrm{id} - K_\mu)^{-1}[\exp(f_{\mu,\mu}/\varepsilon) \cdot \dot{\beta}]\right\rangle_\mu\right).$$

Expressed in this space $\mathcal{H}_c$, independent of μ, the metric tensor $\tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta})$ is jointly continuous in μ and $\dot{\beta}$ (Proposition 4.14) and equivalent to $\|\dot{\beta}\|_{\mathcal{H}_c}^2$ the squared norm on $\mathcal{H}_c$ uniformly in μ (Proposition 4.15). This is in stark contrast with the metric tensor $\mathbf{g}_\mu^0$ of the optimal transport problem (1.3), which depends in a non-smooth way on μ and which does not compare well to any other functional space uniformly on $\mathcal{P}(X)$.

Scaling limit. In Section 4.4 we verify, though without rigorous derivations, that the scaling limits of our metric tensor when $\varepsilon \to 0$ and $\varepsilon \to \infty$ indeed correspond to what is expected, respectively the metric tensor of optimal transport and that of a degenerate MMD distance.

Building a new distance. Once we understand our metric tensor $\mathbf{g}_\mu$ well, we build our new metric on the space of probability measures in Section 5. While in the classical case the "dynamic" formula (1.4) is a consequence of the "static" definition (1.1), here we take it as a definition of the distance. Specifically, in Definition 5.1 we set

$$\mathsf{d}_S(\mu_0, \mu_1) := \left(\inf \int_0^1 \mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)\, \mathrm{d}t\right)^{\frac{1}{2}}, \tag{1.8}$$

where the infimum is taken over a suitable set of admissible paths $(\mu_t)_{t\in[0,1]}$ between μ_0 and μ_1. Theorem 5.2 shows that d_S is a metric which metrizes the weak-$*$ topology on $\mathcal{P}(X)$. We actually prove more: $\mathsf{d}_S(\mu_0, \mu_1)$ is uniformly equivalent to $\|\beta_0 - \beta_1\|_{\mathcal{H}_c}$. This is to be compared to optimal transport, where $\sqrt{\mathrm{OT}_0}$ can be compared to an H^{-1} norm as in [62] and [68, Sect. 5.5.2], but these results require some regularity assumptions on the measures involved and are not valid uniformly on $\mathcal{P}(X)$. Estimates can also be made with linearized optimal transport [25], though this only leads to a bi-Hölder equivalence. The price to pay is that our constants depend on ε and blow up as $\varepsilon \to 0$.

Existence of geodesics. In Theorem 5.4 we prove existence of geodesics, that is, we show that the infimum in (1.8) is attained. In most of the problems which are variations of the Benamou–Brenier formula, such as unbalanced optimal transport [20], variational mean field games [8], or harmonic maps [43], the existence of solutions to problems of the type (1.8) uses the property that $\mathbf{g}^0_\mu(\dot\mu, \dot\mu)$ (or its suitable counterpart) is jointly convex in μ and $\dot\mu$. Thus tools of convex analysis can be used and lower semicontinuity is easy to establish. In our case, however, neither the Sinkhorn divergence $S_\varepsilon(\mu, \nu)$ nor the associated metric tensor $\mathbf{g}_\mu(\dot\mu, \dot\mu)$ are jointly convex: we provide counterexamples that we defer to Section 7.2. To prove that the infimum in (1.8) is attained we still rely on the direct method of calculus of variations, but we have to work a little bit harder: coercivity comes from the change of variables $\mu \leftrightarrow \beta$ studied in Section 4, and lower semicontinuity is proved thanks to the smoothness of the metric tensor in its inputs.

Absolutely continuous paths. The class of admissible paths in the minimization problem (1.8) defining the metric d_S is initially defined via Sobolev regularity of the corresponding path $(\beta_t)_t$ in $\mathcal{H}_c$. These paths end up being precisely the 2-absolutely continuous paths in $(\mathcal{P}(X), \mathsf{d}_S)$. We give a characterization of admissibility in terms of the path $(\mu_t)_t$ itself and some regularity of the potentials in Proposition 5.8. If c is smooth we show in Corollary 5.10 that $\mathcal{C}^m$-perturbations, for which the metric tensor coincides with the Hessian of the Sinkhorn divergence, are admissible paths. We further show that

$$\mathsf{d}_S(\mu_0, \mu_1) \le C \cdot W_1(\mu_0, \mu_1) \quad \text{and} \quad \mathsf{d}_S(\mu_0, \mu_1) \le C \cdot \mathsf{SHK}(\mu_0, \mu_1)$$

for some $C > 0$ depending on ε, where W_1 denotes the 1-Wasserstein distance and SHK the spherical Hellinger–Kantorovich distance (see [42]). In particular, 2-absolutely continuous paths in the p-Wasserstein metric for any $p \ge 1$ are admissible paths for d_S.

The quadratic case. Our theory applies to *any* compact metric space with *any* continuous cost function so long as $k_c = \exp(-c/\varepsilon)$ defines a positive definite universal kernel. The special case of the squared Euclidean distance as the cost function on a subset of $\mathbb{R}^d$ is studied in Section 5.4. For measures μ_0, μ_1 it has been known for a while (see e.g. [10, Lem. 8.8] or [29, proof of Thm. 2.1]) that the following decomposition holds:

$$\mathrm{OT}_0(\mu_0, \mu_1) = \|m_1 - m_0\|^2 + \mathrm{OT}_0(\bar\mu_0, \bar\mu_1),$$

where m_i denotes the mean of the measure μ_i and $\bar\mu_i = (\mathrm{id} - m_i)_\# \mu_i$ is the centered version of μ_i, for $i = 0, 1$. In other words, the transport of the means and the transport of the centered parts decouple. We show that this extends to Sinkhorn divergences and that it can be read as an orthogonal decomposition of the tangent space for the geometry of $\mathbf{g}_\mu$ in Proposition 5.14. We then integrate this inequality and show in Theorem 5.16 that our novel distance satisfies

$$\mathsf{d}_S(\mu_0, \mu_1)^2 = \|m_1 - m_0\|^2 + \mathsf{d}_S(\bar\mu_0, \bar\mu_1)^2.$$

In particular, uniform translations are geodesics, as in the case of classical optimal transport.

Some special cases. Section 6 covers two examples where we can explicitly compute the metric tensor. We start with Gaussians: it is one of the few examples in optimal transport for which there are closed formulas and for which the geometry is well understood [73]; it is also one of the paradigmatic examples in information geometry which relates Gaussian measures to the Poincaré half-plane, see e.g. [23]. Building on an explicit formula for OT_ε between Gaussians from [49], Theorem 6.1 shows that

$$\mathbf{g}_v(\dot{v},\dot{v}) = \frac{1}{4\sqrt{\frac{\varepsilon^2}{16}+v^2}}\dot{v}^2.$$

Here we identify a one-dimensional Gaussian measure $\mathcal{N}(0,v)$ with its variance v. Theorem 6.2 then provides a formula for the geodesic restriction of d_S to Gaussian measures. Note that strictly speaking we are not in our framework since Gaussians are not supported on a compact space. As a second example, we consider the two-point space $X=\{x_1,x_2\}$, where we can compute the metric tensor. In particular, in Proposition 6.3 we study the asymptotic of the distance in two different regimes, namely $c(x_1,x_2)\ll\varepsilon$ and $c(x_1,x_2)\gg\varepsilon$. The formula works for the squared distance cost function and asymptotically recovers the cost of horizontal movement on $\mathbb{R}^d$ as $\|x_1-x_2\|\to 0$.

Negative results. Section 7 collects some negative results on the Sinkhorn divergence and the metric tensor that have been mentioned throughout the introduction. Specifically, we provide examples showing that $\sqrt{S_\varepsilon}$ does not satisfy the triangle inequality, that $(\mu,\nu)\mapsto S_\varepsilon(\mu,\nu)$ is not jointly convex, and that the metric tensor $\mathbf{g}_\mu(\dot{\mu},\dot{\mu})$ is also not jointly convex in μ and $\dot{\mu}$.

Comparison with Schrödinger bridges. We emphasize that our geometry and the corresponding geodesics are fundamentally different from Schrödinger bridges, a notion of displacement interpolation associated with entropic optimal transport [32,45]. In our construction, given two probability measures μ_0, μ_1, and $\varepsilon>0$ there exists $(\mu_t^{\mathrm{geo},\varepsilon})_{t\in[0,1]}$ which is a geodesic in our metric d_S, i.e. a minimizer in (1.8). In contrast, when $c=\|x-y\|^2$ is the quadratic cost on $\mathbb{R}^d$, the Schrödinger bridge $(\mu_t^{\mathrm{Sbr},\varepsilon})_{t\in[0,1]}$ is constructed as follows [45]: Let π_ε be the optimal transport plan, that is, the minimizer in (1.5). Let $q_{t,\varepsilon}(x,y)$ denote the law of the Brownian bridge at time t starting at $t=0$ in x and ending at $t=1$ in y with diffusivity $\sqrt{\varepsilon/2}$. That is, $q_{t,\varepsilon}(x,y)\sim\mathcal{N}((1-t)x+ty,t(1-t)\varepsilon/2)$. Then $\mu_t^{\mathrm{Sbr},\varepsilon}$ is given by

$$\mu_t^{\mathrm{Sbr},\varepsilon} = \iint_{X\times X} q_{t,\varepsilon}(x,y)\,\mathrm{d}\pi_\varepsilon(x,y).$$

It is also a curve which satisfies $\mu_0^{\mathrm{Sbr},\varepsilon}=\mu_0$, $\mu_1^{\mathrm{Sbr},\varepsilon}=\mu_1$ and which interpolates smoothly in between. Nevertheless, the two notions of interpolation are conceptually very different. Let us point out three major differences between $\mu^{\mathrm{geo},\varepsilon}$ and $\mu^{\mathrm{Sbr},\varepsilon}$:

(1) The geodesic interpolation $\mu^{\mathrm{geo},\varepsilon}$ is defined on any compact space X if the continuous cost function c defines a universal positive definite kernel $k_c=\exp(-c/\varepsilon)$,

whereas Schrödinger bridges need the quadratic cost on $\mathbb{R}^d$, or more generally a reference continuous-time Markov process (the Brownian motion in the case of quadratic cost) to be defined [45].

(2) If $\mu_0 = \mu_1 = \mu$, then the geodesic interpolation $\mu^{\mathrm{geo},\varepsilon}$ is constant in time and coincides with μ. In the same case, $\mu^{\mathrm{Sbr},\varepsilon}$ is *not* constant in time, and typically has larger variance for intermediate $t \in (0,1)$ than for $t = 0, 1$. This is linked to the phenomenon that generically $\mathrm{OT}_\varepsilon(\mu,\mu) > 0$; see also the discussion in [46], after Theorem 1.7.

(3) In both cases, for the quadratic cost the parameter $\sqrt{\varepsilon}$ acts as a typical length. In the case of Schrödinger bridges, it is also related to a typical time, whereas for geodesics there is no typical time. To understand this, for $\tau \in (0,1)$, let us look at the curve $(\mu_t)_{t\in[0,\tau]}$ but rescaled to be over $[0,1]$. That is, let us consider $(\mu^{\mathrm{geo},\varepsilon}_{t\tau})_{t\in[0,1]}$ and $(\mu^{\mathrm{Sbr},\varepsilon}_{t\tau})_{t\in[0,1]}$. Then $(\mu^{\mathrm{geo},\varepsilon}_{t\tau})_{t\in[0,1]}$ is a geodesic from μ_0 to μ_τ for any value of $\tau \in (0,1)$, as usual in Riemannian geometry. On the other hand, $(\mu^{\mathrm{Sbr},\varepsilon}_{t\tau})_{t\in[0,1]}$ is a Schrödinger bridge between its end points $\mu^{\mathrm{Sbr},\varepsilon}_0$ and $\mu^{\mathrm{Sbr},\varepsilon}_\tau$, but with a regularization parameter $\varepsilon\tau$. This can be seen by using that the Schrödinger bridge is the marginal of a process which minimizes an entropy over the space of measures on paths [45], and to use the concatenation property of the entropy for Markov processes in this space [7, Lem. 3.4]. The key insight is that, when rescaled from time $[0,\tau]$ to time $[0,1]$, a Brownian motion with diffusivity $\sqrt{\varepsilon/2}$ becomes a Brownian motion with diffusivity $\sqrt{\tau\varepsilon/2}$. So in particular as $\tau \to 0$, that is, as we "zoom in" in time on a Schrödinger bridge, it increasingly resembles a Schrödinger bridge with $\varepsilon = 0$, that is, a Wasserstein geodesic [44, 54].

We also refer ahead to Figure 1 with Gaussian measures where we can do explicit computations for our geodesics and Schrödinger bridges.

Setting and notation. We consider a compact metric space (X, d). We fix $\varepsilon > 0$ and a symmetric, continuous, non-negative cost function $c \in \mathcal{C}(X \times X; [0,\infty))$ that induces a positive definite universal kernel $k_c(x,y) = \exp(-c(x,y)/\varepsilon)$.

We denote the dual pairing between a space and its dual space by $\langle\cdot,\cdot\rangle$ and write $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ for the inner product on a Hilbert space $\mathcal{H}$.

We work with the space of signed Radon measures $\mathcal{M}(X)$, which is identified with the dual space of $\mathcal{C}(X)$ through the Riesz–Markov–Kakutani theorem. We denote the subset of non-negative measures by $\mathcal{M}_+(X)$, the space of "balanced measures" $\nu \in \mathcal{M}(X)$ with $\int \mathbb{1}_X \,\mathrm{d}\nu = 0$ by $\mathcal{M}_0(X)$, and Radon probability measures by $\mathcal{P}(X)$.

When X is the closure of a bounded open set in $\mathbb{R}^d$ we write $\mathcal{C}^m(X)$ for the space of m times differentiable functions over X. This space is endowed with the norm $\|\phi\|_{\mathcal{C}^m} = \sup_{|\alpha|\le m}\|\partial^\alpha\phi\|_\infty$ of uniform convergence of the function and all its (partial) derivatives up to order m, denoted using a multi-index $\alpha \in \mathbb{N}_0^d$. We denote by $\mathcal{C}^{(m,m)}(X,X)$ the space of functions for which mixed partial derivatives up to order m in the first and second variables separately are continuous. With this notation, $\mathcal{C}^0(X) = \mathcal{C}(X)$ and $\mathcal{C}^{(0,0)}(X) = \mathcal{C}(X \times X)$. We define $\mathcal{C}^m(X)^*$ as the dual space of $\mathcal{C}^m(X)$. It is endowed with the dual norm $\|\cdot\|_{\mathcal{C}^{m},*}$ defined as $\|\nu\|_{\mathcal{C}^m,*} = \sup\{\langle\nu,\phi\rangle \mid \|\phi\|_{\mathcal{C}^m} \le 1\}$. In addition to the strong topology on $\mathcal{C}^m(X)^*$ coming from the dual norm $\|\cdot\|_{\mathcal{C}^{m,*}}$, we can also endow it with the weak-$*$ topology. As X is compact, note that $\mathcal{C}^0(X)^* = \mathcal{M}(X)$ and weak-$*$ convergence in $\mathcal{C}^0(X)^*$ is the usual weak-$*$ convergence of measures. In this way, setting $m = 0$ throughout this paper recovers the assumption X compact and c is continuous.

For Landau big-O and little-o notation we use $\mathcal{O}$ and o respectively.

2. Preliminaries on entropic optimal transport

The entropic optimal transport problem, as defined in (1.5), is a constrained convex optimization problem whose dual formulation reads

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

Here the notation $f \oplus g$ stands for the function $(x,y) \mapsto f(x) + g(y)$ defined on $X \times X$. The dual problem admits maximizers $f_{\mu,\nu}$, $g_{\mu,\nu}$, which we call *Schrödinger potentials*, that satisfy

$$\mathrm{OT}_\varepsilon(\mu,\nu) = \int_X f_{\mu,\nu}\,\mathrm{d}\mu + \int_X g_{\mu,\nu}\,\mathrm{d}\nu \tag{2.2}$$

and are solutions of the Schrödinger system

$$\begin{cases} f_{\mu,\nu} = T_\varepsilon(g_{\mu,\nu},\nu), \\ g_{\mu,\nu} = T_\varepsilon(f_{\mu,\nu},\mu), \end{cases} \tag{2.3}$$

where $T_\varepsilon\colon C(X)\times\mathcal{P}(X)\to C(X)$ is defined by

$$T_\varepsilon(f,\mu)(y) := -\varepsilon\log\int_X \exp\Big(\frac{1}{\varepsilon}(f(x) - c(x,y))\Big)\,\mathrm{d}\mu(x). \tag{2.4}$$

We will sometimes write the condition $g_{\mu,\nu} = T_\varepsilon(f_{\mu,\nu},\mu)$ as

$$\int_X \exp\Big(\frac{1}{\varepsilon}(f_{\mu,\nu}(x) + g_{\mu,\nu}(y) - c(x,y))\Big)\,\mathrm{d}\mu(x) = 1 \tag{2.5}$$

for every $y \in X$. The Schrödinger potentials are unique in the space $\mathcal{C}(X)\times\mathcal{C}(X)/\sim$ with the relation $(f,g)\sim(f+\lambda\mathbb{1}_X, g-\lambda\mathbb{1}_X)$ for all $\lambda\in\mathbb{R}$. Alternatively, we can select a canonical representative by setting $f_{\mu,\nu}(x_0) = g_{\mu,\nu}(x_0)$ for some fixed $x_0 \in X$, which we do throughout the paper. This choice yields $f_{\mu,\nu} = g_{\nu,\mu}$ and, in particular, throughout this article $f_{\mu,\mu} = g_{\mu,\mu}$. The optimal potentials inherit the modulus of continuity from the cost function c [57, Lem. 3.1]. The optimal transport plan π in (1.5) is given by

$$\pi = \exp\Big(\frac{1}{\varepsilon}(f_{\mu,\nu}\oplus g_{\mu,\nu} - c)\Big)(\mu\otimes\nu). \tag{2.6}$$

Note that the function $f_{\mu,\nu}\oplus g_{\mu,\nu}$ does not depend on the choice of the representatives since constant shifts cancel out. Optimal $(f_{\mu,\nu}, g_{\mu,\nu})$ can be obtained numerically by iterating the fixed point equation (2.3), which corresponds to the celebrated Sinkhorn algorithm. We refer to [61] for an overview on computational optimal transport, and to [45, 56] for a detailed discussion on entropic optimal transport and related literature.

3. The Hessian of the Sinkhorn divergence

3.1. Statement of the expansion

Theorem 1.1 shows that the Sinkhorn divergence (1.6) is positive definite and convex in each of its input variables. However, as we show in Section 7.1, S_ε, or more precisely $\sqrt{S_\varepsilon}$, in general does not satisfy the triangle inequality. To construct the metric d_S announced in (1.8) in the introduction, we first characterize how S_ε behaves close to the diagonal $\mu = \nu$. Specifically, we want to find an expansion of the form

$$S_\varepsilon(\mu, \mu_t) \sim t^2 \mathbf{g}_\mu(\dot{\mu}, \dot{\mu})$$

as $t \to 0$ for a suitably differentiable path $(\mu_t)_t$ in $\mathcal{P}(X)$ with $\mu_0 = \mu$, $\dot{\mu}_0 = \dot{\mu}$. The term $\mathbf{g}_\mu(\dot{\mu}, \dot{\mu})$ will then serve as the metric tensor and it will be examined in detail in Section 4. By Theorem 1.1 there is good hope that $\mathbf{g}_\mu$ will be a positive quadratic form on a meaningful tangent space to probability measures. The appropriate notion of tangent space will become apparent in Section 4.

Though in our general setting X is a compact space, in this section we will also consider a differentiable setting in which stronger results can be obtained.

Assumption 1 (Differentiable setting). We fix an integer $m \geq 0$. If $m = 0$, we simply assume, as stated above, that X is a compact metrizable space and $c \in \mathcal{C}(X \times X)$ is a continuous function. If $m \geq 1$ we assume that X is the closure of a bounded open set in $\mathbb{R}^d$ and that $c \in \mathcal{C}^{(m,m)}(X \times X)$.

Definition 3.1 ($\mathcal{C}^m$-perturbations). Let $m \geq 0$ and $\tau > 0$. We call a path $(\mu_t)_{t \in (-\tau,\tau)}$ valued in $\mathcal{P}(X)$ a $\mathcal{C}^m$-perturbation of $\mu = \mu_0$ if it is continuously differentiable in the space $\mathcal{C}^m(X)^*$ endowed with the weak-$*$ topology. We write $\dot{\mu} \in \mathcal{C}^m(X)^*$ for the derivative of the curve at time $t = 0$.

As $\mathcal{C}^m(X)^*$ is endowed with its weak-$*$ topology, it means that for any $\phi \in \mathcal{C}^m(X)$ the map $t \mapsto \langle \mu_t, \phi \rangle$ is differentiable, that the derivative, denoted by $\langle \dot{\mu}_t, \phi \rangle$, is linear in ϕ, and that the mapping $(\phi, t) \mapsto \langle \dot{\mu}_t, \phi \rangle$ is jointly continuous from $\mathcal{C}^m(X) \times (-\tau, \tau)$ to $\mathbb{R}$ (see Lemma A.2).

Remark 3.2 (Vertical and horizontal perturbations). For any compact set X, a balanced measure $\nu \in \mathcal{M}_0(X)$ induces a $\mathcal{C}^0$-perturbation of μ in the form of $\mu_t = \mu + t\nu$ if and only if $\nu \ll \mu$ and the Radon–Nikodym derivative $\frac{\mathrm{d}\nu}{\mathrm{d}\mu}$ is uniformly bounded from above and below. We call such perturbations "vertical perturbations" as they correspond to mass disappearing somewhere to be recreated elsewhere. The constraint $\langle \nu, \mathbb{1}_X \rangle = 0$ in the space of balanced measures $\mathcal{M}_0(X)$ comes from the preservation of mass.

When X is the closure of a bounded open set in $\mathbb{R}^d$, we call "horizontal perturbations" curves of the form

$$\mu_t = \Psi(t, \cdot)_{\#}\mu$$

for a continuous function $\Psi: (-\tau, \tau) \times X \to X$ with one continuous time derivative. It defines a $\mathcal{C}^1$-perturbation and $\dot{\mu}_t = -\operatorname{div}(\mu_t v_t)$ with $v_t = \frac{\partial \Psi}{\partial t}(t, \cdot)$. Indeed, for any $\phi \in \mathcal{C}^1(X)$ we have

$$\langle \dot{\mu}_t, \phi \rangle = \int_X \frac{\partial \Psi}{\partial t}(t, x) \cdot \nabla \phi(x) \, \mathrm{d}\mu_t(x) = \int_X \frac{\partial \Psi}{\partial t}(t, \Psi(t, x)) \cdot \nabla \phi(\Psi(t, x)) \, \mathrm{d}\mu_t(x),$$

and the latter is jointly continuous in (ϕ, t) with respect to the norm topology of $\mathcal{C}^1(X)$.

Before stating the main result of this section, we need to introduce integral operators related to the transport kernel. We define the self-transport kernel as follows:

$$k_\mu(x, y) := \exp\Big(\frac{1}{\varepsilon}\Big(f_{\mu,\mu}(x) + f_{\mu,\mu}(y) - c(x, y)\Big)\Big).$$

Recall here that we adopt the convention $f_{\mu,\mu} = g_{\mu,\mu}$ so that k_μ is unambiguous.

Since we assume that $c \in \mathcal{C}^{(m,m)}(X \times X)$, the smoothing effect of the Schrödinger system (2.3) guarantees that the potentials $f_{\mu,\nu}$, $g_{\mu,\nu}$ belong to $\mathcal{C}^m(X)$ (see also Proposition 3.7 below). Thus k_μ has the same regularity class as the cost function c, and it is enough to guarantee that the following operators are well defined (see also Proposition 3.6 below).

Definition 3.3. For $\mu \in \mathcal{P}(X)$ we define the kernel operators K_μ, H_μ by

$$\begin{aligned} &K_\mu: \mathcal{C}(X) \to \mathcal{C}^m(X), && K_\mu[\phi](y) = \int_X \phi(x) k_\mu(x, y) \, \mathrm{d}\mu(x), \\ &H_\mu: \mathcal{C}^m(X)^* \to \mathcal{C}^m(X), && H_\mu[\nu](y) = \langle \nu, k_\mu(\cdot, y) \rangle. \end{aligned}$$

Analogously, we define $H_c: \mathcal{C}^m(X)^* \to \mathcal{C}^m(X)$, $H_c[\nu](y) = \langle \nu, k_c(\cdot, y) \rangle$.

The operator K_μ can be extended to the domain $L^1(X, \mu)$. Note that $H_\mu[\nu](y) = \int_X k_\mu(x, y) \, \mathrm{d}\nu(x)$ when ν is a measure, thus if $\phi \in L^1(X, \mu)$, then $K_\mu[\phi] = H_\mu[\phi\mu]$. The operator H_μ is the kernel mean embedding into the reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mu$ induced by k_μ, and its domain can be extended to $\mathcal{H}_\mu^*$. This will be discussed in Section 4.

Let us call $\mathcal{C}^m(X)_0^*$ the space of balanced distributions, that is, the space of all $\nu \in \mathcal{C}^m(X)^*$ satisfying $\langle \nu, \mathbb{1}_X \rangle = 0$. It is dual to $\mathcal{C}^m(X)/\mathbb{R}$, the space of $\mathcal{C}^m$ functions identified up to shifts by constant functions (identified with $\mathbb{R}$). Hence the pairing $\langle \nu, \phi \rangle$ is well defined between $\nu \in \mathcal{C}^m(X)_0^*$ and $\phi \in \mathcal{C}^m(X)/\mathbb{R}$. Since we regard paths valued in $\mathcal{P}(X)$, all perturbations in Definition 3.1 are of this form. The operator K_μ descends to a well-defined operator

$$K_\mu: \mathcal{C}(X)/\mathbb{R} \to \mathcal{C}^m(X)/\mathbb{R},$$

because $K_\mu[\mathbb{1}_X] = \mathbb{1}_X$ by (2.5). Using the canonical projection $\mathcal{C}^m(X) \to \mathcal{C}^m(X)/\mathbb{R}$ we can always move to the quotient space.

Theorem 3.4 (Hessian of the Sinkhorn divergence). *With Assumption 1 for some $m \geq 0$, let $\mu \in \mathcal{P}(X)$ and $(\mu_t)_t$ be a $\mathcal{C}^m$-perturbation of μ. Then*

$$\lim_{t\to 0} \frac{S_\varepsilon(\mu,\mu_t)}{t^2} = \frac{\varepsilon}{2}\langle \dot{\mu}, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}]\rangle. \tag{3.1}$$

The expression on the right-hand side is well defined, in the sense that $(\mathrm{id} - K_\mu^2)^{-1}$ exists as an operator on $\mathcal{C}^m(X)/\mathbb{R}$ and thus the duality pairing is understood between this space and balanced measures or distributions. In Section 4 we will take the formula on the right-hand side of (3.1) as a starting point and analyze for which class of perturbations $\dot{\mu}$ we can make sense of it.

Remark 3.5 (Discrete measures with finite support). When $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ is a discrete measure with $a_i > 0$ and the vertical perturbation reads $\nu = \sum_{i=1}^n b_i \delta_{x_i}$ with $\sum_i b_i = 0$, the formula can be expressed as follows. With $f = f_{\mu,\mu}$, we introduce the $n \times n$ matrices K, H as

$$\begin{aligned}
K_{ij} &= \exp\Big(\frac{1}{\varepsilon}(f(x_i) + f(x_j) - c(x_i,x_j))\Big)a_j,\\
H_{ij} &= \exp\Big(\frac{1}{\varepsilon}(f(x_i) + f(x_j) - c(x_i,x_j))\Big)
\end{aligned}$$

for indices $i, j \in \{1,\dots,n\}$. With $b = (b_i)_i \in \mathbb{R}^n$, Theorem 3.4 reads

$$\lim_{t\to 0} \frac{S_\varepsilon(\mu,\mu_t)}{t^2} = \frac{\varepsilon}{2} b^\top (\mathrm{id} - K^2)^\dagger H b,$$

where $(\mathrm{id} - K^2)^\dagger$ denotes the Moore–Penrose pseudo-inverse of the matrix $\mathrm{id} - K^2$.

The proof of Theorem 3.4 relies on the differentiability of the Schrödinger potentials along the path $(\mu_t)_t$ and on explicit formulas for the derivatives.

For the technical arguments we follow the line of the works [16, 18] which use the implicit function theorem to compute these derivatives, though these articles are in the more involved context of a multi-marginal optimal transport problem and with different regularity assumptions on the path of measures $(\mu_t)_t$.

We first prove preliminary results on the operators K_μ and H_μ before moving to the core of our strategy and arguments.

3.2. Preliminary results on the operators K_μ and H_μ

In this section we analyze the operators K_μ and H_μ which will appear naturally when differentiating the fixed point equation (2.3) that characterizes the Schrödinger potentials. We then compute the derivatives of the Schrödinger potentials, and finally prove Theorem 3.4 in Section 3.3.

We still fix $m \geq 0$ and work in the setting of Assumption 1. We start by qualitative results on the operators H_μ, K_μ and the Schrödinger potentials.

Proposition 3.6. *Fix $\mu \in \mathcal{P}(X)$. The operators $H_\mu, H_c : \mathcal{C}^m(X)^* \to \mathcal{C}^m(X)$ are weak-$*$-to-norm continuous, as well as norm-to-norm continuous and compact. In particular, the operator $K_\mu : \mathcal{C}(X) \to \mathcal{C}^m(X)$ is norm-to-norm continuous and compact.*

Proof. The kernels k_c and k_μ belong to $\mathcal{C}^{(m,m)}(X)$. The results for H_c, H_μ directly follow by standard arguments on integral kernel operators; see Lemma A.3 in the appendix. As $K_\mu[\phi] = H_\mu[\phi\mu]$, continuity and compactness extend to K_μ as well. ■

Proposition 3.7. *The map $(\mu, \nu) \mapsto f_{\mu,\nu}$ from $\mathcal{P}(X)^2$ to $\mathcal{C}^m(X)$ is continuous and the set $\{f_{\mu,\nu} \mid \mu, \nu \in \mathcal{P}(X)\}$ is compact in $\mathcal{C}^m(X)$.*

By symmetry the same results holds for $(\mu, \nu) \mapsto g_{\mu,\nu}$.

Proof of Proposition 3.7. From continuity of $(\mu, \nu) \mapsto (f_{\mu,\nu}, g_{\mu,\nu})$ in $\mathcal{C}(X)^2$ [58, Prop. B.1], we see that the map $(\mu, \nu) \to g_{\mu,\nu}\nu \in \mathcal{M}(X)$ is weak-$*$-to-weak-$*$ continuous. As $\mathcal{M}(X) = \mathcal{C}^0(X)^*$, we can apply Lemma A.3 and deduce that $(\mu, \nu) \mapsto H_c[g_{\mu,\nu}\nu] = \exp(-f_{\mu,\nu}/\varepsilon)$ is continuous into $\mathcal{C}^m(X)$, and thus continuity as in our claim follows by taking the logarithm. The set $\{f_{\mu,\nu} \mid \mu, \nu \in \mathcal{P}(X)\}$ is compact as a continuous image of a compact set. ■

The following result is a key quantitative estimate which holds uniformly over $\mathcal{P}(X)$. We recall that $\mathcal{C}(X)/\mathbb{R}$ is the quotient space of continuous functions by constant functions. We equip this space with the quotient norm

$$\|\phi\|_{\mathcal{C}/\mathbb{R}} := \inf_{\lambda \in \mathbb{R}} \|\phi - \lambda \cdot \mathbb{1}_X\|_\infty = \frac{1}{2}\Big(\sup_{x \in X} \phi(x) - \inf_{y \in X} \phi(y)\Big).$$

Proposition 3.8 (Contraction property). *Let $q := 1 - \exp(-4\|c\|_\infty/\varepsilon) \in (0, 1)$. Then we get $\|K_\mu[\phi]\|_{\mathcal{C}/\mathbb{R}} \leq q \cdot \|\phi\|_{\mathcal{C}/\mathbb{R}}$ for all $\phi \in \mathcal{C}(X)$ and all $\mu \in \mathcal{P}(X)$.*

Proof. By [26, Thm. 1.2] the Schrödinger potentials $f_{\mu,\mu} \in \mathcal{C}(X)$ satisfy the constraint $\|f_{\mu,\mu}\|_\infty \leq \frac{3}{2}\|c\|_\infty$. With $a = \exp(-4\|c\|_\infty/\varepsilon)$ and using $c \geq 0$, we see that $k_\mu(y, z) \geq a \cdot k_\mu(x, z)$ for any triplet $x, y, z \in X$. For any $x, y \in X$, this, together with (2.5), leads to

$$\begin{aligned}
&K_\mu[\phi](x) - K_\mu[\phi](y) \\
&\quad = \int_{\{\phi>0\}} \phi(z)(k_\mu(x, z) - k_\mu(y, z))\, \mathrm{d}\mu(z) \\
&\qquad + \int_{\{\phi<0\}} -\phi(z)(-k_\mu(x, z) + k_\mu(y, z))\, \mathrm{d}\mu(z) \\
&\quad \leq \int_{\{\phi>0\}} \phi(z)(1 - a)k_\mu(x, z)\, \mathrm{d}\mu(z) + \int_{\{\phi<0\}} -\phi(z)(1 - a)k_\mu(y, z)\, \mathrm{d}\mu(z) \\
&\quad \leq 2(1 - a) \cdot \|\phi\|_\infty.
\end{aligned}$$

The operator norm estimate with respect to the quotient norm $\|.\|_{\mathcal{C}/\mathbb{R}}$ on domain and codomain follows by taking a supremum over x and y and considering the infimum over shifts $\phi - \lambda$ by constants $\lambda \in \mathbb{R}$. Finally, $q = 1 - a$. ∎

From this estimate we prove that the inverse of the map $\mathrm{id} - K_\mu^2$, central in the expansion of the Sinkhorn divergence, actually exists. We use $\mathbb{B}(\mathbb{X})$ to denote the set of bounded linear operators on a normed space $\mathbb{X}$.

Theorem 3.9. *Let $\mu \in \mathcal{P}(X)$. Then $(\mathrm{id} - K_\mu^2)^{-1}$ exists in $\mathbb{B}(\mathcal{C}^m(X)/\mathbb{R})$ and $(\mathrm{id} + K_\mu)^{-1}$ exists in $\mathbb{B}(\mathcal{C}^m(X))$.*

The result will be strengthened in Theorem 4.3 in the next section.

Proof of Theorem 3.9. The operators $\mathrm{id} - K_\mu^2$ and $\mathrm{id} + K_\mu$ are compact perturbations of the identity on, respectively, the Banach spaces $\mathcal{C}^m(X)/\mathbb{R}$ and $\mathcal{C}^m(X)$. To check that they are invertible, by the Fredholm alternative [14, Thm. 6.6] it is enough to verify that the equations $\phi = K_\mu^2[\phi]$ and $\phi = -K_\mu[\phi]$ only have the trivial solution $\phi = 0$. In both cases, Proposition 3.8 directly yields that $\|\phi\|_{\mathcal{C}/\mathbb{R}} = 0$. In the case $\phi = K_\mu^2[\phi]$, this gives $\phi = 0$ in $\mathcal{C}(X)/\mathbb{R}$. For the equation $\phi = -K_\mu[\phi]$, this yields that ϕ is constant, and from $K_\mu[\mathbb{1}_X] = \mathbb{1}_X$ we obtain $\phi = 0$. ∎

Remark 3.10 (A probabilistic interpretation). Note that K_μ can be interpreted as follows. For $\mu \in \mathcal{P}(X)$, we can consider a Markov chain on X with equilibrium measure μ, and whose transition kernel is given by $x \mapsto k_\mu(x, y)\,\mathrm{d}\mu(y)$. The operator K_μ is simply the dual Markov operator of this chain. Indeed, the entropic optimal transport coupling π between μ and itself, given by (2.6), is the law of the pair of random variables (x_1, x_2) following this Markov chain if the initial distribution is μ.

As k_μ is symmetric, we see that the Markov chain is reversible. The spectrum of the operator K_μ is studied in more detail in Proposition 4.2 which builds on Proposition 3.8: all the eigenvalues belong to $[0, 1]$, and the spectral gap of $\mathrm{id} - K_\mu$ is at least $\exp(-4\|c\|_\infty/\varepsilon)$.

3.3. Proof of the expansion

We are now ready for the proof of Theorem 3.4, again under Assumption 1. The starting point is this elementary lemma from calculus.

Lemma 3.11. *Assume that F, defined in a neighborhood of $(0,0) \in \mathbb{R}^2$, is of class $\mathcal{C}^{(1,1)}$. Then*

$$\lim_{t \to 0} \frac{F(0,t) + F(t,0) - F(0,0) - F(t,t)}{2} = -\frac{1}{2}\frac{\partial^2 F}{\partial t\, \partial s}(0,0).$$

Proof. By writing $F(0,t) + F(t,0) - F(0,0) - F(t,t) = F(t,0) - F(0,0) - (F(t,t) - F(t,0))$ and integrating first with respect to the second variable and then the first one,

$$F(0,t) + F(t,0) - F(0,0) - F(t,t) = -\int_0^t \int_0^t \partial_1 \partial_2 F(\tau, \sigma)\,\mathrm{d}\sigma\,\mathrm{d}\tau.$$

The conclusion follows by continuity of $\partial_1 \partial_2 F$. ∎

We will apply this lemma to $F(t,s) := \mathrm{OT}_\varepsilon(\mu_t,\mu_s)$. Indeed, as this function is symmetric we have

$$S_\varepsilon(\mu,\mu_t) = \frac{F(0,t)+F(t,0)-F(0,0)-F(t,t)}{2}.$$

So provided we prove that F is of class $\mathcal{C}^{(1,1)}$ around $(0,0)$, Lemma 3.11 yields

$$\lim_{t\to 0}\frac{S_\varepsilon(\mu,\mu_t)}{t^2} = -\frac{1}{2}\frac{\partial^2 F}{\partial t\,\partial s}(0,0). \tag{3.2}$$

Thus to prove our theorem we need to prove $F \in \mathcal{C}^{(1,1)}$ and differentiate it twice. Taking one derivative of F is standard [28]; we provide the following lemma adapted to our context.

Lemma 3.12. *If the paths $(\mu_t)_{t\in(-\tau,\tau)}$, $(\nu_t)_{t\in(-\tau,\tau)}$ are $\mathcal{C}^m$-perturbations of $\mu,\nu\in\mathcal{P}(X)$, then*

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}\mathrm{OT}_\varepsilon(\mu_t,\nu_t) = \langle\dot\mu, f_{\mu,\nu}\rangle + \langle\dot\nu, g_{\mu,\nu}\rangle.$$

Proof. We follow the proof strategy of [28, Prop. 2]. As seen there (replacing $\delta\mu$ by $(\mu_t-\mu)/t$), taking $f_{\mu_t,\nu_t}, g_{\mu_t,\nu_t}$ as a suboptimal pair in the dual formulation (2.1) of $\mathrm{OT}_\varepsilon(\mu,\nu)$ and likewise $f_{\mu,\nu}, g_{\mu,\nu}$ for $\mathrm{OT}_\varepsilon(\mu_t,\nu_t)$ yields

$$\begin{aligned}\left\langle\frac{\mu_t-\mu}{t}, f_{\mu,\nu}\right\rangle + \left\langle\frac{\nu_t-\nu}{t}, g_{\mu,\nu}\right\rangle + o(1) &\le \frac{1}{t}(\mathrm{OT}_\varepsilon(\mu_t,\nu_t)-\mathrm{OT}_\varepsilon(\mu,\nu))\\ &\le \left\langle\frac{\mu_t-\mu}{t}, f_{\mu_t,\nu_t}\right\rangle + \left\langle\frac{\nu_t-\nu}{t}, g_{\mu_t,\nu_t}\right\rangle + o(1).\end{aligned}$$

Taking the limit $t\to 0$ gives the result. Indeed, as $t\to 0$, $(f_{\mu_t,\nu_t}, g_{\mu_t,\nu_t})$ converges to $(f_{\mu,\nu}, g_{\mu,\nu})$ in $(\mathcal{C}^m(X)/\mathbb{R})^2$ (Proposition 3.7), while $(\mu_t-\mu)/t$ and $(\nu_t-\nu)/t$ converge weak-$*$ to $\dot\mu$ and $\dot\nu$ in $\mathcal{C}^m(X)^*_0$. ∎

From this lemma we obtain $\frac{\partial F}{\partial t}(t,s) = \langle\dot\mu_t, f_{\mu_t,\mu_s}\rangle$. To compute the second derivative of F, we will need to differentiate the Schrödinger potentials with respect to the input measures, as announced. As we will use the fixed point equation (2.3) defining the optimal pair of potentials $(f_{\mu_t,\mu_s}, g_{\mu_t,\mu_s})$, we first need to understand the derivatives of the operator T_ε featured in it. Whereas T_ε is a priori defined on $\mathcal{C}(X)\times\mathcal{P}(X)$, from $T_\varepsilon(f+\lambda\mathbb{1}_X,\mu) = T_\varepsilon(f,\mu)-\lambda\mathbb{1}_X$ for any f, μ, and $\lambda\in\mathbb{R}$, we see that T_ε descends to an operator $\mathcal{C}(X)/\mathbb{R}\times\mathcal{P}(X)\to\mathcal{C}^m(X)/\mathbb{R}$.

Lemma 3.13. *Let the path $(\mu_t)_{t\in(-\tau,\tau)}$ be a $\mathcal{C}^m$-perturbation. Then the parametrized map $\mathcal{C}^m(X)\times(-\tau,\tau)\to\mathcal{C}^m(X)$, $(f,t)\mapsto T_\varepsilon(f,\mu_t)$ built from (2.4) is continuously differentiable. For $\phi\in\mathcal{C}^m(X)$ it satisfies*

$$D_1T_\varepsilon(f_{\mu,\mu},\mu)[\phi] = -K_\mu[\phi],$$
$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0}T_\varepsilon(f_{\mu,\mu},\mu_t) = -\varepsilon H_\mu[\dot\mu].$$

These formulas remain true when $\mathcal{C}^m(X)$ is replaced by $\mathcal{C}^m(X)/\mathbb{R}$.

Proof. We first show that the mapping is continuously differentiable. It is enough to prove that $U(g,t)(y) = \int_X k_c(x,y)g(x)\,\mathrm{d}\mu_t(x)$ is continuously differentiable as a map $\mathcal{C}^m(X) \times (-\tau,\tau) \to \mathcal{C}^m(X)$ as $T_\varepsilon(f,t) = -\varepsilon \log U(\exp(f/\varepsilon),t)$. The mapping U is linear in g and continuous (thus continuously differentiable) in g by Lemma A.3. Regarding the variable t, for any given $g \in \mathcal{C}^m(X)$ and $y \in X$ and any $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le m$ we have $\partial^\alpha \frac{\mathrm{d}}{\mathrm{d}t} U(g,t)(y) = \langle \dot\mu_t, \partial_y^\alpha k_c(\cdot,y) g\rangle$ as $\partial_y^\alpha k_c(\cdot,y)g \in \mathcal{C}^m(X)$. Using the regularity of $\dot\mu_t$ given by Definition 3.1, this expression is continuous in y, g, and t. If $g_n \to g$, Lemma A.1 shows that the convergence of $\partial^\alpha \frac{\mathrm{d}}{\mathrm{d}t} U(g_n,t)(y)$ to $\partial^\alpha \frac{\mathrm{d}}{\mathrm{d}t} U(g,t)(y)$ is uniform in y. Thus we have differentiability of U in its second variable with jointly continuous derivative. The differentiability of T_ε follows.

Next we establish the expressions for the derivatives at $(f_{\mu,\mu},0)$. By the chain rule we have easily

$$D_1 T_\varepsilon(f,\mu)[\phi](y) = -\varepsilon \frac{\int \frac{1}{\varepsilon}\phi(x)\exp\big(\frac{1}{\varepsilon}(f(x)-c(x,y))\big)\,\mathrm{d}\mu(x)}{\int \exp\big(\frac{1}{\varepsilon}(f(x)-c(x,y))\big)\,\mathrm{d}\mu(x)}.$$

As $T_\varepsilon(f_{\mu,\mu},\mu) = f_{\mu,\mu}$, at $f = f_{\mu,\mu}$ the denominator is equal to $\exp(-\frac{1}{\varepsilon} f_{\mu,\mu}(y))$. Thus

$$\begin{aligned} D_1 T_\varepsilon(f_{\mu,\mu},\mu)[\phi](y) &= -\int_X \phi(x)\exp\Big(\frac{1}{\varepsilon}(f_{\mu,\mu}(y) + f_{\mu,\mu}(x) - c(x,y))\Big)\,\mathrm{d}\mu(x) \\ &= -K_\mu[\phi](y). \end{aligned}$$

For the derivative in the second component, a similar computation yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} T_\varepsilon(f_{\mu,\mu},\mu_t)(y) = -\varepsilon \frac{\big\langle \dot\mu, \exp\big(\frac{1}{\varepsilon}(f_{\mu,\mu} - c(\cdot,y))\big)\big\rangle}{\big\langle \mu, \exp\big(\frac{1}{\varepsilon}(f_{\mu,\mu} - c(\cdot,y))\big)\big\rangle}.$$

From the same argument as before, the denominator is equal to $\exp(-\frac{1}{\varepsilon} f_{\mu,\mu}(y))$. Thus the right-hand side coincides with $-\varepsilon H_\mu[\dot\mu]$. Quotienting by constant functions to extend to the case where $f \in \mathcal{C}^m(X)/\mathbb{R}$ instead of $\mathcal{C}^m(X)$ is straightforward. ∎

Proposition 3.14. *Let the path $(\mu_t)_{t\in(-\tau,\tau)}$ be a $\mathcal{C}^m$-perturbation of $\mu \in \mathcal{P}(X)$. Then the Schrödinger potentials $f_{t,s} = f_{\mu_t,\mu_s}$ are continuously differentiable in $\mathcal{C}^m(X)/\mathbb{R}$ with respect to the time parameters in a neighborhood of $(t,s) = (0,0)$. Moreover, they satisfy*

$$\begin{aligned} \frac{\partial f_{0,s}}{\partial s}\Big|_{s=0} &= -\varepsilon(\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot\mu] \in \mathcal{C}^m(X)/\mathbb{R}, \\ \frac{\partial f_{t,0}}{\partial t}\Big|_{t=0} &= \varepsilon K_\mu(\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot\mu] \in \mathcal{C}^m(X)/\mathbb{R}, \\ \frac{\partial f_{t,t}}{\partial t}\Big|_{t=0} &= -\varepsilon(\mathrm{id} + K_\mu)^{-1} H_\mu[\dot\mu] \in \mathcal{C}^m(X). \end{aligned}$$

Proof. We use that $f_{t,s}$ and $g_{t,s} = f_{s,t}$ are characterized by the fixed point equation (2.3). Specifically, we have

$$f_{t,s} = T_\varepsilon(g_{t,s},\mu_s) = T_\varepsilon(T_\varepsilon(f_{t,s},\mu_t),\mu_s) \tag{3.3}$$

for all t and s. We will use the implicit function theorem [27, Thm. 10.2.1] applied to the map

$$\widetilde{T}: \mathcal{C}^m(X)/\mathbb{R} \times (-\tau,\tau)^2 \to \mathcal{C}^m(X)/\mathbb{R}, \quad (f,t,s) \mapsto \big(f - T_\varepsilon(T_\varepsilon(f,\mu_t),\mu_s)\big).$$

The condition $\widetilde{T}(f,t,s) = 0$ is equivalent to $f = f_{t,s}$ in $\mathcal{C}^m(X)/\mathbb{R}$ by (3.3). The assumptions of [27, Thm. 10.2.1] ask that the map $\widetilde{T}$ is of class C^1 and that $D_1\widetilde{T}$ is invertible at the point $(t,s) = (0,0)$, $f = f_{0,0}$. The former is a direct consequence of Lemma 3.13, while for the latter we have $D_1\widetilde{T}(f_{0,0},0,0) = \mathrm{id} - K_\mu^2$ by Lemma 3.13 and this is invertible on $\mathcal{C}^m(X)/\mathbb{R}$ by Theorem 3.9.

We can thus apply [27, Thm. 10.2.1], which ensures that the solution $f = f_{t,s}$ solving $\widetilde{T}(f_t,t,s) = 0$ is continuously differentiable in $\mathcal{C}^m(X)/\mathbb{R}$ with respect to the time parameters in a neighborhood of $(t,s) = (0,0)$, and the partial derivatives are given by

$$\begin{aligned}
\frac{\partial f_{t,0}}{\partial t}\Big|_{t=0} &= -(D_1\widetilde{T}(f_{0,0},0))^{-1}\Big(\frac{\partial \widetilde{T}(f_{0,0},t,0)}{\partial t}\Big|_{t=0}\Big),\\
\frac{\partial f_{0,s}}{\partial s}\Big|_{s=0} &= -(D_1\widetilde{T}(f_{0,0},0))^{-1}\Big(\frac{\partial \widetilde{T}(f_{0,0},0,s)}{\partial s}\Big|_{s=0}\Big).
\end{aligned}$$

Following Lemma 3.13 we obtain

$$\frac{\partial \widetilde{T}(f_{0,0},t,0)}{\partial t}\Big|_{t=0} = -\varepsilon K_\mu H_\mu[\dot{\mu}], \qquad \frac{\partial \widetilde{T}(f_{0,0},0,s)}{\partial s}\Big|_{s=0} = \varepsilon H_\mu[\dot{\mu}].$$

This yields the first two derivatives.

To calculate the time derivative of $f_{t,t}$ we may drop the quotient and use the map $\widetilde{T}'(f,t) := f - T_\varepsilon(f,\mu_t)$ defined on $\mathcal{C}^m(X) \times (-\tau,\tau)$ instead of $\widetilde{T}$. Here, $D_1\widetilde{T}'(f_{0,0},0) = \mathrm{id} + K_\mu$ whose invertibility in the space $\mathcal{C}^m(X)$ can be found once again in Theorem 3.9. Analogous arguments to the previous case give the result. ■

Proof of Theorem 3.4. Denote the Schrödinger potentials by $f_{t,s} := f_{\mu_t,\mu_s}$. Define the symmetric function $F(t,s) := \mathrm{OT}_\varepsilon(\mu_t,\mu_s)$. Recall that with this notation, once we prove that $F \in \mathcal{C}^{(1,1)}$, then (3.2) holds by Lemma 3.11. By Lemma 3.12 we have

$$\frac{\partial F}{\partial t}(t,s) = \langle \dot{\mu}_t, f_{t,s}\rangle.$$

The weak-$*$ continuity of $\dot{\mu}_t$ and continuous differentiability of $f_{t,s}$ by the time parameters (Proposition 3.14) ensure that F is of class $\mathcal{C}^{(1,1)}$ around $(0,0)$. Moreover, we see that

$$\frac{\partial^2 F}{\partial t\,\partial s}(0,0) = \Big\langle \dot{\mu}, \frac{\partial f_{0,s}}{\partial s}\Big|_{s=0}\Big\rangle.$$

Note that on the right-hand side we have the derivative of the "first" Schrödinger potential when the "second" measure is changed. From Proposition 3.14 we obtain

$$\frac{\partial f_{0,s}}{\partial s}\Big|_{s=0} = -\varepsilon(\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \in \mathcal{C}^m(X)/\mathbb{R}.$$

We conclude that

$$\lim_{t\to 0} \frac{S_\varepsilon(\mu,\mu_t)}{t^2} = -\frac{1}{2}\frac{\partial^2 F}{\partial t \partial s}(0,0) = -\frac{1}{2}\Big\langle \dot\mu, \frac{\partial f_{0,s}}{\partial s}\Big|_{s=0}\Big\rangle = \frac{\varepsilon}{2}\langle \dot\mu, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot\mu]\rangle. \blacksquare$$

4. The metric tensor and its regularity

In this section we analyze the expression obtained as the second-order approximation of the Sinkhorn divergence in Theorem 3.4. We provide a way to extend the expression to a larger class of perturbations. Recall that $\mathcal{M}_0(X)$ is the set of balanced signed measures $\nu \in \mathcal{M}(X)$, such that $\langle \nu, \mathbb{1}_X\rangle = 0$.

Definition 4.1 (The metric tensor). For $\mu \in \mathcal{P}(X)$ and $\dot\mu_1$, $\dot\mu_2$ in $\mathcal{M}_0(X)$, we define the metric tensor $\mathbf{g}_\mu(\dot\mu_1,\dot\mu_2)$ as

$$\mathbf{g}_\mu(\dot\mu_1,\dot\mu_2) = \frac{\varepsilon}{2}\langle \dot\mu_1, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot\mu_2]\rangle.$$

In the setting of Assumption 1 we saw that this expression can be extended to perturbations $\dot\mu_1$, $\dot\mu_2$ in $\mathcal{C}^m(X)_0^*$. Even without this assumption we will see that $\mathbf{g}_\mu$ defines an inner product on $\mathcal{M}_0(X)$ whose completion with respect to $\mathbf{g}_\mu$ is canonically identified with (a subset of) the dual of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\mu$.

The RKHS $\mathcal{H}_\mu$, which can be thought of as a tangent space to the space of probability measures at μ, depends on μ. In contrast, the spaces $\mathcal{M}_0(X)$, or even $\mathcal{C}^m(X)_0^*$ for which Theorem 3.4 was established, are independent of μ. We thus do a global change of variables to embed the set of probability distributions $\mathcal{P}(X)$ into the unit sphere of another RKHS $\mathcal{H}_c$. This makes all tangent spaces subspaces of $\mathcal{H}_c$. Further, the metric tensor $\tilde{\mathbf{g}}_\mu$, expressed in these new coordinates, is equivalent to the norm on $\mathcal{H}_c$ (uniformly in μ). This change of variables will be especially useful in Section 5, where we define our new geodesic distance, as it yields uniform coercivity of the metric tensor in these coordinates.

We first recall some results on RKHS before moving to the analysis of the completion of the space $\mathcal{M}_0(X)$ endowed with $\mathbf{g}_\mu$, and then to our global change of coordinates.

4.1. A primer on RKHS, and the operators K_μ and H_μ in the RKHS $\mathcal{H}_\mu$

Take $k\colon X\times X \to [0,\infty)$ a non-negative continuous positive definite and universal kernel on X (see [72, Sect. 2]). In the sequel we will consider $k_c = \exp(-c/\varepsilon)$ or $k_\mu = \exp((f_{\mu,\mu}\oplus f_{\mu,\mu} - c)/\varepsilon)$.

We recall that there exists a unique reproducing kernel Hilbert space (RKHS) induced by the kernel k [9, 60]. It is a subset of $\mathcal{C}(X)$ defined as the completion of the linear span of $\{k(x,\cdot) \mid x \in X\}$ in the norm induced by the inner product

$$\Big\langle \sum_{i=1}^n a_i k(x_i,\cdot), \sum_{j=1}^m b_j k(y_j,\cdot)\Big\rangle_{\mathcal{H}_k} := \sum_{i=1}^n\sum_{j=1}^m a_i b_j k(x_i,y_j).$$

Here, $n, m \in \mathbb{N}_0$, $a_i, b_j \in \mathbb{R}$, and $x_i, y_j \in X$. We write $\mathcal{H}_k$ for this Hilbert space, and use $\mathcal{H}_c$ for the kernel k_c and $\mathcal{H}_\mu$ for the kernel k_μ. By our assumptions on c, both k_c and k_μ have the required properties (recall that a kernel k is positive definite and universal provided $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$ is positive definite and $\mathcal{H}_k$ is dense in $\mathcal{C}(X)$ in supremum norm). Moreover, when k is bounded, the embedding of $\mathcal{H}_k$ into $\mathcal{C}(X)$ is continuous. The evaluation operator is continuous in $\mathcal{H}_k$: for any $\phi \in \mathcal{H}_k$ there holds $\phi(x) = \langle \phi, k(x,\cdot)\rangle_{\mathcal{H}_k}$. More details can be found in Appendix B.

If $\sigma \in \mathcal{H}_k^*$ is a bounded linear functional on $\mathcal{H}_k$, we denote by $H_k[\sigma] \in \mathcal{H}_k$ its representative given by the Riesz theorem. By definition, for any $\phi \in \mathcal{H}_k$ it satisfies

$$\langle \sigma, \phi\rangle = \langle H_k[\sigma], \phi\rangle_{\mathcal{H}_k}. \tag{4.1}$$

Evaluating this formula on $\phi = k(\cdot, y)$ we see that the function $H_k[\sigma]$ is defined by

$$H_k[\sigma](y) = \langle \sigma, k(\cdot, y)\rangle. \tag{4.2}$$

In particular, for $k = k_\mu$ the notation is consistent: H_μ, introduced in Definition 3.3, can be extended into an operator defined on $\mathcal{H}_\mu^*$ with values in $\mathcal{H}_\mu$. The space $\mathcal{H}_k^*$ contains $\mathcal{M}(X)$ as $\mathcal{H}_k \hookrightarrow \mathcal{C}(X)$.

In the setting of Assumption 1 we have $k_\mu \in \mathcal{C}^{(m,m)}(X \times X)$, hence $\mathcal{H}_\mu \hookrightarrow \mathcal{C}^m(X)$ and $\mathcal{C}^m(X)^* \hookrightarrow \mathcal{H}_\mu^*$ (see [71, Cor. 4]).

In addition, we see that the operator K_μ, defined by $H_\mu[\phi\mu]$ for any $\phi \in \mathcal{C}(X)$, actually takes values in $\mathcal{H}_\mu$. We also recall the identity

$$\|\nu\|^2_{\mathcal{H}_k^*} = \|H_k[\nu]\|^2_{\mathcal{H}_k} = \langle \nu, H_k[\nu]\rangle = \iint_{X\times X} k(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y),$$

valid for all $\nu \in \mathcal{M}(X)$, which is usually referred to as a "maximum mean discrepancy distance" on the space $\mathcal{M}(X)$.

Proposition 4.2. *With Assumption 1 for some $m \geq 0$, the operators $H_\mu : \mathcal{C}^m(X)^* \to \mathcal{H}_\mu$ and $K_\mu : \mathcal{C}^m(X) \to \mathcal{H}_\mu$ are compact. Furthermore, $K_\mu : \mathcal{H}_\mu \to \mathcal{H}_\mu$ is self-adjoint. In particular, it has an eigendecomposition. The eigenvalue 1 has multiplicity 1 and all the other eigenvalues belong to $[0, q]$ with $q < 1$ as in Proposition 3.8.*

Proof. Compactness of H_μ follows from Lemma B.2. Compactness of K_μ follows directly.

The self-adjointness of K_μ is straightforward, as for all $\phi, \psi \in \mathcal{H}_\mu$ it holds that

$$\langle K_\mu[\phi], \psi\rangle_{\mathcal{H}_\mu} = \langle H_\mu[\phi\mu], \psi\rangle_{\mathcal{H}_\mu} = \langle \phi, \psi\rangle_{L^2(X,\mu)} = \langle \phi, K_\mu[\psi]\rangle_{\mathcal{H}_\mu}. \tag{4.3}$$

As $\mathcal{H}_\mu$ injects continuously into $\mathcal{C}(X)$, the operator K_μ is also compact from $\mathcal{H}_\mu$ into $\mathcal{H}_\mu$. Thus the spectral theorem [14, Thm. 6.11] applies, giving the eigendecomposition.

We now analyze the spectrum of K_μ. The computation $\langle K_\mu[\phi], \psi\rangle_{\mathcal{H}_\mu} = \langle \phi, \psi\rangle_{L^2(X,\mu)}$ from (4.3) shows that K_μ is non-negative, so all its eigenvalues are non-negative. Now, if $\phi \in \mathcal{H}_\mu \setminus \{0\}$, $\lambda \in [0,\infty)$ with $K_\mu[\phi] = \lambda\phi$, we see that either $\|\phi\|_{\mathcal{C}(X)/\mathbb{R}} = 0$ so that ϕ is constant and thus $\lambda = 1$, or $\|K_\mu[\phi]\|_{\mathcal{C}(X)/\mathbb{R}} = |\lambda| \cdot \|\phi\|_{\mathcal{C}(X)/\mathbb{R}}$ so that $\lambda \in [0, q]$ by Proposition 3.8. ∎

Based on these observations we now prove a strengthening of Theorem 3.9. As $\mathbb{1}_X = H_\mu[\mu] \in \mathcal{H}_\mu$, we can consider $\mathcal{H}_\mu/\mathbb{R}$ the space $\mathcal{H}_\mu$ quotiented by constant functions, equipped with the quotient norm $\|\phi\|_{\mathcal{H}_\mu/\mathbb{R}} = \inf_{\lambda\in\mathbb{R}} \|\phi - \lambda\mathbb{1}_X\|_{\mathcal{H}_\mu}$. The Banach space $\mathcal{H}_\mu/\mathbb{R}$ is isometric to $\{\mathbb{1}_X\}^\perp \subseteq \mathcal{H}_\mu$ and thus canonically is a Hilbert space.

Theorem 4.3. *Let $\mu \in \mathcal{P}(X)$ and $n \geq 1$. On the Hilbert space $\mathcal{H}_\mu/\mathbb{R}$ equipped with the quotient norm the operators $(\mathrm{id} - K_\mu^n)^{-1}$ are continuous and satisfy*

$$\mathrm{id} \leq (\mathrm{id} - K_\mu^n)^{-1} \leq \frac{1}{1-q^n}\mathrm{id}$$

in the partial ordering of self-adjoint operators, with $q < 1$ as in Proposition 3.8. The same conclusion holds in the space $L^2(X,\mu)/\mathbb{R}$.

These operators appear in the metric tensor $\mathbf{g}_\mu$ for $n = 2$ and in $\tilde{\mathbf{g}}_\mu$ for $n = 1$.

Proof of Theorem 4.3. Note that $\mathcal{H}_\mu/\mathbb{R}$ is isometric to the subspace orthogonal to the eigenspace for the eigenvalue 1. So in this space K_μ is self-adjoint and compact with its spectrum being contained in $[0, q]$. The estimate directly follows. Exactly the same arguments carry through to $L^2(X,\mu)/\mathbb{R}$. ∎

Remark 4.4 (Comparison with the differentiable setting). To make the link to the case of $\mathcal{C}^m$-perturbations in the setting of Assumption 1 from the previous section, we recall the following standard bound. Not only $\mathcal{C}^m(X)^* \hookrightarrow \mathcal{H}_\mu^*$, but Lemma B.1 shows that $\|\nu\|_{\mathcal{H}_\mu^*} \leq \|\nu\|_{\mathcal{C}^{m,*}} \cdot \|k_\mu\|_{\mathcal{C}^{(m,m)}}^{1/2}$ for any $\nu \in \mathcal{C}^m(X)^*$. Using Proposition 3.7, we can make this bound uniform in μ. Thus there exists a constant C depending only on ε, c, and m such that for all $\mu \in \mathcal{P}(X)$ and all $\nu \in \mathcal{C}^m(X)^*$, we have

$$\|\nu\|_{\mathcal{H}_\mu^*} \leq C \cdot \|\nu\|_{\mathcal{C}^{m,*}}. \tag{4.4}$$

4.2. The metric tensor in the RKHS $\mathcal{H}_\mu$

We again work under Assumption 1 for some $m \geq 0$ throughout this section. Recall that perturbations integrate to zero against $\mathbb{1}_X$ because of mass conservation, something that we also need to take into account here. For the sake of clarity we state without proof the following result, which would work for any Banach space of functions on X containing $\mathbb{1}_X$.

Lemma 4.5. *The dual of $\mathcal{H}_\mu/\mathbb{R}$ is $\mathcal{H}_{\mu,0}^*$ the space of bounded linear functionals σ on $\mathcal{H}_\mu$ such that $\langle\sigma, \mathbb{1}_X\rangle = 0$.*

In particular, H_μ can be restricted to an operator

$$H_\mu\colon \mathcal{H}^*_{\mu,0} \to \mathcal{H}_\mu/\mathbb{R},$$

which still corresponds to mapping $\sigma \in \mathcal{H}^*_{\mu,0}$ to its representative given by the Riesz theorem: $\langle \sigma, \phi\rangle = \langle H_\mu[\sigma], \phi\rangle_{\mathcal{H}_\mu/\mathbb{R}}$ for any $\phi \in \mathcal{H}_\mu/\mathbb{R}$. From here we can rewrite the metric tensor as follows: If $\dot\mu_1, \dot\mu_2 \in \mathcal{C}^m(X)^*_0$, we see that $H_\mu[\dot\mu_1], H_\mu[\dot\mu_2] \in \mathcal{H}_\mu/\mathbb{R}$. As $(\mathrm{id} - K^2_\mu)^{-1}$ is a bounded operator on this space, we have

$$\begin{aligned}\mathbf{g}_\mu(\dot\mu_1, \dot\mu_2) &= \frac{\varepsilon}{2}\langle \dot\mu_1, (\mathrm{id} - K^2_\mu)^{-1} H_\mu[\dot\mu_2]\rangle \\ &= \frac{\varepsilon}{2}\langle H_\mu[\dot\mu_1], (\mathrm{id} - K^2_\mu)^{-1} H_\mu[\dot\mu_2]\rangle_{\mathcal{H}_\mu/\mathbb{R}}. \end{aligned} \tag{4.5}$$

This formula directly extends to and characterizes the completion of the space $\mathcal{C}^m(X)^*_0$ endowed with $\mathbf{g}_\mu$.

Theorem 4.6. *On $\mathcal{C}^m(X)^*_0$, the quadratic form $\mathbf{g}_\mu$ is positive definite. The completion of $\mathcal{C}^m(X)^*_0$ with respect to this quadratic form coincides with $\mathcal{H}^*_{\mu,0}$, the space of bounded linear functionals σ on $\mathcal{H}_\mu$ such that $\langle\sigma, \mathbb{1}_X\rangle = 0$. Moreover, for any $\dot\mu \in \mathcal{H}^*_{\mu,0}$ one has*

$$\frac{\varepsilon}{2}\|\dot\mu\|^2_{\mathcal{H}^*_{\mu,0}} \le \mathbf{g}_\mu(\dot\mu, \dot\mu) \le \frac{\varepsilon}{2}\frac{1}{1-q^2}\|\dot\mu\|^2_{\mathcal{H}^*_{\mu,0}}, \tag{4.6}$$

with q as in Proposition 3.8.

Proof. We start from formula (4.5), which is a rewriting of Definition 4.1 in the language of RKHS. Combining it with Theorem 4.3 we obtain

$$\frac{\varepsilon}{2}\|H_\mu[\dot\mu]\|^2_{\mathcal{H}_\mu/\mathbb{R}} \le \mathbf{g}_\mu(\dot\mu, \dot\mu) \le \frac{\varepsilon}{2}\frac{1}{1-q^2}\|H_\mu[\dot\mu]\|^2_{\mathcal{H}_\mu/\mathbb{R}},$$

at least for $\dot\mu \in \mathcal{C}^m(X)^*_0$. As H_μ is an isometry from $\mathcal{H}^*_{\mu,0}$ to the space $\mathcal{H}_\mu/\mathbb{R}$, we deduce that (4.6) is valid for all $\dot\mu \in \mathcal{C}^m(X)^*_0$. Thus $\mathbf{g}_\mu$ is positive definite.

It remains to show that the completion of $\mathcal{C}^m(X)^*_0$ is $\mathcal{H}^*_{\mu,0}$. Given what we have just proved, it is enough to prove that $\mathcal{M}_0(X)$ is dense in $\mathcal{H}^*_{\mu,0}$ for the norm topology. We use the isometry of the latter space to $\mathcal{H}_\mu/\mathbb{R}$ via $\mathcal{H}_\mu$. The set $H_\mu[\mathcal{M}(X)]$ is dense in $\mathcal{H}_\mu$, as discrete measures already yield all linear combinations of $\{k_\mu(x,\cdot) \mid x \in X\}$. To conclude, we observe that $\mathcal{M}(X) = \mathbb{R}\cdot\mu \oplus \mathcal{M}_0(X)$ and $H_\mu[\mu] = \mathbb{1}_X$. ∎

4.3. The metric tensor in the RKHS $\mathcal{H}_c$

Although Theorem 4.6 exhibits a very neat structure, it has the main drawback that the ambient space of the tangent space $\mathcal{H}^*_{\mu,0} \cong \mathcal{H}_\mu/\mathbb{R}$ depends on the point μ. The space $\mathcal{H}_\mu$ is however in a simple bijection with the RKHS $\mathcal{H}_c$, built on the kernel $k_c = \exp(-c/\varepsilon)$, as the next proposition shows. In the following we will abbreviate

$$a_\mu := \exp(f_{\mu,\mu}/\varepsilon).$$

Proposition 4.7. *The spaces $\mathcal{H}_c$ and $\mathcal{H}_\mu$ are isometric to one another under the map*

$$\iota_\mu\colon \mathcal{H}_c \to \mathcal{H}_\mu, \quad \phi \mapsto \phi \cdot \exp\Big(\frac{1}{\varepsilon} f_{\mu,\mu}\Big) = \phi \cdot a_\mu.$$

Proof. We can directly verify $\exp(f_{\mu,\mu}(y)/\varepsilon)k_c(x,y) = \exp(-f_{\mu,\mu}(x)/\varepsilon)k_\mu(x,y)$ and hence

$$\begin{aligned}\Big\| \exp\Big(\frac{1}{\varepsilon} f_{\mu,\mu}\Big) \sum_{i=1}^n b_i k_c(x_i,\cdot)\Big\|_{\mathcal{H}_\mu}^2 &= \sum_{i,j=1}^n b_i b_j \cdot a_\mu^{-1}(x_i) a_\mu^{-1}(x_j) k_\mu(x_i,x_j) \\ &= \sum_{i,j=1}^n b_i b_j k_c(x_i,x_j) = \Big\| \sum_{i=1}^n b_i k_c(x_i,\cdot)\Big\|_{\mathcal{H}_c}^2. \quad \blacksquare\end{aligned}$$

We consider the maps $A\colon \mathcal{P}(X) \to \mathcal{M}(X)$ and $B = H_c \circ A\colon \mathcal{P}(X) \to \mathcal{H}_c$ given by

$$\begin{aligned} A(\mu) &= \exp\Big(\frac{1}{\varepsilon} f_{\mu,\mu}\Big)\mu, \\ B(\mu) &= H_c[A(\mu)] = \int_X \exp\Big(\frac{1}{\varepsilon}(f_{\mu,\mu}(y) - c(\cdot,y))\Big)\,\mathrm{d}\mu(y) = \exp\Big(-\frac{1}{\varepsilon} f_{\mu,\mu}\Big).\end{aligned}$$

Recall that the RKHS embeddings $H_c\colon \mathcal{H}_c^* \to \mathcal{H}_c$ and $H_\mu\colon \mathcal{H}_\mu^* \to \mathcal{H}_\mu$ are characterized by (4.1) and (4.2). From a path $(\mu_t)_t \in \mathcal{P}(X)$ one obtains the paths $(H_\mu[\mu_t])_t \in \mathcal{H}_\mu$, $(\alpha_t)_t \in \mathcal{M}_+(X)$, and $(\beta_t)_t \in \mathcal{H}_c$ with $\alpha_t = A(\mu_t)$, $\beta_t = B(\mu_t) = H_c[A(\mu_t)]$. The change of variables $\mu \leftrightarrow \alpha$ was previously used by Feydy et al. in the proof of [28, Prop. 3]. Note that, for a fix measure ν, the space $\mathcal{P}(X)$ embeds into $\mathcal{H}_\nu$ via H_ν, and that the latter can be mapped into $\mathcal{H}_c$ with ι_ν^{-1}. However, this embedding $\mu \mapsto \iota_\nu^{-1}(H_\nu[\mu]) = H_c[a_\nu \cdot \mu]$ of $\mathcal{P}(X)$ into $\mathcal{H}_c$ differs from $B\colon \mu \mapsto H_c[a_\mu \cdot \mu]$.

The measures α_t are naturally interpreted as linear functionals on the common space $\mathcal{H}_c$ (represented by β_t). One can hope that differentiability on this fixed space is well behaved. We will see that the metric tensor can be expressed in the β variable and that it is equivalent to the squared norm on the space $\mathcal{H}_c$ in Theorem 4.12.

We first prove that $\mu \leftrightarrow \alpha$ and $\mu \leftrightarrow \beta$ are valid changes of variables in the sense that A and B are homeomorphisms onto their images. Recall that P_1 stands for the projection map $(x,y) \mapsto x$.

Theorem 4.8. *The maps A and B are homeomorphisms onto their images. Explicitly, we have the following:*

(1) *A is a weak-$*$ homeomorphism onto $\mathcal{A} := \{\alpha \in \mathcal{M}_+(X) \mid \|H_c[\alpha]\|_{\mathcal{H}_c} = 1\}$ with inverse $A^{-1}(\alpha) = (\mathrm{P}_1)_\#(k_c \cdot \alpha \otimes \alpha)$. In particular, $\mathcal{A}$ is weak-$*$-compact.*

(2) *B is weak-$*$-to-weak homeomorphic onto $\mathcal{B} := \{\beta \in H_c[\mathcal{M}_+(X)] \mid \|\beta\|_{\mathcal{H}_c} = 1\}$. Weak convergence and norm convergence agree on $\mathcal{B}$, and the set $\mathcal{B}$ is weakly and norm compact in $\mathcal{H}_c$.*

Note that $\mathcal{B}$ is the intersection of the convex cone $H_c[\mathcal{M}_+(X)]$ with the unit sphere. The proof for the weak compactness of $\mathcal{B} \subset \mathcal{H}_c$ relies heavily on the positivity constraint of the cone $H_c[\mathcal{M}_+(X)]$. In contrast, the weak closure of the whole unit sphere in infinite dimension is the unit ball, which is not norm compact.

Remark 4.9. As a sidenote which might be interesting on its own, the characterization of $\mathcal{B}$ also gives non-negativity of the entropic self-transport potential of any measure with respect to any non-negative cost function inducing a positive definite universal kernel. By Theorem 4.8, we have $\|\exp(-f_{\mu,\mu}/\varepsilon)\|_{\mathcal{H}_c} = 1$ and hence the estimate (B.1) implies that $\|\exp(-f_{\mu,\mu}/\varepsilon)\|_\infty \le \|\exp(-f_{\mu,\mu}/\varepsilon)\|_{\mathcal{H}_c} \cdot \sup_{x\in X} k_c(x,x) \le 1$. In particular,

$$f_{\mu,\mu}(x) \ge 0 \quad \text{for all } x \in X.$$

With this we can slightly strengthen the contraction bound in Proposition 3.8 to $q = 1 - \exp(-\frac{5}{2}\|c\|_\infty/\varepsilon)$.

Proof of Theorem 4.8. We first treat the measure-valued map A and then extend the results to the map B using properties of the RKHS embedding H_c.

The map A and its image. Denote $\alpha = A(\mu) = a_\mu \mu$. Note that the map A is injective since μ can be reconstructed as $\mu = F(\alpha) := (\mathrm{P}_1)_\#(k_c \cdot \alpha \otimes \alpha)$ thanks to (2.6). This also implies that the inverse map is weak-$*$-continuous: Let $\alpha_n \to \alpha$ converge weak-$*$ with corresponding $\mu_n = (\mathrm{P}_1)_\#(k_c \cdot \alpha_n \otimes \alpha_n)$. Then $\alpha_n \otimes \alpha_n$ converges weak-$*$ to $\alpha \otimes \alpha$ (see e.g. [68, Lem. 7.3]). Weak convergence of μ_n to μ easily follows. To prove that the forward map is continuous, it suffices to use the continuity of $f_{\mu,\mu}$ as a function of μ in supremum norm (see Proposition 3.7).

We now characterize the image of A: Clearly $A(\mu) \ge 0$ and

$$\|H_c[A(\mu)]\|^2_{\mathcal{H}_c} = \langle a_\mu \mu \otimes a_\mu \mu, k_c\rangle = \langle \mu \otimes \mu, k_\mu\rangle = 1$$

for all $\mu \in \mathcal{P}(X)$. The map $F\colon \mathcal{A} \to \mathcal{P}(X)$ is a left inverse of $A\colon \mathcal{P}(X) \to \mathcal{A}$. To show that F is also a right inverse, we may equivalently show that F is injective. We have $F(\alpha) = \mu$ if and only if $k_c \cdot \alpha \otimes \alpha \in \Pi(\mu,\mu)$. For every such $\alpha \in \mathcal{A}$ it must hold that $\alpha \ll \mu$. Thus $k_c \cdot \alpha \otimes \alpha = \exp(\phi \oplus \phi) k_c \cdot \mu \otimes \mu$ for $\phi = \log \frac{\mathrm{d}\alpha}{\mathrm{d}\mu}\colon X \to [-\infty, \infty)$. A transport plan of this form is necessarily optimal for the entropic optimal transport problem (1.1) (see [56, Thm. 2.1]). Hence $\phi = f_{\mu,\mu}/\varepsilon$ and $\alpha = A(\mu)$. Finally, $\mathcal{A}$ is weak-$*$-compact as the weak-$*$-continuous image of the weak-$*$-compact set $\mathcal{P}(X)$.

The map B and its image. The map A is injective. The map H_c is a bijection between $\mathcal{H}_c^*$ and $\mathcal{H}_c$ by definition, so it is injective. Thus B is injective. It is clear that $\mathcal{B}$ is the image of $\mathcal{A}$ by H_c. Thus $\mathcal{B}$ is the image of $\mathcal{P}(X)$ by B.

As the injection of $\mathcal{H}_c$ into $\mathcal{C}(X)$ is continuous by (B.1), the map H_c is continuous from $\mathcal{M}(X)$ endowed with the weak-$*$ topology into $\mathcal{H}_c$ endowed with the weak topology. Thus B is weak-$*$-to-weak continuous and $\mathcal{B}$ is weakly compact as the image of the weak-$*$-compact set $\mathcal{A}$. As a continuous bijection between compact Hausdorff spaces, B is a weak-$*$-to-weak homeomorphism.

Finally, if $(\beta_n)_n$ is any sequence in $\mathcal{B}$ which converges weakly to a limit β in $\mathcal{H}_c$, as $\mathcal{B}$ is weakly compact we must have $\beta \in \mathcal{B}$ so $\|\beta\|_{\mathcal{H}_c} = 1 = \lim_{n\to\infty}\|\beta_n\|_{\mathcal{H}_c}$. This implies norm convergence of the sequence $(\beta_n)_n$. ∎

After the investigation of the topological properties of these changes of variables, we move on to the differential properties. For a path $(\mu_t)_t$ in $\mathcal{P}(X)$, we denote $\alpha_t = A(\mu_t)$, $\beta_t = B(\mu_t)$. By definition this means $\langle \alpha_t, \phi\rangle = \langle \beta_t, \phi\rangle_{\mathcal{H}_c}$ for all $\phi \in \mathcal{H}_c$. We will express the metric tensor in the variable $\dot\beta_t := \frac{\mathrm{d}}{\mathrm{d}t}\beta_t = \frac{\mathrm{d}}{\mathrm{d}t}\exp(-f_{\mu_t,\mu_t}/\varepsilon) \in \mathcal{H}_c$. Let us first define the push-forward of the metric tensor before making a rigorous link to Definition 4.1 in Theorem 4.12.

Definition 4.10 (The metric tensor on $\mathcal{H}_c$). If $\mu \in \mathcal{P}(X)$ with $\beta = B(\mu) \in \mathcal{B}$, and $\dot\beta \in \mathcal{H}_c$ with $\langle \beta, \dot\beta\rangle_{\mathcal{H}_c} = 0$, we define the quadratic form $\tilde{\mathbf{g}}_\mu(\dot\beta, \dot\beta)$ by

$$\tilde{\mathbf{g}}_\mu(\dot\beta, \dot\beta) = \frac{\varepsilon}{2}(\|\dot\beta\|^2_{\mathcal{H}_c} + 2\langle a_\mu \cdot \dot\beta, (\mathrm{id} - K_\mu)^{-1}[a_\mu \cdot \dot\beta]\rangle_{L^2(X,\mu)}).$$

To explicitly link $\mathbf{g}$ and $\tilde{\mathbf{g}}$, we need to consider curves valued in $\mathcal{P}(X)$ and their corresponding curves in $\mathcal{B}$. The following definition fixes the convention about what we mean by a *differentiable path*.

Definition 4.11 (Differentiability of paths). Let $\mathbb{X}$ be a vector space with a topology τ. We say that a path $(x_s)_{s\in I}$ in $\mathbb{X}$ is τ-differentiable at t when t is an accumulation point of I and the difference quotients $(x_s - x_t)/(s-t)$ converge in τ to some $\dot x_t$ as $s \to t$. When $\mathbb{X}$ is a normed vector space and τ is the norm topology, we drop the τ in the notation.

We define $\dot\mu_t$ by $\langle \dot\mu_t, \phi\rangle = \frac{\mathrm{d}}{\mathrm{d}s}|_{s=t}\langle \mu_s, \phi\rangle$ for $\phi \in \mathcal{H}_{\mu_t}$ whenever this is a bounded linear functional on $\mathcal{H}_{\mu_t}$, i.e. when the path $(\mu_s)_s$, interpreted in the dual space $\mathcal{H}^*_{\mu_t}$, is weakly differentiable in $\mathcal{H}^*_{\mu_t}$ at t. Note that automatically $\dot\mu_t \in \mathcal{H}^*_{\mu_t,0}$.

In the sequel, for paths $(\mu_t)_t$, $(\alpha_t)_t$, $(\beta_t)_t$ we write $\mu = \mu_0$, $\dot\mu = \dot\mu_0$ and similarly $\alpha = \alpha_0$, $\dot\alpha = \dot\alpha_0$ and $\beta = \beta_0$, $\dot\beta = \dot\beta_0$.

Theorem 4.12. *Let $(\mu_t)_t$ be a path in $\mathcal{P}(X)$, such that the associated path $(\beta_t)_t = (B(\mu_t))_t$ is weakly differentiable in $\mathcal{H}_c$ at $t = 0$. Then $\langle \dot\beta, \beta\rangle_{\mathcal{H}_c} = 0$, the path $(\mu_t)_t$ is weakly differentiable in $\mathcal{H}^*_\mu$ at $t = 0$, and*

$$\mathbf{g}_\mu(\dot\mu, \dot\mu) = \tilde{\mathbf{g}}_\mu(\dot\beta, \dot\beta).$$

To prove this theorem we first prove a preliminary lemma relating $\dot\beta$ and $\dot\mu$. Note that since $\beta_t = \exp(-f_{\mu_t,\mu_t}/\varepsilon)$, we have $\dot\beta_t = -\frac{1}{\varepsilon}\beta_t \cdot \frac{\mathrm{d}}{\mathrm{d}t} f_{\mu_t,\mu_t}$ and (4.7) below is formally equivalent to the last equation of Proposition 3.14. The differences are the regularity assumptions and the functional spaces involved.

Lemma 4.13. *Let $(\mu_t)_t$ be a weak-$*$ continuous path in $\mathcal{P}(X)$ and denote $f_{t,t} = f_{\mu_t,\mu_t}$. Then the following are equivalent:*

(1) *The corresponding path $(\beta_t)_t$ is weakly (resp. norm-) differentiable in $\mathcal{H}_c$ at $t = 0$.*

(2) *$(\mu_t)_t$ is weakly (resp. norm-) differentiable in $\mathcal{H}_\mu^*$ at $t = 0$ and the difference quotients $(f_{t,t} - f_{0,0})/t$ are bounded in $\mathcal{C}(X)$ as $t \to 0$.*

In this case,

$$H_\mu[\dot{\mu}] = (\mathrm{id} + K_\mu)[a_\mu \cdot \dot{\beta}]. \tag{4.7}$$

Explicitly, this means that for all $\phi \in \mathcal{H}_\mu$ it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \langle \mu_t, \phi\rangle = \langle (\mathrm{id} + K_\mu)[a_\mu \dot{\beta}], \phi\rangle_{\mathcal{H}_\mu} = \langle \dot{\beta}, \beta \cdot (\mathrm{id} + K_\mu)[\phi]\rangle_{\mathcal{H}_c}.$$

The condition of boundedness of the difference quotients $(f_{t,t} - f_{0,0})/t$ in the second point is not so natural. However, it is not clear how to remove it because we would need to linearize the Schrödinger system around (μ, μ) in the space $\mathcal{H}_\mu$ (which depends on μ), and we did not find a way to apply the implicit function theorem in this case.

Proof of Lemma 4.13. We split the proof into two parts.

(1) $\Rightarrow$ (2). Recall that we have $\mu_t = \alpha_t \exp(-f_{\mu_t,\mu_t}/\varepsilon)$ and $\beta_t = a_{\mu_t}^{-1} = \exp(-f_{\mu_t,\mu_t}/\varepsilon)$, thus $\mu_t = \alpha_t \beta_t$. We will apply Leibniz's rule to this formula, but we rederive it given the low regularity of the objects involved. We write

$$H_\mu\Big[\frac{\mu_t - \mu}{t}\Big] = H_\mu\Big[\alpha_t \cdot \frac{\beta_t - \beta}{t}\Big] + H_\mu\Big[\frac{\alpha_t - \alpha}{t} \cdot \beta\Big]. \tag{4.8}$$

In the first term, $\frac{\beta_t - \beta}{t} \to \dot{\beta}$ weakly in $\mathcal{H}_c$, hence uniformly, while $\alpha_t \to \alpha$ weak-$*$ in $\mathcal{M}(X)$. In particular, $\alpha_t \cdot \frac{\beta_t - \beta}{t} \to \alpha \cdot \dot{\beta}$ weak-$*$ in $\mathcal{M}(X)$ and its RKHS embedding converges in norm by Lemma B.2. The limit is $H_\mu[\alpha \cdot \dot{\beta}] = K_\mu[\dot{\beta}/\beta]$.

With $H_\mu[\nu] = 1/\beta \cdot H_c[\nu/\beta]$ for all $\nu \in \mathcal{M}(X)$ the second term can be rewritten as $H_c[\frac{\alpha_t - \alpha}{t}]/\beta = \frac{\beta_t - \beta}{t}/\beta$, which converges weakly (resp. in norm) to $\dot{\beta}/\beta$. The self-transport potentials are even differentiable in time in $\mathcal{C}(X)$, which is immediate from weak differentiability of $(\beta_t)_t$ in $\mathcal{H}_c$ and Lemma B.2.

(2) $\Rightarrow$ (1). We use $\beta_t = H_c[\mu_t/\beta_t]$ and $H_\mu[\nu] = 1/\beta \cdot H_c[\nu/\beta]$ for $\nu \in \mathcal{M}(X)$ to calculate

$$\begin{aligned}\frac{\beta_t - \beta}{t} &= H_c\Big[\frac{(\beta_t^{-1} - \beta^{-1})\mu_t + \beta^{-1}(\mu_t - \mu)}{t}\Big] \\ &= H_c\Big[\frac{\beta_t^{-1} - \beta^{-1}}{t} \cdot \mu_t\Big] + \beta \cdot H_\mu\Big[\frac{\mu_t - \mu}{t}\Big].\end{aligned} \tag{4.9}$$

We now show the desired convergence of these difference quotients in $\mathcal{H}_c$ as $t \to 0$. For the second summand this is due to Proposition 4.7 and the weak (resp. norm-) differentiability of $(\mu_t)_t$. The first summand is trickier. Boundedness of the difference quotients and weak-$*$ convergence of μ_t to μ give $\|\cdot\|_{\mathcal{C}(X)^*}$-boundedness of the measure in the argument of H_c. The embedding $H_c\colon \mathcal{M}(X) \to \mathcal{H}_c$ is compact by Lemma B.2. Thus a convergent subsequence at times $(t_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} t_k = 0$ can be extracted along which the first summand converges in norm in $\mathcal{H}_c$ to a limit. This yields weak (resp. norm-) convergence of the left-hand side in $\mathcal{H}_c$. In particular, we obtain uniform convergence of the difference quotients $(\beta_{t_k} - \beta)/t_k \to v$ for some $v \in \mathcal{H}_c$ and $(\beta_{t_k}^{-1} - \beta^{-1})/t_k \to -v/\beta^2$ in $\mathcal{C}(X)$. Taking the limit $k \to \infty$ in (4.9) yields

$$v = -H_c[v/\beta^2 \cdot \mu] + \beta \cdot H_\mu[\dot{\mu}] = \beta \cdot (H_\mu[\dot{\mu}] - K_\mu[v/\beta]), \tag{4.10}$$

or equivalently $v/\beta = (\mathrm{id} + K_\mu)^{-1} H_\mu[\dot{\mu}]$. As all subsequences of $(\beta_t - \beta)/t$ have convergent subsequences with the same limit v, we have overall convergence as $t \to 0$ and $v = \dot{\beta}$. With self-adjointness of K_μ on $\mathcal{H}_\mu$ (Proposition 4.2) and $a_\mu = 1/\beta$, the explicit formula follows from (4.10). ∎

Proof of Theorem 4.12. The equation $\langle \dot{\beta}, \beta \rangle_{\mathcal{H}_c} = 0$ is directly obtained by differentiating the equation $\|\beta_t\|^2_{\mathcal{H}_c} = 1$ in time.

We then plug the formula for $\dot{\mu}$ from Lemma 4.13 into the metric tensor

$$\begin{aligned} \mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) &= \frac{\varepsilon}{2} \langle \dot{\mu}, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot{\mu}] \rangle \\ &= \frac{\varepsilon}{2} \langle (\mathrm{id} + K_\mu)[\dot{\beta}/\beta], (\mathrm{id} - K_\mu^2)^{-1} (\mathrm{id} + K_\mu)[\dot{\beta}/\beta] \rangle_{\mathcal{H}_\mu}. \end{aligned}$$

The operator $(\mathrm{id} + K_\mu)$ is self-adjoint on $\mathcal{H}_\mu$. Furthermore,

$$(\mathrm{id} + K_\mu)(\mathrm{id} - K_\mu^2)^{-1}(\mathrm{id} + K_\mu) = (\mathrm{id} + K_\mu)(\mathrm{id} - K_\mu)^{-1} = \mathrm{id} + 2K_\mu(\mathrm{id} - K_\mu)^{-1}.$$

To conclude, use $\langle \phi, K_\mu[\psi] \rangle_{\mathcal{H}_\mu} = \langle \phi, \psi \rangle_{L^2(X,\mu)}$ and Proposition 4.7. ∎

We finish this section by studying the continuity properties of the metric tensor $\tilde{\mathbf{g}}$, and by proving that it can be uniformly compared to the norm of $\mathcal{H}_c$.

Proposition 4.14. *The metric tensor $\tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta})$ is jointly continuous (resp. lower semicontinuous) with respect to weak-$*$ convergence of μ in $\mathcal{P}(X)$ and norm convergence (resp. weak convergence) of $\dot{\beta}$ in $\mathcal{H}_c$.*

Proof. The first part $\frac{\varepsilon}{2}\|\dot{\beta}\|^2_{\mathcal{H}_c}$ of the metric tensor $\tilde{\mathbf{g}}_\mu$ does not depend on μ and is of course continuous (resp. lower semicontinuous) with respect to norm convergence (resp. weak convergence) of $\dot{\beta}$ in $\mathcal{H}_c$.

For the second part, we actually claim that the map

$$\begin{aligned} &\left\{(\mu, \phi) \in \mathcal{P}(X) \times \mathcal{H}_c \;\middle|\; \langle B(\mu), \phi \rangle_{\mathcal{H}_c} = 0 \right\} \to \mathbb{R}, \\ &\qquad\qquad (\mu, \phi) \mapsto \langle a_\mu \cdot \phi, (\mathrm{id} - K_\mu)^{-1}[a_\mu \cdot \phi] \rangle_{L^2(X,\mu)} \end{aligned}$$

is jointly continuous with respect to weak-$*$ convergence on $\mathcal{P}(X)$ and weak convergence on $\mathcal{H}_c$. Indeed, let $\mu_n \to \mu$ weak-$*$ in $\mathcal{P}(X)$, $\phi_n \to \phi$ weakly in $\mathcal{H}_c$, and $\langle B(\mu_n), \phi_n \rangle = 0$. In particular, we have uniform convergence $f_n = f_{\mu_n,\mu_n} \to f_{\mu,\mu}$, hence $a_{\mu_n} \to a_\mu$, and by Lemma B.2 also $\phi_n \to \phi$ uniformly. Let us write

$$u_n = (\mathrm{id} - K_{\mu_n})^{-1}[a_{\mu_n} \cdot \phi_n] \quad \Leftrightarrow \quad u_n = a_{\mu_n} \cdot \phi_n + K_{\mu_n} u_n.$$

We show that the sequence u_n converges to $u = (\mathrm{id} - K_\mu)^{-1}[a_\mu \cdot \phi]$ in $\mathcal{C}(X)/\mathbb{R}$. Indeed, the uniform convergence of $a_{\mu_n} \cdot \phi_n$ and the operator norm estimate in Proposition 3.8 show that the functions u_n are uniformly bounded in $\mathcal{C}(X)/\mathbb{R}$.

By Proposition 3.7 and the Arzelà–Ascoli theorem, the potentials f_{μ_n,μ_n} are equi-bounded and equi-continuous.

Thus it is clear that the functions $K_{\mu_n} u_n$ are equi-continuous and equi-bounded. Hence, up to extraction of a subsequence, $K_{\mu_n} u_n$ converges uniformly to a limit. Now also $u_n = a_{\mu_n} \cdot \phi_n + K_{\mu_n} u_n$ converges uniformly to a limit u. Once we have this information, we can identify this limit with $(\mathrm{id} - K_\mu)^{-1}[a_\mu \cdot \phi]$, because

$$u_n = a_{\mu_n} \cdot \phi_n + K_{\mu_n} u_n \to a_\mu \cdot \phi + K_\mu u = u.$$

Once we have this convergence,

$$\langle a_{\mu_n} \cdot \phi_n, (\mathrm{id} - K_{\mu_n})^{-1}[a_{\mu_n} \cdot \phi_n] \rangle_{L^2(X,\mu_n)} = \int_X a_{\mu_n} \phi_n u_n \, \mathrm{d}\mu_n,$$

and the integrand $a_{\mu_n} \phi_n u_n$ converges uniformly to $a_\mu \phi u$, while the measure μ_n converges weak-$*$ to μ. It is easy to pass to the limit using Lemma A.2. ∎

Proposition 4.15 (Equivalence to the norm on $\mathcal{H}_c$). *Let $\mu \in \mathcal{P}(X)$ and $\dot{\beta} \in \mathcal{H}_c$ with $\langle B(\mu), \dot{\beta} \rangle_{\mathcal{H}_c} = 0$. It holds that*

$$\frac{\varepsilon}{2} \|\dot{\beta}\|_{\mathcal{H}_c}^2 \leq \tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta}) \leq C \frac{\varepsilon}{2} \|\dot{\beta}\|_{\mathcal{H}_c}^2$$

for $C = 1 + 2\exp(\frac{11}{2\varepsilon} \|c\|_\infty)$.

Proof. From Theorem 4.3, on the space $L^2(X,\mu)/\mathbb{R}$ we have $\mathrm{id} \leq (\mathrm{id} - K_\mu)^{-1} \leq \frac{1}{1-q} \mathrm{id}$ with $q = 1 - \exp(-4\|c\|_\infty/\varepsilon)$ as in Proposition 3.8 (and in particular $(\mathrm{id} - K_\mu)^{-1}$ is non-negative). Thus we get

$$\frac{\varepsilon}{2} \|\dot{\beta}\|_{\mathcal{H}_c}^2 \leq \tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta}) \leq \frac{\varepsilon}{2} \Big(\|\dot{\beta}\|_{\mathcal{H}_c}^2 + \frac{2}{1-q} \|a_\mu \cdot \dot{\beta}\|_{L^2(X,\mu)/\mathbb{R}}^2 \Big).$$

We then control the $L^2(X,\mu)/\mathbb{R}$ norm by the supremum norm, using that $\|\dot{\beta}\|_\infty \leq \|\dot{\beta}\|_{\mathcal{H}_c}$ as the cost function is non-negative. Thus

$$\frac{\varepsilon}{2} \|\dot{\beta}\|_{\mathcal{H}_c}^2 \leq \tilde{\mathbf{g}}_\mu(\dot{\beta}, \dot{\beta}) \leq \frac{\varepsilon}{2} \Big(\|\dot{\beta}\|_{\mathcal{H}_c}^2 + \frac{2}{1-q} \|a_\mu\|_\infty \|\dot{\beta}\|_{\mathcal{H}_c}^2 \Big).$$

Finally, using [26, Thm. 1.2] we have $\|a_\mu\|_\infty \leq \exp(\|f_{\mu,\mu}\|_\infty/\varepsilon) \leq \exp(\frac{3}{2}\|c\|_\infty/\varepsilon)$, and together with the definition of q we conclude the proof. ∎

4.4. The metric tensor as ε varies

Let us now discuss how the metric tensor varies as a function of ε. Though it is not the main goal of our work, which rather focuses on showing that, for ε fixed, S_ε induces a Riemannian geometry, it can give some intuition on the influence of the parameter ε. For this discussion we fix X to be a bounded convex subset of $\mathbb{R}^d$ with non-empty interior, and $c(x, y) = \|x - y\|^2$ is the quadratic cost. We add a superscript ε to stress the fact that, contrary to the rest of this work, ε is not fixed anymore. At the level of the Sinkhorn divergence, as the Kullback–Leibler divergence is invariant under reparametrization,

$$S_\varepsilon(\mu, \nu) = \varepsilon S_1(L^\varepsilon_\# \mu, L^\varepsilon_\# \nu), \tag{4.11}$$

with $L^\varepsilon(x) = x/\sqrt{\varepsilon}$; see [52]. It shows how $\sqrt{\varepsilon}$ acts as a typical length scale.

We first show that, for $\varepsilon \in (0, \infty)$, the metric tensor changes smoothly with respect to ε (however, it is not a monotone function of ε for μ, $\dot{\mu}$ fixed: this can be seen in Figure 3).

Proposition 4.16. *Fix $\mu \in \mathcal{P}(X)$ and $\dot{\mu} \in \mathcal{C}^m(X)^*_0$ for some $m \geq 0$. Then the function $\varepsilon \mapsto \mathbf{g}^\varepsilon_\mu(\dot{\mu}, \dot{\mu})$ defined on $(0, \infty)$ is continuous.*

In this proposition we restrict to tangent vectors in $\mathcal{C}^m(X)^*_0$, which embed into the ε-dependent tangent space $\mathcal{H}^*_{\mu,0}$ for any ε.

Proof of Proposition 4.16. By checking that it satisfies the Schrödinger system, we have, with $L_\varepsilon(x) = x/\sqrt{\varepsilon}$,

$$f^\varepsilon_{\mu,\mu}(x) = \varepsilon f^1_{L^\varepsilon_\# \mu, L^\varepsilon_\# \mu}\left(\frac{x}{\sqrt{\varepsilon}}\right).$$

With Proposition 3.7, it gives that $f^\varepsilon_{\mu,\mu}$ varies continuously with ε in $\mathcal{C}^m(X)$. From this and the bound of Proposition 3.8 which is locally uniform in ε, we see that both $(\mathrm{id} - K^\varepsilon_\mu)^{-1}$ and H^ε_μ vary continuously in operator norm with ε. The conclusion follows from the explicit expression of $\mathbf{g}^\varepsilon_\mu(\dot{\mu}, \dot{\mu})$; see Definition 4.1. ∎

We now move to the limit cases $\varepsilon \to 0$ and $\varepsilon \to \infty$. The connection between optimal transport, regularized optimal transport, and MMD distances in these limit cases has been investigated in the literature [64]. It is thus a natural question to study what happens to our metric tensor in these same limiting cases. We first focus on the large ε limit as it is the easiest case.

Theorem 4.17. *For $\mu \in \mathcal{P}(X)$ and $\dot{\mu} = -\mathrm{div}(\mu v)$ with $v \in L^1(X, \mu; \mathbb{R}^d)$, it holds that*

$$\lim_{\varepsilon \to \infty} \mathbf{g}^\varepsilon_\mu(\dot{\mu}, \dot{\mu}) = \left\| \int_X v(x) \, \mathrm{d}\mu(x) \right\|^2.$$

The limit of the metric tensor is degenerate, as $\|\int v(x) \, \mathrm{d}\mu(x)\| = \|\dot{m}_\mu\|$ with $m_\mu \in \mathbb{R}^d$ the center of mass of μ. Thus in the limit only the motion of the center of mass matters.

Proof of Proposition 4.17. From equation (4.6), as $q \to 0$ when $\varepsilon \to \infty$, we have

$$\begin{aligned}\mathbf{g}^{\varepsilon}_{\mu}(\dot\mu,\dot\mu) \sim \frac{\varepsilon}{2}\|\dot\mu\|^2_{\mathcal{H}^{*,\varepsilon}_{\mu,0}} &= \frac{\varepsilon}{2}\iint_{X\times X} k^{\varepsilon}_{\mu}(x,y)\,\mathrm{d}\dot\mu(x)\,\mathrm{d}\dot\mu(y)\\ &= \frac{\varepsilon}{2}\iint_{X\times X} [\nabla_{xy}k^{\varepsilon}_{\mu}(x,y)]\cdot[v(x)\otimes v(y)]\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y),\end{aligned}$$

where the last equality follows from integration by parts. We then compute

$$\nabla_{xy}k^{\varepsilon}_{\mu}(x,y) = \left(\frac{2}{\varepsilon}\mathrm{id} + \frac{(\nabla f^{\varepsilon}_{\mu,\mu}(x) - 2(x-y))\otimes(\nabla f^{\varepsilon}_{\mu,\mu}(y) - 2(y-x))}{\varepsilon^2}\right)k^{\varepsilon}_{\mu}(x,y).$$

But $k^{\varepsilon}_{\mu} \to 1$ in $\mathcal{C}(X\times X)$ and $\pi_{\mu,\varepsilon} \to \mu\otimes\mu$ as $\varepsilon\to\infty$. This is clear as $\|f^{\varepsilon}_{\mu,\mu}\|_\infty \le 3/2\|c\|_\infty$ for any ε. Moreover, $2y - \nabla f^{\varepsilon}_{\mu,\mu}(y) = 2\int k^{\varepsilon}_{\mu}(y,z)z\,\mathrm{d}\mu(z)$, which is a classical identity that we rederive below in (5.10). We deduce that $2\,\mathrm{id} - \nabla f^{\varepsilon}_{\mu,\mu}$ converges to $2m_\mu$, as $\varepsilon\to\infty$, hence is bounded. As moreover X is bounded, putting these considerations together, as a (matrix-valued) function,

$$\frac{\varepsilon}{2}\nabla_{xy}k^{\varepsilon}_{\mu}(x,y) \to \mathrm{id}, \quad \text{as } \varepsilon\to\infty$$

uniformly on $X\times X$. Plugging this into the previous computations, we find

$$\lim_{\varepsilon\to\infty}\mathbf{g}^{\varepsilon}_{\mu}(\dot\mu,\dot\mu) = \lim_{\varepsilon\to\infty}\frac{\varepsilon}{2}\|\dot\mu\|^2_{\mathcal{H}^{\varepsilon,*}_{\mu,0}} = \iint_{X\times X} v(x)\cdot v(y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y). \qquad \blacksquare$$

We then turn to the regime of vanishing ε. In this case we only do informal computations, as making the expansion rigorous would require a fine analysis of the behavior of the Schrödinger potentials as $\varepsilon\to 0$ in the compact case. We identify a measure with its density with respect to the Lebesgue measure. Writing Δ for the Laplacian (the trace of the Hessian matrix) and $Lf := -\Delta f - \nabla f\cdot\nabla\log\mu$, we claim that it should hold that

$$\mathbf{g}^{\varepsilon}_{\mu}(\dot\mu,\dot\mu) \to \left\langle\frac{\dot\mu}{\mu}, L^{-1}\left[\frac{\dot\mu}{\mu}\right]\right\rangle_{L^2(X,\mu)}, \quad \text{as } \varepsilon\to 0. \tag{4.12}$$

Remark 4.18. In the case where $\dot\mu = -\operatorname{div}(\mu\nabla\psi)$, we see that the limit in (4.12) is the metric tensor $\mathbf{g}^{0}_{\mu}$ of classical optimal transport as defined in (1.3). Indeed, in this setting $L^{-1}[\dot\mu/\mu] = \psi$, which can be checked by expanding the divergence:

$$\frac{\dot\mu}{\mu} = -\frac{\nabla\mu\cdot\nabla\psi}{\mu} - \Delta\psi = -\nabla\log\mu\cdot\nabla\psi - \Delta\psi = L\psi.$$

Thus we obtain from (4.12),

$$\mathbf{g}^{\varepsilon}_{\mu}(\dot\mu,\dot\mu) \to \langle\operatorname{div}(\mu\nabla\psi),\psi\rangle = \int_X |\nabla\psi|^2\,\mathrm{d}\mu = \mathbf{g}^{0}_{\mu}(\dot\mu,\dot\mu).$$

We now discuss the sketch of the proof of the claim (4.12). The main observation is that if one has an operator expansion $K^{\varepsilon}_{\mu} = \mathrm{id} + \varepsilon \tilde{L} + o(\varepsilon)$ for some invertible operator $\tilde{L}$, then it holds that $(K^{\varepsilon}_{\mu})^2 = \mathrm{id} + 2\varepsilon \tilde{L} + o(\varepsilon)$. Therefore, as $H^{\varepsilon}_{\mu}[\cdot] = K^{\varepsilon}_{\mu}[\cdot/\mu]$, we have

$$\begin{aligned}\frac{\varepsilon}{2}\Big\langle \dot{\mu}, (\mathrm{id} - (K^{\varepsilon}_{\mu})^2)^{-1} K^{\varepsilon}_{\mu}\Big[\frac{\dot{\mu}}{\mu}\Big]\Big\rangle &= \frac{1}{4}\Big\langle \dot{\mu}, (-\tilde{L} + o(1))^{-1}(\mathrm{id} + \varepsilon\tilde{L} + o(\varepsilon))\Big[\frac{\dot{\mu}}{\mu}\Big]\Big\rangle \\ &= \frac{1}{4}\Big\langle \dot{\mu}, (-\tilde{L})^{-1}\Big[\frac{\dot{\mu}}{\mu}\Big]\Big\rangle + o(1),\end{aligned}$$

and the conclusion follows if we check that $L = -4\tilde{L}$. Thus we need to identify $\tilde{L}$, i.e. the first-order term in the expansion of K^{ε}_{μ} as $\varepsilon \to 0$. It has already been characterized in some cases: in [50], where the ambient space is a Riemannian manifold, and in [2], which gives a probabilistic interpretation of the resulting expansion. However, it needs measures supported on the whole space $\mathbb{R}^d$. None of these works prove a limit in a sense strong enough to rigorously deduce (4.12) as far as we understand. We leave it for future work to give a mathematical proof of (4.12).

We will only explain how one can guess the limit, following the arguments of [55]. To identify $\tilde{L}$ at least informally, we need a better understanding of the self-dual potentials. Identifying a measure with its density, in the limit $\varepsilon \to 0$ the Laplace formula gives

$$H^{\varepsilon}_{c}[\phi](y) = \int_X \exp\Big(-\frac{1}{\varepsilon}\|x-y\|^2\Big)\phi(x)\,\mathrm{d}x = (\pi\varepsilon)^{d/2}\Big(\phi(x) + \frac{\varepsilon}{4}\Delta\phi(y) + o(\varepsilon)\Big).$$

With $a_{\varepsilon} = \exp(f^{\varepsilon}_{\mu,\mu}/\varepsilon)$, the condition $K^{\varepsilon}_{\mu}[\mathbb{1}_X] = \mathbb{1}_X$ reads $a_{\varepsilon} H^{\varepsilon}_{c}[a_{\varepsilon}\mu] = \mathbb{1}_X$, that is,

$$a_{\varepsilon}^2 = \Big((\pi\varepsilon)^{d/2}\Big(\mu + \frac{\varepsilon}{4}\frac{\Delta(a_{\varepsilon}\mu)}{a_{\varepsilon}} + o(\varepsilon)\Big)\Big)^{-1} = \frac{1}{(\pi\varepsilon)^{d/2}\mu}\Big(1 - \frac{\varepsilon}{4}\frac{\Delta(a_{\varepsilon}\mu)}{a_{\varepsilon}\mu} + o(\varepsilon)\Big).$$

Thus at first order $a_{\varepsilon} \sim (\pi\varepsilon)^{-d/4}\mu^{-1/2}$, which plugged again into the expansion in the right-hand side yields

$$a_{\varepsilon} = \frac{1}{(\pi\varepsilon)^{d/4}\sqrt{\mu}}\Big(1 - \frac{\varepsilon}{8}\frac{\Delta\sqrt{\mu}}{\sqrt{\mu}} + o(\varepsilon)\Big).$$

As $K^{\varepsilon}_{\mu}[\phi] = a_{\varepsilon} H^{\varepsilon}_{c}[a_{\varepsilon}\mu\phi]$, we can again use the Laplace expansion, keeping only the terms of order 1 in ε and noticing that the scaling factors $(\pi\varepsilon)^{d/2}$ cancel:

$$\begin{aligned}K^{\varepsilon}_{\mu}[\phi] &= (\pi\varepsilon)^{d/2} a_{\varepsilon}\Big(a_{\varepsilon}\mu\phi + \frac{\varepsilon}{4}\Delta(a_{\varepsilon}\mu\phi) + o(\varepsilon)\Big) \\ &= \frac{1}{\sqrt{\mu}}\Big(1 - \frac{\varepsilon}{8}\frac{\Delta\sqrt{\mu}}{\sqrt{\mu}}\Big)\Big(\sqrt{\mu}\Big(1 - \frac{\varepsilon}{8}\frac{\Delta\sqrt{\mu}}{\sqrt{\mu}}\Big)\phi + \frac{\varepsilon}{4}\Delta(\sqrt{\mu}\phi)\Big) + o(\varepsilon) \\ &= \phi + \frac{\varepsilon}{4\sqrt{\mu}}(-\phi\Delta\sqrt{\mu} + \Delta(\sqrt{\mu}\phi)) + o(\varepsilon).\end{aligned}$$

As $\Delta(\sqrt{\mu}\phi) - \phi\Delta\sqrt{\mu} = \sqrt{\mu}\Delta(\phi) + 2\nabla\phi\cdot\nabla\sqrt{\mu}$ and $2\nabla\sqrt{\mu} = \sqrt{\mu}\nabla\log\mu$, we obtain

$$K^{\varepsilon}_{\mu}[\phi] = \phi + \frac{\varepsilon}{4}(\Delta\phi + \nabla\log\mu\cdot\nabla\phi) + o(\varepsilon),$$

which is sufficient to identify $\tilde{L} = \frac{1}{4}(\Delta + \nabla \log \mu \cdot \nabla)$ and to see that $L = -4\tilde{L}$ coincides with the expression given above (4.12).

5. The induced path metric

5.1. Definition and metrization of the weak-$*$ topology

Definition 5.1. Our metric is constructed in the following way:

(1) We call a path $(\mu_t)_{t\in[0,1]}$ admissible when the corresponding path $(\beta_t)_{t\in(0,1)} = (B(\mu_t))_{t\in(0,1)}$ belongs to the Sobolev space $\mathcal{H}^1((0,1);\mathcal{H}_c)$ (defined in Appendix C). We denote the set of admissible paths as P and write $P(\mu_0,\mu_1)$ for the subset of P with fixed start and end points $\mu_0,\mu_1 \in \mathcal{P}(X)$.

(2) The energy of a path $(\mu_t)_t \in P$ is defined as

$$E((\mu_t)_t) := \int_0^1 \tilde{\mathbf{g}}_{\mu_t}(\dot{\beta}_t,\dot{\beta}_t)\,\mathrm{d}t = \int_0^1 \mathbf{g}_{\mu_t}(\dot{\mu}_t,\dot{\mu}_t)\,\mathrm{d}t.$$

(These two integrals indeed coincide if $(\mu_t)_t \in P$ thanks to Theorem 4.12.)

(3) For $\mu_0,\mu_1 \in \mathcal{P}(X)$ we define

$$\mathsf{d}_S(\mu_0,\mu_1) := \Big(\inf_{(\mu_t)_t\in P(\mu_0,\mu_1)} E((\mu_t)_t)\Big)^{\frac{1}{2}}. \tag{5.1}$$

Theorem 4.8 tells us that P corresponds to $\mathcal{H}^1((0,1);\mathcal{B})$ and the paths $(\mu_t)_t$ are weak-$*$-continuous.

If the cost function is of class $\mathcal{C}^{(m,m)}$, then $\mathcal{C}^m$-perturbations are admissible, as we will see later in Corollary 5.10. Intuitively, $\mathcal{C}^m$-perturbations (which include combinations of vertical and horizontal perturbations) would seem sufficiently rich to describe $\dot{\mu}$ for all relevant trajectories, but we extend admissible paths to P to guarantee completeness. We now show that d_S is a distance (Theorem 5.2) which is geodesic (Theorem 5.4).

Theorem 5.2. *The function d_S is a metric on $\mathcal{P}(X)$ that metrizes the weak-$*$ topology. Further, for all $\mu_0,\mu_1 \in \mathcal{P}(X)$, $\beta_0 = B(\mu_0)$, $\beta_1 = B(\mu_1)$ there holds*

$$\sqrt{\frac{\varepsilon}{2}}\,\|\beta_0-\beta_1\|_{\mathcal{H}_c} \le \mathsf{d}_S(\mu_0,\mu_1) \le \frac{\pi}{2}\sqrt{\frac{\varepsilon}{2}C}\,\|\beta_0-\beta_1\|_{\mathcal{H}_c}$$

with $C = 1 + 2\exp(\frac{11}{2\varepsilon}\|c\|_\infty)$ as in Proposition 4.15.

Proof. It is easy to see that $\mathsf{d}_S(\mu_0,\mu_1) = \mathsf{d}_S(\mu_1,\mu_0) \ge 0$. We show the triangle inequality and the relation to the $\mathcal{H}_c$ norm. The latter also implies finiteness, positive definiteness, and the metrization of the weak-$*$ topology.

The proof of the triangle inequality is analogous to that of [13, Thm. 1]. Let $\mu,\nu,\eta \in \mathcal{P}(X)$, $(\mu_t)_t \in P(\mu,\nu)$, $(\nu_t)_t \in P(\nu,\eta)$ and denote $\beta_t^1 = B(\mu_t)$, $\beta_t^2 = B(\nu_t)$. Assume that

$\mathrm{d}_S(\mu,\nu) > 0$ and $\mathrm{d}_S(\nu,\eta) > 0$. Let $\tau \in (0,1)$ and set $\gamma_t = \mu_{t/\tau}$ for $t \in [0,\tau]$, $\gamma_t = \nu_{\frac{t-\tau}{1-\tau}}$ for $t \in [\tau,1]$. Using Theorem C.2 and the weak-$*$ continuity of $(\gamma_t)_t$, one can see that $(\gamma_t)_t \in P(\mu,\eta)$ and by simple rescaling there holds

$$\begin{aligned} E((\gamma_t)_t) &= \frac{1}{\tau}\int_0^1 \tilde{\mathbf{g}}_{\mu_t}(\dot\beta_t^1,\dot\beta_t^1)\,\mathrm{d}t + \frac{1}{1-\tau}\int_0^1 \tilde{\mathbf{g}}_{\nu_t}(\dot\beta_t^2,\dot\beta_t^2)\,\mathrm{d}t \\ &= \frac{1}{\tau}E((\mu_t)_t) + \frac{1}{1-\tau}E((\nu_t)_t). \end{aligned}$$

Setting $\tau = E((\mu_t)_t)^{\frac12}/(E((\mu_t)_t)^{\frac12} + E((\nu_t)_t)^{\frac12})$ gives

$$\frac{1}{\tau}E((\mu_t)_t) + \frac{1}{1-\tau}E((\nu_t)_t) = (E((\mu_t)_t)^{\frac12} + E((\nu_t)_t)^{\frac12})^2.$$

Taking the infimum over $(\mu_t)_t \in P(\mu,\nu)$ and $(\nu_t)_t \in P(\nu,\eta)$ yields the triangle inequality for d_S.

To show equivalence to the $\mathcal{H}_c$-norm, we use that $\frac{\varepsilon}{2}\|\dot\beta_t\|^2_{\mathcal{H}_c} \le \tilde{\mathbf{g}}_{\mu_t}(\dot\beta_t,\dot\beta_t) \le C\frac{\varepsilon}{2}\|\dot\beta_t\|^2_{\mathcal{H}_c}$ from Proposition 4.15.

For the lower bound, let $(\mu_t)_t \in P(\mu_0,\mu_1)$ and denote $\beta_t = B(\mu_t)$. Then Jensen's inequality, the norm estimate from Theorem C.1, and the fundamental theorem of calculus in $\mathcal{H}^1((0,1);\mathcal{H}_c)$ from Theorem C.2 show

$$\frac{2}{\varepsilon}\int_0^1 \tilde{\mathbf{g}}_{\mu_t}(\dot\beta_t,\dot\beta_t)\,\mathrm{d}t \ge \int_0^1 \|\dot\beta_t\|^2_{\mathcal{H}_c}\,\mathrm{d}t \ge \left\|\int_0^1 \dot\beta_t\,\mathrm{d}t\right\|^2_{\mathcal{H}_c} = \|\beta_1-\beta_0\|^2_{\mathcal{H}_c}.$$

This gives the lower bound when we take the infimum over all curves.

For the upper bound, let $(\beta_t)_t$ be the arc connecting β_0 to β_1 with $\|\beta_t\|_{\mathcal{H}_c} = 1$ parametrized by arc length of $(\beta_t)_t$ in $\mathcal{H}_c$. By Theorem 4.8, this is an admissible path. Proposition 4.15 and a basic geometry argument in the two-dimensional Hilbert space $\operatorname{span}\{\beta_0,\beta_1\}$ yield

$$\sqrt{\frac{2}{C\varepsilon}}\mathrm{d}_S(\mu_0,\mu_1) \le \left(\int_0^1 \|\dot\beta_t\|^2_{\mathcal{H}_c}\,\mathrm{d}t\right)^{\frac12} = \int_0^1 \|\dot\beta_t\|_{\mathcal{H}_c}\,\mathrm{d}t = 2\arcsin\left(\frac12\|\beta_1-\beta_0\|_{\mathcal{H}_c}\right),$$

which, in turn, is bounded by $\frac{\pi}{2}\|\beta_1-\beta_0\|_{\mathcal{H}_c}$.

Finally, for $\mu,\mu_n \in \mathcal{P}(X)$, the established estimate and Theorem 4.8 show that

$$\mathrm{d}_S(\mu_n,\mu) \to 0 \quad \Leftrightarrow \quad \|B(\mu_n)-B(\mu)\|_{\mathcal{H}_c} \to 0 \quad \Leftrightarrow \quad \mu_n \to \mu \text{ weak-}*. \qquad \blacksquare$$

Remark 5.3. Theorem 5.2 yields a metric equivalence between the flat norm on $\mathcal{H}_c$ and d_S. In particular, these two distances define the same class of absolutely continuous curves. However, the geometry induced by the metric tensor $\tilde{\mathbf{g}}$ differs from the flat one. We give one example of this: the shortest path (a.k.a. minimizing geodesic) between β_0 and β_1 in $\mathcal{B} \subseteq \mathcal{H}_c$ in the RKHS geometry is given by arc interpolation. This path lies entirely in the plane spanned by β_0, β_1. The equivalent holds for the corresponding path between α_0 and α_1 in $\mathcal{M}_+(X)$. Hence the support of the intermediate measures α_t (a fortiori of μ_t) is contained in the union of the support of the initial and final measure. In contrast, geodesics in d_S can have horizontal movement of mass that leaves the original support, for instance in the case of translations as in Section 5.4.

5.2. Existence of geodesics

Given the definition of d_S as infimization over the energy of paths, it is natural to ask whether geodesics in the sense of energy-minimizing paths exist. The set of admissible paths P has been chosen carefully to obtain existence of geodesics for d_S.

Theorem 5.4. *The variational problem* (5.1) *defining* d_S *has a minimizer. Hence there exist Riemannian geodesics for the metric* d_S.

The proof will follow the direct method of calculus of variations. Indeed, we first show lower semicontinuity of E on $\mathcal{H}^1((0,1);\mathcal{H}_c)$ by rewriting $E((\mu_t)_t)$ as a supremum of weakly lower semicontinuous problems in Lemma 5.6. Using that boundedness in E implies boundedness in the Sobolev norm, we then find that the weak cluster points of any minimizing sequence for the least energy problem (5.1) are minimizers, hence geodesic. This geodesic belongs to P, as $\mathcal{H}^1((0,1);\mathcal{B})$ is weakly closed.

Remark 5.5. We can obtain the existence of geodesics for d_S in the metric sense in a much simpler way by showing that in $(\mathcal{P}(X),\mathsf{d}_S)$ all pairs of points almost admit a midpoint (see [13, Def. 3.6]) and applying the Hopf–Rinow theorem [13, Lem. B.3]. For any $\mu_0,\mu_1\in\mathcal{P}(X)$ this would give a path $(\mu_t)_t$ in $\mathcal{P}(X)$, such that for all $t,s\in[0,1]$ it holds that

$$\mathsf{d}_S(\mu_t,\mu_s)=|t-s|\cdot\mathsf{d}_S(\mu_0,\mu_1).$$

However, it is not immediately clear whether the path $(\mu_t)_t$ belongs to $P(\mu_0,\mu_1)$ and whether $\mathsf{d}_S(\mu_0,\mu_1)=E((\mu_t)_t)$, i.e. whether $(\mu_t)_t$ is also a geodesic in the Riemannian sense as a minimizer of the energy in (5.1).

Lemma 5.6. *Let* $(\mu_t)_t\in P$ *and define* $\boldsymbol{\mu}\in\mathcal{M}([0,1]\times X)$ *via* $\langle\boldsymbol{\mu},\psi\rangle=\iint\psi(t,\cdot)\,\mathrm{d}\mu_t\,\mathrm{d}t$ *for any* $\psi\in\mathcal{C}([0,1]\times X)$. *The energy of the path* $(\mu_t)_t$ *can be rewritten as*

$$E((\mu_t)_t)=\frac{\varepsilon}{2}\int_0^1\|\dot\beta_t\|^2_{\mathcal{H}_c}\,\mathrm{d}t+2\varepsilon\cdot\sup_{\psi}\left\{\int_0^1\langle H_c[\psi(t,\cdot)A(\mu_t)],\dot\beta_t\rangle_{\mathcal{H}_c}\,\mathrm{d}t\right.\qquad(5.2)$$
$$\left.-\frac{1}{2}\int_0^1\langle\psi(t,\cdot),(\mathrm{id}-K_{\mu_t})\psi(t,\cdot)\rangle_{L^2(X,\mu_t)}\,\mathrm{d}t\right\},$$

where the supremum in ψ *is taken over any space of functions dense in* $L^2([0,1]\times X,\boldsymbol{\mu})$, *e.g. the jointly continuous functions over* $[0,1]\times X$.

Proof. It is a well-known fact in convex analysis that if T is a symmetric, coercive, and continuous operator on a Hilbert space $\mathcal{H}$, it holds that

$$\frac{1}{2}\langle x, T^{-1}x\rangle_{\mathcal{H}} = \sup_y \langle x, y\rangle_{\mathcal{H}} - \frac{1}{2}\langle y, Ty\rangle_{\mathcal{H}} \tag{5.3}$$

for all $x \in \mathcal{H}$. This is Fenchel–Legendre duality, or could also be seen as a corollary of the Lax–Milgram theorem [14, Cor. 5.8]. Continuity of T allows the restriction of the supremum to a dense subset of $\mathcal{H}$. We apply (5.3) to

- the space $\mathcal{H} = \{\psi \in L^2([0,1] \times X, \boldsymbol{\mu}) \mid \langle \mu_t, \psi(t,\cdot)\rangle = 0 \text{ a.e.}\}$, i.e. the orthogonal complement of functions which are μ_t-a.e. constant in space at almost all $t \in [0,1]$,
- the operator $T[\psi](t,y) = (\mathrm{id} - K_{\mu_t})[\psi(t,\cdot)](y)$,
- the point $x = (a_{\mu_t}\dot{\beta}_t)_t \in \mathcal{H}$.

The operator T is coercive and continuous on $\mathcal{H}$ by Theorem 4.3. In this notation,

$$E((\mu_t)_t) = \int_0^1 \tilde{\mathbf{g}}_{\mu_t}(\dot{\beta}_t, \dot{\beta}_t)\,\mathrm{d}t = \frac{\varepsilon}{2}\int_0^1 \|\dot{\beta}_t\|^2_{\mathcal{H}_c}\,\mathrm{d}t + \varepsilon\langle x, T^{-1}x\rangle_{\mathcal{H}}.$$

We then use that

$$\langle a_{\mu_t}\dot{\beta}_t, \psi(t,\cdot)\rangle_{L^2(X,\mu_t)} = \langle \psi(t,\cdot)a_{\mu_t}\mu_t, \dot{\beta}_t\rangle = \langle H_c[\psi(t,\cdot)A(\mu_t)], \dot{\beta}_t\rangle_{\mathcal{H}_c}.$$

With $\langle \dot{\beta}_t, \beta_t\rangle = 0$ and $(\mathrm{id} - K_{\mu_t})[\mathbb{1}_X] = 0$, one sees that the expression on the right-hand side of (5.2) gives the same value on functions whose difference is constant in space for almost all times t. Hence the supremum remains the same when taken on $\mathcal{H}$ or $L^2([0,1] \times X; \boldsymbol{\mu})$. ∎

Proof of Theorem 5.4. We apply the direct method of calculus of variations. Let $(\mu_t^n)_t \in P(\mu_0, \mu_1)$ with corresponding $(\beta_t^n)_t \in \mathcal{H}^1((0,1); \mathcal{H}_c)$ be such that

$$\mathrm{d}_S(\mu_0, \mu_1)^2 = \lim_{n\to\infty}\int_0^1 \tilde{\mathbf{g}}_{\mu_t^n}(\dot{\beta}_t^n, \dot{\beta}_t^n)\,\mathrm{d}t.$$

From $\|\beta_t^n\|_{\mathcal{H}_c} = 1$ and Proposition 4.15 we see that the sequence of paths $(\beta_t^n)_t$ must be bounded in $\mathcal{H}^1((0,1); \mathcal{H}_c)$. Hence, up to taking a subsequence, $(\beta_t^n)_t$ converges weakly in $\mathcal{H}^1((0,1); \mathcal{H}_c)$ to some limit path $(\beta_t)_t$. It remains to check that $\beta_t \in \mathcal{B}$ for all $t \in (0,1)$ and that the corresponding path of $\mu_t = B^{-1}(\beta_t)$ satisfies $E((\mu_t)_t) = \mathrm{d}_S(\mu_0, \mu_1)^2$.

Note that the sequence (β_t^n) is uniformly equi-continuous: For any $n \in \mathbb{N}$, $s < t \in [0,1]$ set $v = (\beta_t^n - \beta_s^n)/\|\beta_t^n - \beta_s^n\|_{\mathcal{H}_c}$. Then Theorem C.2 and the Cauchy–Schwarz inequality give

$$\begin{aligned}\|\beta_t^n - \beta_s^n\|_{\mathcal{H}_c} = \int_s^t \langle \dot{\beta}_\tau^n, v\rangle\,\mathrm{d}\tau &= \langle (\dot{\beta}_\tau^n)_\tau, v \cdot \mathbb{1}_{[s,t]}\rangle_{L^2([0,1];\mathcal{H}_c)} \\ &\le \|(\beta_\tau^n)_\tau\|_{\mathcal{H}^1((0,1);\mathcal{H}_c)} \cdot \sqrt{|t-s|}.\end{aligned}$$

Thus, again up to taking a subsequence, we can assume that (β_t^n) converges uniformly, hence the limit satisfies $\beta_t \in \mathcal{B}$ for all $t \in (0,1)$.

We now estimate the energy. By lower semicontinuity of the norm for weak convergence, we have directly

$$\int_0^1 \|\dot{\beta}_t\|_{\mathcal{H}_c}^2 \,\mathrm{d}t \leq \liminf_{n\to\infty} \int_0^1 \|\dot{\beta}_t^n\|_{\mathcal{H}_c}^2 \,\mathrm{d}t.$$

Thus by Lemma 5.6, using $\alpha_t^n = A(\mu_t^n) = a_{\mu_t^n}\mu_t^n$ and rewriting the operator $K_{\mu_t^n}$ using $k_{\mu_t^n} = k_c \cdot a_{\mu_t^n} \otimes a_{\mu_t^n}$, to conclude the proof it suffices to show that for all $\psi \in \mathcal{C}([0,1] \times X)$ we have

$$\begin{aligned}
\liminf_{n\to\infty} &\int_0^1 \langle H_c[\psi(t,\cdot)\alpha_t^n], \dot{\beta}_t^n\rangle_{\mathcal{H}_c} \,\mathrm{d}t - \frac{1}{2}\int_0^1 \int_X \psi(t,x)^2 \,\mathrm{d}\mu_t^n(x)\,\mathrm{d}t \\
&+ \frac{1}{2}\int_0^1 \iint_{X\times X} k_c(x,y)\psi(t,x)\psi(t,y)\,\mathrm{d}\alpha_t^n(x)\,\mathrm{d}\alpha_t^n(y)\,\mathrm{d}t \\
\geq &\int_0^1 \langle H_c[\psi(t,\cdot)\alpha_t], \dot{\beta}_t\rangle_{\mathcal{H}_c} \,\mathrm{d}t - \frac{1}{2}\int_0^1 \int_X \psi(t,x)^2 \,\mathrm{d}\mu_t(x)\,\mathrm{d}t \\
&+ \frac{1}{2}\int_0^1 \iint_{X\times X} k_c(x,y)\psi(t,x)\psi(t,y)\,\mathrm{d}\alpha_t(x)\,\mathrm{d}\alpha_t(y)\,\mathrm{d}t.
\end{aligned}$$

As β_t^n converges to β_t, we know from Theorem 4.8 that $\mu_t^n \to \mu_t$, $\alpha_t^n \to \alpha_t$ weak-$*$ in $\mathcal{M}(X)$. In addition, we can bound the mass of α_t^n and the supremum of $\psi(t,\cdot)$ uniformly in t and n, and so the second and third integrals are easy to pass to the limit. For the first one note that

$$(H_c[\psi(t,\cdot)\alpha_t^n])_t \to (H_c[\psi(t,\cdot)\alpha_t])_t.$$

The convergence is pointwise in t thanks to Lemma B.2, and also in norm in $L^2([0,1];\mathcal{H}_c)$ thanks to the Lebesgue dominated convergence theorem. Since the first integral can be rewritten as

$$\int_0^1 \langle H_c[\psi(t,\cdot)\alpha_t^n], \dot{\beta}_t^n\rangle_{\mathcal{H}_c}\,\mathrm{d}t = \langle (H_c[\psi(t,\cdot)\alpha_t^n])_t, (\dot{\beta}_t^n)_t\rangle_{L^2([0,1];\mathcal{H}_c)},$$

convergence to the desired limit holds as a pairing between weakly and strongly convergent sequences in $L^2([0,1];\mathcal{H}_c)$ (see Lemma A.2). ■

5.3. Absolutely continuous curves

Let us first recall the definition of absolutely continuous curves: in Hilbert spaces they correspond to a.e. differentiable curves for which the fundamental theorem of calculus holds (Theorem C.2).

Definition 5.7 ([4, Def. 1.1.1]). Let $(\mathcal{X}, \mathrm{d})$ be a complete metric space and let $p \in [1,\infty]$. A curve $\gamma\colon (a,b) \to \mathcal{X}$ defined on a real interval belongs to $\mathrm{AC}^p((a,b);\mathcal{X})$, if there exists a function $m \in L^p(a,b)$ such that for all $a < s \le t \le b$ it holds that

$$\mathsf{d}(\gamma(s),\gamma(t)) \le \int_s^t m(\tau)\,\mathrm{d}\tau.$$

Due to the equivalence between the norm on $\mathcal{H}_c$ and the metric d_S from Theorem 5.2, and as $\mathcal{H}^1((0,1);\mathcal{H}_c) = \mathrm{AC}^2((0,1);\mathcal{H}_c)$ from Theorem C.2, if $\beta_t = B(\mu_t)$ we have

$$(\mu_t)_t \text{ is admissible} \quad \Leftrightarrow \quad (\beta_t)_t \in \mathrm{AC}^2((0,1);\mathcal{H}_c) \quad \Leftrightarrow \quad (\mu_t)_t \in \mathrm{AC}^2((0,1);\mathsf{d}_S).$$

Building on Lemma 4.13 we can characterize the set of admissible paths in terms of differentiability of the path of measures $(\mu_t)_t$ in duality with each respective RKHS $\mathcal{H}_{\mu_t}$ and regularity of the path of self-transport potentials f_{μ_t,μ_t} in the space of continuous functions as opposed to the less intuitive space $\mathcal{H}_c$.

Proposition 5.8. *Let $(\mu_t)_t$ be a weak-$*$-continuous path in $\mathcal{P}(X)$. Denote $f_{t,t} = f_{\mu_t,\mu_t}$ and $\beta_t = B(\mu_t)$. Then $(\mu_t)_t$ is an admissible path for our metric, i.e. $(\beta_t)_t \in \mathcal{H}^1((0,1);\mathcal{H}_c)$, or equivalently $(\mu_t)_t \in \mathrm{AC}^2((0,1);(\mathcal{P}(X),\mathsf{d}_S))$, if and only if the following conditions are satisfied:*

1. *$(\mu_t)_t$ is strongly differentiable, i.e. $\frac{\mu_t-\mu_s}{t-s} \to \dot\mu_s$ in norm in $\mathcal{H}^*_{\mu_s}$ at time $t = s$ for almost all $s \in [0,1]$.*
2. *$\int_0^1 \mathbf{g}_{\mu_t}(\dot\mu_t,\dot\mu_t)\,\mathrm{d}t < \infty$.*
3. *$(f_{t,t})_t \in \mathrm{AC}^1((0,1);\mathcal{C}(X))$.*

Further, when these are satisfied the following integration by parts formula holds for all functions of the form $\psi_t = \exp(f_{t,t}/\varepsilon)\cdot\phi_t$ with $\phi \in \mathcal{C}^\infty_c((0,1);\mathcal{H}_c)$:

$$\int_0^1 \langle \mu_t, \partial_t\psi_t\rangle\,\mathrm{d}t = -\int_0^1 \langle \dot\mu_t,\psi_t\rangle\,\mathrm{d}t. \tag{5.4}$$

In this equation, $\psi_t \in \mathcal{H}_{\mu_t}$ for every t by Lemma 4.7, so that the pairing $\langle\dot\mu_t,\psi_t\rangle$ is well defined.

Proof. The proof is divided into three parts.

Direct implication. Let $(\beta_t)_t \in \mathcal{H}^1((0,1);\mathcal{H}_c)$. Then $\dot\beta_t$ exists in $\mathcal{H}_c$ for almost all $t \in (0,1)$. From Lemma 4.13, $(\mu_t)_t$ is strongly differentiable in $\mathcal{H}^*_{\mu_s}$ at time $t = s$ for almost all $s \in (0,1)$. This is condition (1). By Proposition 4.15, we have

$$\int_0^1 \mathbf{g}_{\mu_t}(\dot\mu_t,\dot\mu_t)\,\mathrm{d}t = \int_0^1 \tilde{\mathbf{g}}_{\mu_t}(\dot\beta_t,\dot\beta_t)\,\mathrm{d}t \le C\frac{\varepsilon}{2}\|(\beta_t)_t\|^2_{\mathcal{H}^1((0,1);\mathcal{H}_c)} < \infty,$$

proving condition (2). Due to the embedding $\mathcal{H}_c \to \mathcal{C}(X)$ and equi-boundedness of the potentials $f_{t,t}$, we obtain for $s \le t$,

$$\|f_{t,t} - f_{s,s}\|_\infty \le C\cdot\|\beta_t - \beta_s\|_\infty \le C\cdot\|\beta_t-\beta_s\|_{\mathcal{H}_c} \le C\cdot\int_s^t \|\dot\beta_\tau\|_{\mathcal{H}_c}\,\mathrm{d}\tau$$

for some $C \in (0,\infty)$, where the last inequality is due to Theorem C.1. As $(\|\dot\beta_t\|)_t \in L^2((0,1)) \subset L^1((0,1))$, this proves condition (3).

Converse implication. We show that $(\beta_t)_t \in \mathcal{H}^1((0,1);\mathcal{H}_c)$ using the characterization of Theorem C.2. Specifically, we show that $(\beta_t)_t$ is strongly differentiable almost everywhere, $(\dot\beta_t)_t \in L^2((0,1);\mathcal{H}_c)$, and that $\beta_t - \beta_s = \int_s^t \dot\beta_\tau \,\mathrm{d}\tau$ for all $t, s \in (0,1)$.

Condition (3) gives $\|(f_{t,t} - f_{s,s})/(t-s)\|_\infty \le \frac{1}{t-s}\int_s^t g(\tau)\,\mathrm{d}\tau$ for some $g \in L^1((0,1);\mathbb{R})$. By the Lebesgue differentiation theorem [65, Thm. 7.10], the right-hand side converges to $g(s)$ as $t \to s$ for almost every $s \in (0,1)$. In particular, the difference quotients are bounded in the limit for almost all s. Differentiability of $(\beta_t)_t$ almost everywhere follows from Lemma 4.13 and condition (1). By Theorem 4.12 and Proposition 4.15 we have $\frac{\varepsilon}{2}\|\dot\beta_s\|^2_{\mathcal{H}_c} \le \tilde{\mathbf{g}}_{\mu_s}(\dot\beta_s,\dot\beta_s) = \mathbf{g}_{\mu_s}(\dot\mu_s,\dot\mu_s)$ almost everywhere. Thus condition (2) yields $(\dot\beta_t)_t \in L^2((0,1);\mathcal{H}_c)$. It only remains to check the compatibility condition: for all t, s,

$$\beta_t - \beta_s = \int_s^t \dot\beta_\tau \,\mathrm{d}\tau \tag{5.5}$$

as a Bochner integral in $\mathcal{H}_c$ (see [40, Sect. 2.1]). The integral in the right-hand side of (5.5) exists in $\mathcal{H}_c$; see Theorem C.1. To identify the left-hand side, as $(f_{t,t})_t \in \mathrm{AC}^1((0,1);\mathcal{C}(X))$ and the potentials are equi-bounded, we obtain $(\beta_t)_t \in \mathrm{AC}^1((0,1);\mathcal{C}(X))$. In particular, for any $x \in X$ the curve $(\beta_t(x))_t$ belongs to $\mathrm{AC}^1((0,1);\mathbb{R})$ and thus (5.5), evaluated at $x \in X$, holds as a classical integral. We conclude to (5.5) as Bochner integral in $\mathcal{H}_c$ as the evaluation maps are continuous in $\mathcal{H}_c$.

Integration by parts. Let $(\beta_t)_t \in \mathcal{H}^1((0,1);\mathcal{H}_c)$. Let $(\psi_t)_t$ and $(\phi_t)_t$ be as in (5.4). By our assumptions on $(\beta_t)_t$, the rescaled function $\psi_t = \phi_t/\beta_t$ is differentiable in $\mathcal{C}(X)$ for almost every t. Lemma 4.13 yields

$$\begin{aligned}\int_0^1 \langle \dot\mu_t, \psi_t\rangle \,\mathrm{d}t &= \int_0^1 \langle \dot\beta_t, \beta_t\cdot\psi_t + \beta_t\cdot K_{\mu_t}[\psi_t]\rangle_{\mathcal{H}_c}\,\mathrm{d}t\\ &= \int_0^1 \langle \dot\beta_t, \phi_t + H_c[\phi_t/\beta_t^2\cdot\mu_t]\rangle_{\mathcal{H}_c}\,\mathrm{d}t.\end{aligned}$$

Lemma C.3 gives $\int_0^1 \langle\dot\beta_t,\phi_t\rangle_{\mathcal{H}_c}\,\mathrm{d}t = -\int_0^1\langle\beta_t,\partial_t\phi_t\rangle_{\mathcal{H}_c}\,\mathrm{d}t$. Further, we have

$$\langle\beta_t,\partial_t\phi_t\rangle_{\mathcal{H}_c} = \langle\alpha_t,\partial_t\phi_t\rangle = \langle\mu_t,(\partial_t\phi_t)/\beta_t\rangle.$$

Using $\langle\dot\beta_t, H_c[\phi_t/\beta_t^2\cdot\mu_t]\rangle_{\mathcal{H}_c} = \langle\mu_t,\phi_t\cdot\dot\beta_t/\beta_t^2\rangle$ we can compute the remaining part. Together this gives

$$\int_0^1\langle\dot\mu_t,\psi_t\rangle = -\int_0^1\langle\mu_t,(\partial_t\phi_t)/\beta_t - \phi_t\cdot\dot\beta_t/\beta_t^2\rangle\,\mathrm{d}t = -\int_0^1\langle\mu_t,\partial_t\psi_t\rangle\,\mathrm{d}t. \qquad \blacksquare$$

The previous proposition gives a full characterization of the set of admissible paths and requires a light notion of differentiability of the path $(\mu_t)_t$ together with a robust regularity assumption on the path of self-transport potentials $(f_{t,t})_t$ in $\mathcal{C}(X)$.

In the differentiable setting, Proposition 5.9 below gives a simpler way to show that a path is admissible. In it we utilize a stronger regularity assumption on $(\mu_t)_t$ by requiring $\dot\mu_t \in \mathcal{C}^m(X)^*$ and avoid the extra assumption on the potentials.

Proposition 5.9. *For $m \geq 0$, in the setting of Assumption 1, there exists $C > 0$ depending on c, ε, and m such that for all $\mu_0, \mu_1 \in \mathcal{P}(X)$,*

$$\mathsf{d}_S(\mu_0, \mu_1) \leq C \cdot \|\mu_1 - \mu_0\|_{\mathcal{C}^{m,*}}. \tag{5.6}$$

Notably, $(\mu_t)_t \in \mathrm{AC}^2((0,1); (\mathcal{C}^m(X)^*, \|\cdot\|_{\mathcal{C}^{m,*}}))$ yields $(\mu_t)_t \in \mathrm{AC}^2((0,1); (\mathcal{P}(X), \mathsf{d}_S))$ and any such path is admissible for d_S.

Proof. We clearly only need to prove (5.6). The vertical interpolation $\mu_t = (1-t)\mu_0 + t\mu_1$ is an admissible path with $\dot{\mu}_t = \mu_1 - \mu_0$. We have, using successively the estimates (4.6) and (4.4),

$$\mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t) \leq C_1 \|\mu_1 - \mu_0\|^2_{\mathcal{H}^*_{\mu,0}} \leq C_2 \|\mu_1 - \mu_0\|^2_{\mathcal{C}^{m,*}},$$

for some $C_1, C_2 \geq 0$, where the constant C_2 may depend on m but neither one depends on μ. Integrating this between 0 and 1 yields the estimate as the left-hand side becomes an upper bound for $\mathsf{d}_S(\mu_0, \mu_1)^2$ by definition. ∎

We can now show that paths which essentially follow a $\mathcal{C}^m$-perturbation for time $[0,1]$ are admissible paths for d_S. For these paths the metric tensor is explicitly given by the Hessian of the Sinkhorn divergence. This provides the link between Section 3 and Section 5.1.

Corollary 5.10. *In the setting of Assumption 1 for some $m \geq 0$, let $(\mu_t)_t$ be a path that is continuously differentiable from $(0,1)$ to the space $\mathcal{C}^m(X)^*$ endowed with the weak-$*$ topology, such that the derivative extends weak-$*$ continuously to the end points. Then it is $\sup_{t\in[0,1]} \|\dot{\mu}_t\|_{\mathcal{C}^{m,*}} < \infty$, and we have $(\mu_t)_t \in \mathrm{AC}^2((0,1); (\mathcal{C}^m(X)^*, \|\cdot\|_{\mathcal{C}^{m,*}}))$. In particular, $(\mu_t)_t$ is an admissible path for d_S.*

Proof. There holds $\sup_{t\in[0,1]} \|\dot{\mu}_t\|_{\mathcal{C}^{m,*}} < \infty$, as weak-$*$-continuity gives boundedness of $\|\dot{\mu}_{t_n}\|_{\mathcal{C}^{m,*}}$ along any convergent sequence $t_n \to t$ and $[0,1]$ is compact. Then by [40, Prop. 3.8, (iv) $\Rightarrow$ (ii)] with the choice of separating subset $\mathcal{C}^m(X) \subseteq (\mathcal{C}^m(X)^*)^*$, we obtain that $(\mu_t)_t$ is of class $\mathrm{AC}^2((0,1); (\mathcal{C}^m(X)^*, \|\cdot\|_{\mathcal{C}^{m,*}}))$ if we show that for any $\phi \in \mathcal{C}^m(X)$, the real-valued function $t \mapsto \langle \mu_t, \phi\rangle$ is continuously differentiable with derivative $\langle \dot{\mu}_t, \phi\rangle$ and $(\dot{\mu}_t)_t \in L^2((0,1); \mathcal{C}^m(X)^*)$. All of these are already established. ∎

Remark 5.11 (Comparison between d_S, Wasserstein, and Hellinger–Kantorovich). In the setting of Assumption 1 with $m = 1$, Proposition 5.9 yields a $C > 0$ such that

$$\mathsf{d}_S(\mu_0, \mu_1) \leq C \cdot \|\mu_1 - \mu_0\|_{\mathcal{C}^{1,*}} \leq C \cdot W_1(\mu_0, \mu_1) \tag{5.7}$$

for any $\mu_0, \mu_1 \in \mathcal{P}(X)$. Here, W_p denotes the p-Wasserstein distance [68, Chap. 5] for $p \geq 1$. Thus 2-absolutely continuous paths in W_1 (hence W_p for any $p \in [1, \infty)$) are admissible paths for our metric. The converse does not hold, as shown in Example 5.12 below.

In the Euclidean setting the formal Riemannian tensor of the Hellinger–Kantorovich distance [20, 39, 47] is given by

$$\mathbf{g}_\mu^{\mathrm{HK}}(\dot\mu,\dot\mu) = \int_X \Big(|\nabla\psi(x)|^2 + \frac{1}{4}|\zeta(x)|^2\Big)\,\mathrm{d}\mu(x), \quad \dot\mu = -\operatorname{div}(\mu\nabla\psi) + \mu\cdot\zeta, \tag{5.8}$$

where $\dot\mu$ is explicitly decomposed into a horizontal and a vertical component. This induces a Riemannian distance on non-negative Radon measures with variable mass. The geodesic restriction to probability measures, the so-called spherical Hellinger–Kantorovich distance SHK, is studied in [42]. In particular, $\mathbf{g}_\mu^{\mathrm{HK}}(\dot\mu,\dot\mu)$ is finite, even when $\dot\mu$ contains a vertical component, which enables the "teleportation" of mass from one point to another without moving it through the base space. This is similar to the metric tensor $\mathbf{g}_\mu(\dot\mu,\dot\mu)$ of d_S and different from the tensor for the standard Wasserstein distance, (1.3). The same C as in (5.7) gives

$$\mathbf{g}_\mu(\dot\mu,\dot\mu) \le C^2\cdot\|\dot\mu\|^2_{\mathcal{C}^{1,*}} \le 16C^2\cdot\Big(\|\nabla\phi\|_{L^2(X,\mu)} + \frac{1}{4}\|\zeta\|_{L^2(X,\mu)}\Big)^2 \le 32C^2\cdot\mathbf{g}_\mu^{\mathrm{HK}}(\dot\mu,\dot\mu)$$

for all perturbations $\dot\mu$ as in (5.8). Hence we have the generalization of (5.7),

$$\mathsf{d}_S(\mu_0,\mu_1) \le 4\sqrt{2}C\cdot\mathsf{SHK}(\mu_0,\mu_1).$$

However, beyond this similarity, there are several significant differences: $\mathbf{g}_\mu^{\mathrm{HK}}$ is spatially local in the functions $\nabla\psi$ and ζ, which parametrize $\dot\mu$, whereas $\mathbf{g}_\mu$ from Definition 4.1 is non-local due to the operators K_μ and H_μ and does not rely on an explicit decomposition of $\dot\mu$ into vertical and horizontal terms. For $\mathbf{g}_\mu^{\mathrm{HK}}$, the cost of teleportation is independent of the distance between the start and end points in base space. In fact, when the supports of μ_0 and μ_1 are separated by more than $\pi/2$, their geodesic with respect to $\mathbf{g}_\mu^{\mathrm{HK}}$ solely relies on the vertical teleportation term and becomes equal to that of the Hellinger distance. In contrast, for $c(x,y) = \|x-y\|^2$ we find below (Section 6.2) that in $\mathbf{g}_\mu$ teleportation can become exponentially expensive at large distances, whereas at small distances horizontal and vertical perturbations have comparable cost.

Example 5.12. The path given by $\mu_t = \frac{1}{2}(\delta_{-\sqrt{t}} + \delta_{\sqrt{t}})$ on $X = [-1,1] \subseteq \mathbb{R}$ is continuously differentiable in $\mathcal{C}^2(X)^*$ endowed with the norm topology for $t \in (0,1]$ and weak-$*$-differentiable in $\mathcal{C}^2(X)^*$ at $t = 0$. Indeed, we have $\dot\mu_t = \frac{1}{2\sqrt{t}}(\delta'_{\sqrt{t}} - \delta'_{-\sqrt{t}})$, thus the fundamental theorem of calculus yields $\|\dot\mu_t\|_{\mathcal{C}^{2,*}} \le 1$ for any $t \in (0,1]$. At $t = 0$, we can actually compute $\langle\dot\mu_0,\phi\rangle = \phi''(0)$. Proposition 5.9 yields that $(\mu_t)_t$ is an admissible path for d_S when $c \in \mathcal{C}^{(2,2)}(X\times X)$, such as for the quadratic cost.

On the other hand, as $\dot\mu_t = -\operatorname{div}(\mu_t v_t)$ with $|v_t(x)| = 1/\sqrt{t}$, we have $\mathbf{g}^0_{\mu_t}(\dot\mu_t,\dot\mu_t) = \mathbf{g}^{\mathrm{HK}}_{\mu_t}(\dot\mu_t,\dot\mu_t) = 1/t$. It shows that $(\mu_t)_t$ has infinite action on $[0,1]$ both for the classical optimal transport geometry and the Hellinger–Kantorovich geometry.

5.4. Mean estimate for the squared Euclidean distance

In this section, $c(x,y) = \|x-y\|^2$ is the squared Euclidean distance and we work with measures in $\mathbb{R}^d$ with bounded support. For such a measure μ, we write $m = \langle\mu,\mathrm{id}\rangle \in \mathbb{R}^d$

for its mean, and $\bar{\mu} = (\mathrm{id} - m)_{\#}\mu$ for the centered measure. As recalled in the introduction, it is well understood that for classical optimal transport the cost OT_0 between μ_0 and μ_1 decomposes as

$$\mathrm{OT}_0(\mu_0, \mu_1) = \|m_1 - m_0\|^2 + \mathrm{OT}_0(\bar{\mu}_0, \bar{\mu}_1).$$

The same proof carries through for the Sinkhorn divergence (Proposition 5.13). We then extend this orthogonal decomposition result to the metric tensor in Proposition 5.14 and to the distance d_S in Theorem 5.16, which in particular implies that translations $\mu_t = (\mathrm{id} + t \cdot v)_{\#}\mu_0$ for $v \in \mathbb{R}^d$ are geodesics for d_S.

Proposition 5.13. *On $\mathbb{R}^d$ let $c(x, y) = \|x - y\|^2$. Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ with bounded support and means $m_0, m_1 \in \mathbb{R}^d$. Their centered versions $\bar{\mu}_0$, $\bar{\mu}_1$ satisfy*

$$S_\varepsilon(\mu_0, \mu_1) = \|m_1 - m_0\|^2 + S_\varepsilon(\bar{\mu}_0, \bar{\mu}_1).$$

Proof. The proof is basically the same as for plain optimal transport, where $\varepsilon = 0$. For the squared Euclidean distance, optimal couplings for OT_ε remain optimal under translations in the following sense: for all $\pi \in \Pi(\mu_0, \mu_1)$ it holds that $(\mathrm{id} - m_0, \mathrm{id} - m_1)_{\#}\pi \in \Pi(\bar{\mu}_0, \bar{\mu}_1)$ and

$$\begin{aligned}
\langle (\mathrm{id}-m_0, \mathrm{id}-m_1)_{\#}\pi, c\rangle &= \int_{X\times X} \|x - m_0 - (y - m_1)\|^2 \,\mathrm{d}\pi(x, y) \\
&= \int_{X\times X} \|x-y\|^2 - 2\langle x-y, m_0-m_1\rangle + \|m_1-m_0\|^2 \,\mathrm{d}\pi(x, y) \\
&= \langle \pi, c\rangle - \|m_1 - m_0\|^2.
\end{aligned}$$

The KL-divergence is invariant under such shifts, i.e.

$$\mathrm{KL}((\mathrm{id} - m_0, \mathrm{id} - m_1)_{\#}\pi \mid (\mathrm{id} - m_0)_{\#}\mu_0 \otimes (\mathrm{id} - m_1)_{\#}\mu_1) = \mathrm{KL}(\pi \mid \mu_0 \otimes \mu_1).$$

Therefore, minimizing over π yields $\mathrm{OT}_\varepsilon(\mu_0, \mu_1) = \|m_1 - m_0\|^2 + \mathrm{OT}_\varepsilon(\bar{\mu}_0, \bar{\mu}_1)$. Analogously one sees that $\mathrm{OT}_\varepsilon(\mu_i, \mu_i) = \mathrm{OT}_\varepsilon(\bar{\mu}_i, \bar{\mu}_i)$ for $i = 0, 1$. From the definition of S_ε in (1.6) we obtain the result. ∎

We now extend this decomposition to the level of the metric tensor. If $v \in \mathbb{R}^d$, the uniform translation of a measure μ by v corresponds to the derivative $\dot{\mu} = -\operatorname{div}(v\mu) \in \mathcal{H}^*_{\mu,0}$: this is the continuity equation. For $\mu \in \mathcal{P}(X)$ let $\mathcal{S}_\mu := \{-\operatorname{div}(v\mu) \mid v \in \mathbb{R}^d\} \subseteq \mathcal{H}^*_{\mu,0}$ be the set of constant shift perturbations of μ. Denote the orthogonal projection onto $\mathcal{S}_\mu$ with respect to the inner product $\mathbf{g}_\mu$ by

$$P_{\mathcal{S}_\mu} \colon \mathcal{H}^*_{\mu,0} \to \mathcal{S}_\mu.$$

Proposition 5.14. *Let $X \subseteq \mathbb{R}^d$ be compact, $c(x, y) = \|x - y\|^2$. Let $\mu \in \mathcal{P}(X)$ and let $\dot{\mu} \in \mathcal{H}^*_{\mu,0}$. Then*

(1) *the mean $\dot{m} = \langle \dot{\mu}, \mathrm{id}\rangle$ exists in $\mathbb{R}^d$ and is characterized by $P_{\mathcal{S}_\mu}(\dot{\mu}) = -\operatorname{div}(\dot{m}\mu)$,*

(2) *there holds* $\mathbf{g}_\mu(\dot\mu, \dot\mu) = \|\dot m\|^2 + \mathbf{g}_\mu(\dot\mu + \operatorname{div}(\dot m \mu), \dot\mu + \operatorname{div}(\dot m\mu))$.

Proof. For $v \in \mathbb{R}^d$ denote $\dot\mu^v = -\operatorname{div}(v\mu)$. As a distribution of order one it belongs to $\mathcal{H}_\mu^*$ (see Remark 4.4). We compute

$$
\begin{aligned}
H_\mu[\dot\mu^v](x) &= \int_X \langle v, \nabla_y k_\mu(x,y)\rangle \, \mathrm{d}\mu(y) \\
&= \frac{1}{\varepsilon}\int_X \langle v, \nabla f_{\mu,\mu}(y) - 2y + 2x\rangle k_\mu(x,y)\, \mathrm{d}\mu(y) \\
&= \frac{1}{\varepsilon} K_\mu[\langle v, \nabla f_{\mu,\mu} - 2\,\mathrm{id}\rangle](x) + \frac{2}{\varepsilon}\langle v, x\rangle. \qquad (5.9)
\end{aligned}
$$

Differentiating the relation $f_{\mu,\mu}(y) = T_\varepsilon(f_{\mu,\mu},\mu)(y)$ in y yields

$$
2y - \nabla f_{\mu,\mu}(y) = \int_X k_\mu(y,z) 2z \, \mathrm{d}\mu(z) = 2K_\mu[\mathrm{id}](y). \qquad (5.10)
$$

Thus $\langle v, \nabla f_{\mu,\mu} - 2\,\mathrm{id}\rangle = -2K_\mu[\langle v, \mathrm{id}\rangle]$. Applying this in (5.9), we deduce the identity $\frac{\varepsilon}{2}(\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot\mu^v] = \langle v, \mathrm{id}\rangle$ in $\mathcal{H}_\mu/\mathbb{R}$. In particular, the function $\langle v, \mathrm{id}\rangle$ belongs to $\mathcal{H}_\mu/\mathbb{R}$ for any $v \in \mathbb{R}^d$. Since every coordinate function $x \mapsto x_i$ belongs to $\mathcal{H}_\mu/\mathbb{R}$, the mean $\dot m = \langle \dot\mu, \mathrm{id}\rangle$ is well defined as an element of $\mathbb{R}^d$.

Given the expression of the metric tensor (Definition 4.1), we see that

$$
\mathbf{g}_\mu(\dot\mu, \dot\mu^v) = \langle \dot\mu, \langle v, \mathrm{id}\rangle\rangle = \langle v, \dot m\rangle = \mathbf{g}_\mu(\dot\mu^{\dot m}, \dot\mu^v).
$$

Therefore, $\dot\mu^{\dot m}$ is the orthogonal projection of $\dot\mu$ to $\mathcal{S}_\mu$ with respect to the metric tensor $\mathbf{g}_\mu$. The second point of the proposition is a direct consequence of the Pythagorean identity for the inner product $\mathbf{g}_\mu$. ∎

Remark 5.15. For $\mathcal{C}^m$-perturbations (Definition 3.1) for which the metric tensor is linked to the Sinkhorn divergence via Theorem 3.4, the proof of Proposition 5.14 can be done directly as an application of Proposition 5.13.

The orthogonal decomposition at the level of the metric tensor from Proposition 5.14 can be integrated and extended to the level of the distance d_S.

Theorem 5.16. *On* $\mathbb{R}^d$ *let* $c(x,y) = \|x-y\|^2$. *Let* $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^d)$ *with bounded support and means* $m_0, m_1 \in \mathbb{R}^d$ *and centered measures* $\bar\mu_0$, $\bar\mu_1$. *Then*

$$
\mathsf{d}_S(\mu_0,\mu_1)^2 = \|m_1 - m_0\|^2 + \mathsf{d}_S(\bar\mu_0, \bar\mu_1)^2.
$$

In particular, $\mathsf{d}_S(\mu_0,\mu_1) \geq \|m_1 - m_0\|$ *with equality if and only if* μ_1 *is a translation of* μ_0. *In this case,* $\mu_t = (\mathrm{id} + t(m_1 - m_0))_\# \mu_0$ *is the unique geodesic between them.*

Before diving into the proof of the theorem, we need an additional technical lemma. While Proposition 5.14 is concerned with what is happening at the level of the metric tensor, we need to check additionally that the map $\mu \mapsto \bar\mu = (\mathrm{id} - m)_\#\mu$ with $m = \langle \mu, \mathrm{id}\rangle$ preserves the class of admissible curves.

Lemma 5.17. *Let $(\mu_t)_t \in P$ with support in some bounded set and write m_t and $\bar{\mu}_t$ for the mean and centered part of μ_t. Then $(\bar{\mu}_t) \in P$ (with support in some possibly larger bounded set) and $(m_t) \in \mathcal{H}^1((0,1);\mathbb{R}^d)$.*

Proof. We start with a preliminary remark. We proved in Proposition 5.14 that for any $v \in \mathbb{R}^d$ and any μ we have $\langle v, \mathrm{id}\rangle = \frac{\varepsilon}{2}(\mathrm{id} - K_\mu^2)^{-1} H_\mu[-\operatorname{div}(\mu v)]$. Thus with the help of Theorem 4.3 and the estimate (4.4) we have $\langle v, \mathrm{id}\rangle \in \mathcal{H}_\mu$ and

$$\|\langle v, \mathrm{id}\rangle\|_{\mathcal{H}_\mu} \le C_1 \|\operatorname{div}(\mu v)\|_{\mathcal{H}_\mu^*} \le C_2 \|\operatorname{div}(\mu v)\|_{\mathcal{C}^{1,*}} \le C_2 \|v\| \tag{5.11}$$

for constants C_1, C_2 which only depend on the radius of the support of $\mu \in \mathcal{P}(X)$.

We first prove that $m_t \in \mathcal{H}^1([0,1],\mathbb{R}^d)$. It is enough to show that $\langle v, m_t\rangle \in \mathcal{H}^1([0,1],\mathbb{R})$ for any $v \in \mathbb{R}^d$. Recall from reformulating (4.8) that for any $\phi \in \mathcal{H}_{\mu_s}$ and any t, s,

$$\langle \mu_t - \mu_s, \phi\rangle = \langle \beta_t - \beta_s, a_{\mu_s}^{-1}\phi\rangle_{\mathcal{H}_c} + \langle \beta_t - \beta_s, H_c[\alpha_t \phi]\rangle_{\mathcal{H}_c}.$$

For $\phi = \langle v, \mathrm{id}\rangle$, with the help of the Cauchy–Schwarz inequality and using that multiplication by $a_{\mu_s}^{-1}$ is an isometry from $\mathcal{H}_{\mu_s}$ to $\mathcal{H}_c$, this gives

$$|\langle v, m_t - m_s\rangle| \le \|\beta_t - \beta_s\|_{\mathcal{H}_c} (\|\langle v, \mathrm{id}\rangle\|_{\mathcal{H}_{\mu_s}} + \|H_c[\alpha_t \langle v, \mathrm{id}\rangle]\|_{\mathcal{H}_c}).$$

Given that $(\beta_t)_t \in \mathcal{H}([0,1];\mathcal{H}_c)$, we need only to prove that $\|\langle v, \mathrm{id}\rangle\|_{\mathcal{H}_{\mu_s}} + \|H_c[\alpha_t\phi]\|_{\mathcal{H}_c}$ can be bounded independently on t, s to conclude with Theorem C.2. This is a consequence of (5.11) for the first term, while for the second term it is enough to notice that α_t is bounded in total variation uniformly in t.

We now proceed to prove that $\bar{\mu}_t \in P$. By uniqueness of the Schrödinger potentials we see that $f_{\bar{\mu}_t,\bar{\mu}_t}(x) = f_{\mu_t,\mu_t}(x + m_t)$. Writing $\bar{\beta}_t = B(\bar{\mu}_t)$, this gives $\bar{\beta}_t(x) = \beta_t(x + m_t)$. Thus for any t, s,

$$\begin{aligned}\|\bar{\beta}_t - \bar{\beta}_s\|_{\mathcal{H}_c} &= \|\beta_t(\cdot + m_t) - \beta_s(\cdot + m_s)\|_{\mathcal{H}_c} \\ &\le \|\beta_t(\cdot + m_t) - \beta_s(\cdot + m_t)\|_{\mathcal{H}_c} + \|\beta_s(\cdot + m_t) - \beta_s(\cdot + m_s)\|_{\mathcal{H}_c}.\end{aligned}$$

Using that the norm on $\mathcal{H}_c$ is invariant by translation, because so is the kernel, the first term coincides with $\|\beta_t - \beta_s\|_{\mathcal{H}_c}$. For the second term, we can write

$$\beta_s(x - m_t) - \beta_s(x - m_s) = \int_s^t \dot{m}_r \cdot \nabla\beta_s(x - m_r)\,\mathrm{d}r,$$

and so we only need to check that $\nabla\beta_s(\cdot - m_r) \in \mathcal{H}_c$ with a controlled norm to justify that the equality above holds as equality of functions in $\mathcal{H}_c$. Recalling $\beta = \exp(-f_{\mu,\mu}/\varepsilon)$ and recalling that we have already computed $\nabla f_{\mu,\mu}$ in (5.10), we find

$$\dot{m}_r \cdot \nabla\beta_s = \frac{2}{\varepsilon}(K_{\mu_s}[\langle \dot{m}_r, \mathrm{id}\rangle] - \langle \dot{m}_r, \mathrm{id}\rangle)\beta_s.$$

Using invariance by translation of the norm in $\mathcal{H}_c$, and that multiplication by $\beta_s = a_{\mu_s}^{-1}$ is an isometry between $\mathcal{H}_{\mu_s}$ and $\mathcal{H}_c$,

$$\|\dot{m}_r \cdot \nabla \beta_s(\cdot - m_s)\|_{\mathcal{H}_c} = \|\dot{m}_r \cdot \nabla \beta_s\|_{\mathcal{H}_c} \leq \frac{2}{\varepsilon}(\|K_{\mu_s}[\langle \dot{m}_r, \mathrm{id}\rangle]\|_{\mathcal{H}_{\mu_s}} + \|\langle \dot{m}_r, \mathrm{id}\rangle\|_{\mathcal{H}_{\mu_s}}).$$

We use that K_{μ_s} is a contraction (see Proposition 4.2), and again (5.11) to conclude that the right-hand side can be bounded by $C \cdot \|\dot{m}_r\|$ for some $C > 0$ independently on s. Thus

$$\|\bar{\beta}_t - \bar{\beta}_s\|_{\mathcal{H}_c} \leq \|\beta_t - \beta_s\|_{\mathcal{H}_c} + C \int_t^s \|\dot{m}_r\| \,\mathrm{d}r.$$

From Theorem C.2 we conclude that $(\bar{\beta}_t)_t \in \mathcal{H}^1([0,1]; \mathcal{H}_c)$ so that $(\bar{\mu}_t)_t \in P$. ∎

Proof of Theorem 5.16. If $(\mu_t)_t \in P(\mu_0, \mu_1)$, then by Lemma 5.17, $(B(\bar{\mu}_t))_t \in P$ and $(m_t)_t$ exists in $\mathcal{H}^1((0,1); \mathbb{R}^d)$. Writing $\dot{\bar{\mu}}_t$ for the derivative of the curve $t \mapsto \bar{\mu}_t$ and $\bar{\dot{\mu}}_t$ for the translation of the perturbation $\dot{\mu}_t$ by $-m_t$, it easy to check that $\dot{\bar{\mu}}_t = \bar{\dot{\mu}}_t + \mathrm{div}(\dot{m}_t \bar{\mu}_t)$ holds in $\mathcal{H}^*_{\mu_t,0}$ for any t when all three derivatives exist. In particular, translation invariance of the metric tensor yields

$$\begin{aligned}\mathbf{g}_{\mu_t}(\dot{\mu} + \mathrm{div}(\dot{m}\mu), \dot{\mu}_t + \mathrm{div}(\dot{m}_t\mu_t)) &= \mathbf{g}_{\bar{\mu}_t}(\bar{\dot{\mu}}_t + \mathrm{div}(\dot{m}_t\bar{\mu}_t), \bar{\dot{\mu}}_t + \mathrm{div}(\dot{m}_t\bar{\mu}_t)) \\ &= \mathbf{g}_{\bar{\mu}_t}(\dot{\bar{\mu}}_t, \dot{\bar{\mu}}_t).\end{aligned}$$

From Proposition 5.14 we deduce that

$$\begin{aligned}\int_0^1 \mathbf{g}_{\mu_t}(\dot{\mu}_t, \dot{\mu}_t)\,\mathrm{d}t &= \int_0^1 \|\dot{m}_t\|^2\,\mathrm{d}t + \int_0^1 \mathbf{g}_{\mu_t}(\dot{\mu} + \mathrm{div}(\dot{m}\mu), \dot{\mu}_t + \mathrm{div}(\dot{m}_t\mu_t))\,\mathrm{d}t \\ &= \int_0^1 \|\dot{m}_t\|^2\,\mathrm{d}t + \int_0^1 \mathbf{g}_{\bar{\mu}_t}(\dot{\bar{\mu}}_t, \dot{\bar{\mu}}_t)\,\mathrm{d}t.\end{aligned}$$

Conversely, from a curve of centered measures $(\bar{\mu}_t)_t$ and a path $(m_t)_t \in \mathcal{H}^1([0,1], \mathbb{R}^d)$ of means we can reconstruct $\mu_t = (\mathrm{id} + m_t)_\# \bar{\mu}_t$. We get the decomposition of the distance by minimizing over $(\mu_t)_t$.

The inequality follows, and if we have equality, then necessarily $(\bar{\mu}_t)_t$ is a constant curve and $(m_t)_t$ is a geodesic in $\mathbb{R}^d$. ∎

Corollary 5.18 (Energy of a traveling Dirac measure). *Let $c(x,y) = \|x-y\|^2$. Let $(x_t)_t \in \mathcal{H}^1((0,1); \mathbb{R}^d)$ and set $\mu_t = \delta_{x_t}$. Then*

$$E((\mu_t)_t) = \int_0^1 \|\dot{x}_t\|^2\,\mathrm{d}t.$$

In particular, $\mathsf{d}_S(\delta_{x_0}, \delta_{x_1}) = \|x_0 - x_1\|$ agrees with the metric on $\mathbb{R}^d$ and the unique geodesic is $(\delta_{y_t})_{t\in[0,1]}$ with $y_t = (1-t)x_0 + tx_1$.

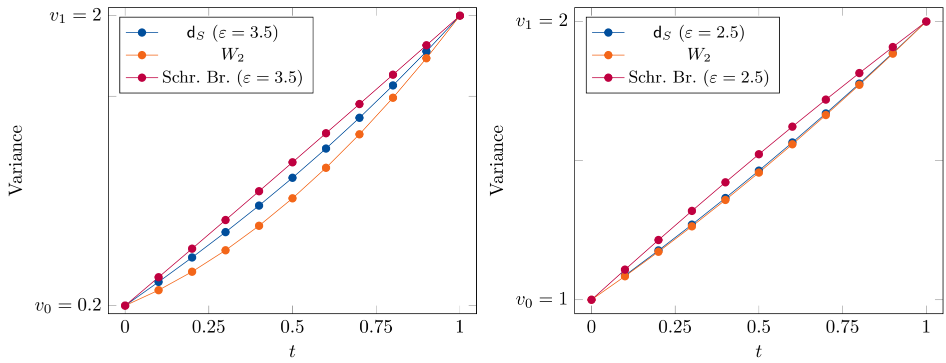

Figure 1. Representation of the Wasserstein and the (Gaussian-restricted) Sinkhorn-based geodesics for one-dimensional centered Gaussian distributions parametrized by their variance v_t, $t \in [0, 1]$, for different choices of v_0, v_1, and ε. For comparison we further show the ε-Schrödinger bridge.

6. Examples in the metric d_S

6.1. Explicit formula for Gaussian measures

Throughout this section we fix $c(x, y) = |x - y|^2$ as the squared Euclidean distance on $\mathbb{R}$. We derive the metric tensor as well as an explicit formula for the geodesic restriction of the distance d_S to the family of Gaussian measures in dimension one. By virtue of Theorem 5.16 we can restrict ourselves to centered measures. Strictly speaking, our theory a priori only applies to compactly supported measures, but we can still evaluate the metric tensor on Gaussians. We parametrize Gaussian measures by their variance v rather than their standard deviation $\sigma = \sqrt{v}$. Writing $\mathcal{N}(0, v)$ for the centered Gaussian measure with variance v, we recall from [29, 73] that

$$\mathrm{OT}_0(\mathcal{N}(0, v_0), \mathcal{N}(0, v_1)) = \left(\sqrt{v_1} - \sqrt{v_0}\right)^2,$$

which corresponds to the unregularized metric tensor $\mathbf{g}^0_{\mathcal{N}(0,v)}(\dot{v}, \dot{v}) = \frac{1}{4v}\dot{v}^2$. As discussed in the introduction, geodesics with respect to d_S differ from an ε-Schrödinger bridge, for which an explicit formula for Gaussian measures can be found in [15]: the result is displayed in Figure 1. With our metric we obtain the following.

Theorem 6.1. *Consider a path $(v_t)_t$ in $(0, \infty)$ that is differentiable around $t = 0$ and denote by $\dot{v}_t$ its derivative at t. Identifying the Gaussian distribution $\mathcal{N}(0, v_t)$ with its variance v_t and writing $v = v_0$, $\dot{v} = \dot{v}_0$, the metric tensor reads*

$$\mathbf{g}_v(\dot{v}, \dot{v}) = g(v)\dot{v}^2 \quad \textit{with } g(v) = \frac{1}{4\sqrt{\frac{\varepsilon^2}{16} + v^2}}. \tag{6.1}$$

Clearly, as $\varepsilon \to 0$ the metric tensor $\mathbf{g}_v(\dot{v}, \dot{v})$ converges to $\mathbf{g}^0_{\mathcal{N}(0,v)}(\dot{v}, \dot{v})$. For $\varepsilon \to \infty$ it degenerates and no longer penalizes changes in the variance: this is consistent with the claims of Section 4.4.

Proof of Theorem 6.1. We identify the Gaussian $\mu_t = \mathcal{N}(0, v_t)$ with its variance v_t. As shown in [49, Cor. 1], the entropic optimal transport cost between one-dimensional Gaussian distributions with variance v_t, v_s is explicitly given by

$$\mathrm{OT}_\varepsilon(v_t, v_s) = v_t + v_s - \frac{\varepsilon}{2}(\kappa_{ts} - \log \kappa_{ts} + \log 2 - 2), \quad \kappa_{ts} := 1 + \sqrt{1 + \frac{16}{\varepsilon^2} v_t v_s}. \tag{6.2}$$

Let $v_0 = v$, $\dot{v}_0 = \dot{v}$. Let us introduce $f(x) = 1 + \sqrt{1 + 16\varepsilon^{-2} x}$, so that $\kappa_{ts} = f(v_t v_s)$. We have

$$\partial_t \kappa_{ts} = \dot{v}_t v_s f'(v_t v_s), \quad \partial_s \partial_t \kappa_{ts} = \dot{v}_s \dot{v}_t f'(v_t v_s) + \dot{v}_s \dot{v}_t v_s v_t f''(v_t v_s).$$

As in the proof of Theorem 3.4, we now compute $\mathbf{g}_v(\dot{v}, \dot{v}) = -\frac{1}{2} \partial_t|_{t=0} \partial_s|_{s=0} \mathrm{OT}_\varepsilon(v_t, v_s)$. To this end, observe that

$$\partial_t \partial_s \mathrm{OT}_\varepsilon(v_t, v_s) = \frac{\varepsilon}{2}\Big(-\partial_t \partial_s \kappa_{ts} + \frac{\partial_t \partial_s \kappa_{ts}}{\kappa_{ts}} - \frac{\partial_t \kappa_{ts} \partial_s \kappa_{ts}}{\kappa_{ts}^2}\Big).$$

In the new notation this gives

$$\partial_t|_{t=0} \partial_s|_{s=0} \mathrm{OT}_\varepsilon(v_t, v_s) = \frac{\varepsilon}{2} \dot{v}^2 \Big((f'(v^2) + v^2 f''(v^2)) \Big(\frac{1}{f(v^2)} - 1 \Big) - v^2 \frac{f'(v^2)^2}{f(v^2)^2} \Big).$$

The metric tensor can thus be rewritten as

$$\begin{aligned} \mathbf{g}_v(\dot{v}, \dot{v}) &= g(v) \dot{v}^2, \\ g(v) &= -\frac{\varepsilon}{4} \Big\{ v^2 \Big(f''(v^2) \frac{1 - f(v^2)}{f(v^2)} - \frac{f'(v^2)^2}{f(v^2)^2} \Big) + f'(v^2) \frac{1 - f(v^2)}{f(v^2)} \Big\}. \end{aligned}$$

It remains to simplify the expression for g. To this end, let us compute the first and second derivatives of f:

$$f'(x) = \frac{8}{\varepsilon^2}\Big(1 + \frac{16}{\varepsilon^2} x\Big)^{-1/2}, \quad f''(x) = -\frac{64}{\varepsilon^4}\Big(1 + \frac{16}{\varepsilon^2} x\Big)^{-3/2}.$$

To simplify notation, introduce the variable $u = 1 + 16\varepsilon^{-2} v^2$ in such a way that

$$f(v^2) = 1 + \sqrt{u}, \quad f'(v^2) = \frac{8}{\varepsilon^2 \sqrt{u}}, \quad f''(v^2) = -\frac{64}{\varepsilon^4 u^{3/2}},$$

and in particular

$$\frac{1 - f(v^2)}{f(v^2)} = -\frac{\sqrt{u}}{1 + \sqrt{u}}.$$

Plugging this into the expression for g yields

$$g(v) = \frac{\varepsilon}{4}\Big\{-v^2\Big(\frac{64}{\varepsilon^4}\frac{1}{u(1+\sqrt{u})} - \frac{64}{\varepsilon^4}\frac{1}{u(1+\sqrt{u})^2}\Big) + \frac{8}{\varepsilon^2}\frac{1}{1+\sqrt{u}}\Big\}$$
$$= -\frac{16v^2}{\varepsilon^3}\frac{1}{\sqrt{u}(1+\sqrt{u})^2} + \frac{2}{\varepsilon}\frac{1}{1+\sqrt{u}}.$$

To lighten the notation further, use $\frac{16}{\varepsilon^2}v^2 = u - 1$ to get

$$g(v) = -\frac{1}{\varepsilon}\frac{u-1}{\sqrt{u}(1+\sqrt{u})^2} + \frac{2}{\varepsilon}\frac{1}{1+\sqrt{u}} = \frac{1}{\varepsilon}\frac{1+2\sqrt{u}+u}{\sqrt{u}(1+\sqrt{u})^2} = \frac{1}{\varepsilon\sqrt{u}}.$$

The claim follows when unpacking $u = 1 + \frac{16}{\varepsilon^2}v^2$. ■

We now turn to the metric on $(0,\infty)$ induced by the tensor $\mathbf{g}$ of Theorem 6.1. Said differently, we look at the geodesic restriction of d_S to Gaussian measures. At this point we do not know whether the set of Gaussian measures is in fact geodesically convex.

Theorem 6.2. *Let F be the antiderivative of the function $x \mapsto (1+x^2)^{-1/4}$. Let $v_0, v_1 \geq 0$. Then the geodesic restriction $\widehat{\mathsf{d}_S}$ of d_S to Gaussians can be expressed as*

$$\widehat{\mathsf{d}_S}(v_0, v_1) = \frac{\sqrt{\varepsilon}}{4}\Big|F\Big(\frac{4}{\varepsilon}v_1\Big) - F\Big(\frac{4}{\varepsilon}v_0\Big)\Big|.$$

The geodesic $(v_t)_{t\in[0,1]}$ joining v_0 to v_1 is given by

$$v_t = \frac{\varepsilon}{4}F^{-1}\Big[(1-t)F\Big(\frac{4}{\varepsilon}v_0\Big) + tF\Big(\frac{4}{\varepsilon}v_1\Big)\Big]$$

and the map $t \mapsto v_t$ is convex.

Proof. Given the metric tensor (6.1) on $(0,\infty)$, geodesics correspond to minimizers of the energy

$$E((v_t)_t) = \int_0^1 g(v_t)\dot{v}_t^2 \,\mathrm{d}t.$$

It is easy to see that geodesics must be monotone and since g is smooth on the interval $[v_0, v_1]$ we can use standard tools from smooth calculus of variations on one-dimensional curves to obtain the minimizer. Let us call $q_t = \dot{v}_t$, and $L(v,q) = g(v)q^2$ the Lagrangian. Recall that the Euler–Lagrange equation is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial L}{\partial q}(v_t, \dot{v}_t)\Big) = \frac{\partial L}{\partial v}(v_t, \dot{v}_t).$$

Expanding it yields

$$g(v_t)\ddot{v}_t = -\frac{g'(v_t)}{2}\dot{v}_t^2.$$

A simple inspection reveals that g is non-negative and decreasing. Thus $g' \leq 0$ and the equation tells us that $\ddot{v} \geq 0$, which is equivalent to the fact that $(v_t)_t$ is convex in time.

Note that if $v_0 \neq v_1$, then $\dot{v}$ cannot vanish: if it were the case, by the unique solvability of the above ODE, v should be a constant function. Thus $(v_t)_t$ is monotone, and without loss of generality we focus on $v_0 \leq v_1$. In this case $(v_t)_t$ is increasing in time. For solutions of the above ODE one has that $C = g(v_t)\dot{v}_t^2$ is constant in time (as usual for geodesics) and non-negative and thus v satisfies the first-order equation

$$\dot{v}_t = \sqrt{\frac{C}{g(v_t)}} = 2\sqrt{C}\Big(\frac{\varepsilon^2}{16} + v_t^2\Big)^{1/4} = \sqrt{C}\,\varepsilon\Big(1 + \Big(\frac{4v_t}{\varepsilon}\Big)^2\Big)^{1/4}.$$

This equation is separable and thus can be integrated. Recalling the definition of F above, the ODE yields

$$\frac{\mathrm{d}}{\mathrm{d}t} F\Big(\frac{4}{\varepsilon} v_t\Big) = 4\sqrt{\frac{C}{\varepsilon}}, \tag{6.3}$$

and thus integrating between 0 and 1 gives

$$C = \frac{\varepsilon}{16}\Big[F\Big(\frac{4}{\varepsilon}v_1\Big) - F\Big(\frac{4}{\varepsilon}v_0\Big)\Big]^2.$$

Now $\widehat{\mathsf{d}_S}(v_0, v_1) = \sqrt{C}$ and the formula for v_t follows from (6.3). ∎

6.2. The two-point space

In this section, we analyze the Riemannian tensor $\mathbf{g}_\mu$ and the metric d_S for a two-point space $X = \{x_1, x_2\}$, where probability measures are fully specified by the mass $m \in [0,1]$ located at x_1 and the metric is fully characterized by the distance $d(x_1, x_2)$. The precise result is stated in Proposition 6.3 below. When $d(x_1, x_2)$ is much smaller than the blur scale $\sqrt{\varepsilon}$ we find that $\mathsf{d}_S(\delta_{x_1}, \delta_{x_2}) \approx d(x_1, x_2)$, which is the same as one would get on an interval $[x_1, x_2] \subset \mathbb{R}$ of length $d(x_1, x_2)$. Thus, it seems that the metric d_S may "gloss over" the discreteness of X at very small length scales. Conversely, for $d(x_1, x_2) \gg \sqrt{\varepsilon}$ the diameter of the space $(\mathcal{P}(X), \mathsf{d}_S)$ grows at least like $\exp(d(x_1, x_2)^2/(2\varepsilon))$. As $d(x_1, x_2)$ increases, the blur in the entropic optimal transport underlying the construction of d_S decreases. It is well known that $\mathcal{P}(X)$ equipped with the quadratic Monge–Kantorovich distance is not a length space [48, Rem. 2.1]. Thus this exponential divergence of the diameter is consistent with the conjecture that d_S converges to the unregularized quadratic Monge–Kantorovich distance as $\varepsilon \to 0$. See Section 8 for a slightly longer discussion and some related references.

Proposition 6.3. *Let $X = \{x_1, x_2\}$. For a parameter $r \in (0, \infty)$ let $d(x_1, x_2) = r$, $c = d^2$.*

(1) *For $m \in (0,1)$, $\mu = m\delta_{x_1} + (1-m)\delta_{x_2}$, $\dot{m} \in \mathbb{R}$, $\dot{\mu} = \dot{m}(\delta_{x_1} - \delta_{x_2})$, denote by p the amount of mass moving from x_1 to x_2 in the optimal self-transport plan from μ to itself (see equation (6.5) below). Then the metric tensor is given by*

$$\mathbf{g}_\mu(\dot{\mu}, \dot{\mu}) = \frac{\varepsilon}{2}\dot{m}^2 \frac{m(1-m) - p}{p(2m(1-m) - p)}.$$

(2) *For* $m\colon [0,1] \to [0,1]$ *continuously differentiable set* $\mu_t = m_t \delta_{x_1} + (1-m_t)\delta_{x_2}$ *for* $t \in [0,1]$. *Then, as* $r \to 0$,

$$E((\mu_t)_t) = r^2 \cdot \int_0^1 \dot{m}_t^2 \, \mathrm{d}t + \mathcal{O}\Big(\frac{1}{\varepsilon}\|\dot{m}\|_\infty^2 r^4\Big). \tag{6.4}$$

(3) *Let* $C > 0$. *In the setting of point (1), when* $\sqrt{m(1-m)} \geq C \cdot \exp(-2r^2/\varepsilon)$ *it holds that*

$$\mathbf{g}_\mu(\dot{\mu},\dot{\mu}) = \frac{\varepsilon}{2}\dot{m}^2 \cdot \bigg(\frac{\exp(r^2/\varepsilon)}{2\sqrt{m(1-m)}} + \mathcal{O}\big((m(1-m))^{-1}\big)\bigg)$$

with the constant in the $\mathcal{O}$*-term only depending on* C.

In particular, for small distances, $\dot{m} = -1$ in (6.4) leads to $E(((1-t)\delta_{x_1} + t\delta_{x_2})_t) = r^2 + \mathcal{O}(\frac{1}{\varepsilon}r^4)$.

Proof of Proposition 6.3. Let $m \in (0,1)$, $\mu = m\delta_{x_1} + (1-m)\delta_{x_2}$, $M = \sqrt{m(1-m)}$. We compute the optimal self-transport plan π and obtain k_μ as the density of π with respect to $\mu \otimes \mu$. Since X only contains two points, we may use the matrix vector notation of the discrete setting. To avoid ambiguity, we denote the entry-wise multiplication of two matrices A, B by $A \odot B$. Denoting $a_\mu = \exp(f_{\mu,\mu}/\varepsilon)$ as a vector $a = (a_1, a_2)^\top$, we see that $\pi = k_\mu \odot \mu \otimes \mu = (a \cdot a^\top) \odot k_c \odot \mu \otimes \mu$. The set of admissible transport plans is described entirely by one parameter p describing the exchange of mass between the two points. With $\exp(-2d(x_1,x_2)^2/\varepsilon) =: b \in (0,1)$,

$$\pi = \begin{pmatrix} m-p & p \\ p & (1-m)-p \end{pmatrix}, \quad k_c \odot \mu \otimes \mu = \begin{pmatrix} m^2 & M^2\sqrt{b} \\ M^2\sqrt{b} & (1-m)^2 \end{pmatrix}.$$

We have $p^2/((m-p)(1-m-p)) = (a_1 a_2 M^2 \sqrt{b})^2/(a_1^2 m^2 a_2^2 (1-m)^2) = b$. This is a quadratic equation whose solution is

$$p = \frac{1}{2(1-b)}[-b + \sqrt{b^2 + 4M^2 b(1-b)}]. \tag{6.5}$$

We can express the matrix representing K_μ as a scaling of π:

$$K_\mu = \begin{pmatrix} m^{-1} & 0 \\ 0 & (1-m)^{-1} \end{pmatrix} \cdot \pi = \begin{pmatrix} \frac{m-p}{m} & \frac{p}{m} \\ \frac{p}{1-m} & \frac{1-m-p}{1-m} \end{pmatrix}.$$

Unsurprisingly, $(1,1)^\top$ is an eigenvector associated to the eigenvalue $\lambda_1 = 1$. We know that the other eigenvector of K_μ should be orthogonal to $(1,1)^\top$ in the inner product of $L^2(X,\mu)$, so it must be $(1-m,-m)^\top$. The corresponding eigenvalue is computed via

$$K_\mu \begin{pmatrix} 1-m \\ -m \end{pmatrix} = (1 - p/M^2)\begin{pmatrix} 1-m \\ -m \end{pmatrix}. \tag{6.6}$$

In particular, on this eigenspace $(\mathrm{id} - K_\mu^2)^{-1}$ acts as multiplication by $(1-\lambda_2^2)^{-1}$ with $\lambda_2 = 1 - p/M^2$. To compute the metric tensor $\mathbf{g}_\mu(\dot\mu, \dot\mu) = \langle \dot\mu, (\mathrm{id} - K_\mu^2)^{-1} H_\mu[\dot\mu]\rangle$ for some vertical perturbation $\dot\mu = \dot m \cdot (1, -1)^\top$, we need to compute $H_\mu[\dot\mu] = K_\mu[\mathrm{d}\dot\mu/\mathrm{d}\mu]$. As $\mathrm{d}\dot\mu/\mathrm{d}\mu = \dot m/M^2(1-m, -m)^\top$, by the previous computation

$$H_\mu \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1 - p/M^2}{M^2} \begin{pmatrix} 1-m \\ -m \end{pmatrix}. \tag{6.7}$$

The formulas (6.6) and (6.7) lead to

$$\mathbf{g}_\mu(\dot\mu, \dot\mu) = \frac{\varepsilon}{2} \frac{\dot m^2 \lambda_2}{M^2(1-\lambda_2^2)} = \frac{\varepsilon}{2} \dot m^2 \frac{M^2 - p}{p(2M^2 - p)}. \tag{6.8}$$

Small r asymptotic. The small parameter in the expansion (6.5) is $\tau = 2d(x_1, x_2)^2/\varepsilon$, so that we have $(1-b) = \tau + \mathcal{O}(\tau^2)$. A second-order Taylor expansion of the square root leads to $p = M^2 - \tau M^4 + \mathcal{O}(\tau^2)$, which gives $\lambda_2 = \tau M^2 + \mathcal{O}(\tau^2)$. Thus $(1-\lambda_2^2)^{-1} = 1 + \mathcal{O}(\tau^2)$ and (6.8) gives

$$\mathbf{g}_\mu(\dot\mu, \dot\mu) = \frac{\varepsilon}{2}(\tau \dot m^2 + \mathcal{O}(\tau^2 \dot m^2)) = \dot m^2 \cdot d(x_1, x_2)^2 + \mathcal{O}\Big(\frac{1}{\varepsilon} \dot m^2 d(x_1, x_2)^4\Big).$$

Big r asymptotic. We set $\varrho := \sqrt{b} = \exp(-d(x_1, x_2)^2/\varepsilon)$. Provided that $M \geq C\varrho$ for some $C > 0$, rewriting (6.5) gives

$$p = \frac{1}{2(1-\varrho^2)}\Big(-\varrho^2 + 2M\varrho\sqrt{1 + \varrho^2(1-4M^2)/(4M^2)}\Big) = M\varrho + \mathcal{O}(\varrho^2),$$

with the constant in the $\mathcal{O}$-term only depending on C. This gives

$$\lambda_2 = 1 - \frac{p}{M^2} = 1 - \frac{\varrho}{M} + \mathcal{O}((\varrho/M)^2),$$

and using (6.8) it leads to

$$\mathbf{g}_\mu(\dot\mu, \dot\mu) = \frac{\varepsilon}{2}\dot m^2 \cdot \Big(\frac{1}{2M\varrho} + \mathcal{O}(1/M^2)\Big) = \frac{\varepsilon}{2}\dot m^2 \cdot \Big(\frac{\exp(r^2/\varepsilon)}{2M} + \mathcal{O}(1/M^2)\Big). \qquad \blacksquare$$

For small M/ϱ one can use analogous methods to find $p = M^2 + \mathcal{O}(M^4/\varrho^2)$, which leads to $\lambda_2 = \mathcal{O}(1/\varrho^2)$ and $\mathbf{g}_\mu(\dot\mu, \dot\mu) = \mathcal{O}(\varepsilon \dot m^2/\varrho^2)$.

7. Negative results on the Sinkhorn divergence and the metric tensor

7.1. $\sqrt{S_\varepsilon}$ violates the triangle inequality

One of the main goals of this paper is to find a "Riemannian" distance associated with the Sinkhorn divergence S_ε. The reader may wonder why one could not directly use S_ε as a distance, since it enjoys several nice features, as recalled in Theorem 1.1. However, despite being positive definite, even with the squared Euclidean distance as the cost function the Sinkhorn divergence, or rather $\sqrt{S_\varepsilon}$ (since S_ε converges to the square of the Monge–Kantorovich distance), is not itself a distance! Although this fact is well known in the community, we have not been able to find an explicit statement of this property. For this reason, we provide an elementary proof, which will actually exclude any convex power of $\sqrt{S_\varepsilon}$.

Theorem 7.1. *On $\mathbb{R}$ let $c(x, y) = |x - y|^2$. Then S_ε^α fails to satisfy the triangle inequality for any $\alpha \geq \frac{1}{2}$.*

The case $0 < \alpha < \frac{1}{2}$ is unsuitable for our purposes too, as $S_\varepsilon^\alpha(\delta_x, \delta_y) = \|x - y\|^{2\alpha}$ is a concave function of the Euclidean distance. Hence, if S_ε^α were a metric at all, it would induce a Finslerian rather than Riemannian geometry.

Proof of Theorem 7.1. It is readily verified that S_ε^α cannot be a metric for any $\alpha > 1/2$, as it is sufficient to consider $\mu_i = \delta_{m_i}$ and observe that $S_\varepsilon^\alpha(\mu_0, \mu_1) = \|m_0 - m_1\|^{2\alpha}$ does not satisfy the triangle inequality.

Finally, $\sqrt{S_\varepsilon}$ is not a distance either. The idea is to look at the Sinkhorn divergence between one-dimensional Gaussian distributions, for which an explicit formula is available as a direct consequence of (6.2) (see [49, Cor. 1]). Indeed, given $\mu_i = \mathcal{N}(m_i, v_i)$ with $m_i \in \mathbb{R}$, $v_i \geq 0$ (for $i = 0, 1$), the Sinkhorn divergence between them is explicitly given by

$$S_\varepsilon(\mu_0, \mu_1) = \|m_0 - m_1\|^2 + \frac{\varepsilon}{4}\Big(\kappa_{00} - 2\kappa_{01} + \kappa_{11} + \log\Big(\frac{\kappa_{01}^2}{\kappa_{00}\kappa_{11}}\Big)\Big),$$

where $\kappa_{ij} := 1 + (1 + 16\varepsilon^{-2} v_i v_j)^{1/2}$, for $i, j \in \{0, 1\}$, as in (6.2). Choose $\mu_0 = \delta_0$, $\mu_1 = \mathcal{N}(0, \frac{\varepsilon}{4}x)$, and $\mu_2 = \mathcal{N}(0, \frac{\varepsilon}{4}v)$, for a fixed $v > 0$: we claim that

$$\sqrt{S_\varepsilon(\mu_0, \mu_2)} > \sqrt{S_\varepsilon(\mu_0, \mu_1)} + \sqrt{S_\varepsilon(\mu_1, \mu_2)} \tag{7.1}$$

when x is sufficiently small and v sufficiently large. To prove the claim, observe that

$$4\varepsilon^{-1} S_\varepsilon(\mu_0, \mu_2) = \sqrt{1 + v^2} - \log(1 + \sqrt{1 + v^2}) + \log 2 - 1.$$

To lighten the notation, let us introduce

$$\begin{aligned} \zeta(x) &:= 4\varepsilon^{-1} S_\varepsilon(\mu_0, \mu_1) = \sqrt{1 + x^2} - \log(1 + \sqrt{1 + x^2}) + \log 2 - 1, \\ \xi(x) &:= 4\varepsilon^{-1} S_\varepsilon(\mu_1, \mu_2) = \sqrt{1 + v^2} - 2\sqrt{1 + vx} + 2\log(1 + \sqrt{1 + vx}) \\ &\qquad - \log(1 + \sqrt{1 + v^2}) - \log 2 + 1 + \zeta(x). \end{aligned}$$

Squaring both sides, and after algebraic manipulations, (7.1) is thus equivalent to

$$\Psi(x) := 2\log(1 + \sqrt{1 + vx}) - 2\sqrt{1 + vx} - 2\log 2 + 2 + 2\zeta(x) + 2\sqrt{\zeta(x)\xi(x)} < 0$$

for x sufficiently close to 0. To prove this inequality, let us compute

$$\Psi'(x) = -\frac{v}{1+\sqrt{1+vx}} + \frac{2x}{1+\sqrt{1+x^2}} + \Big(\frac{\xi(x)}{\zeta(x)}\Big)^{1/2} \frac{x}{1+\sqrt{1+x^2}} + \Big(\frac{\zeta(x)}{\xi(x)}\Big)^{1/2}\Big(-\frac{v}{1+\sqrt{1+vx}} + \frac{x}{1+\sqrt{1+x^2}}\Big)$$

and note that $\zeta(0) = 0$, so that $\Psi(0) = 0$ and

$$\lim_{x\to 0} \Psi'(x) = -\frac{v}{2} + \frac{\sqrt{\xi(0)}}{2} \lim_{x\to 0} \frac{x}{\sqrt{\zeta(x)}}$$

with $\xi(0) = \sqrt{1+v^2} - 1 + \log 2 - \log(1+\sqrt{1+v^2}) > 0$. It is now sufficient to observe that $\zeta'(0) = 0$ and $\zeta''(0) = 1/2$: by a Taylor expansion of $\zeta(x)$ up to order 2 we finally get

$$\Psi'(0) = -\frac{v}{2} + \big(\sqrt{1+v^2} - 1 + \log 2 - \log\big(1+\sqrt{1+v^2}\big)\big)^{1/2},$$

which is negative provided v is sufficiently large. We have thus shown that $\Psi(0) = 0$ and $\Psi'(0) < 0$, hence the conclusion. ∎

Remark 7.2. One can also use the measures $\mu_r = \frac{1}{2}\delta_r + \frac{1}{2}\delta_{-r}$ from Section 7.2 to show that the square root of the Sinkhorn divergence does not satisfy the triangle inequality: Choose $r = \sqrt{\varepsilon}$. Then Lemma 7.5 gives

$$\frac{1}{\varepsilon} S_\varepsilon(\mu_0, \mu_r) = 1 - \frac{1}{2}\log 2 + \frac{1}{2}\log\Big(1 + \frac{1}{e^4}\Big),$$
$$\frac{1}{\varepsilon} S_\varepsilon(\mu_r, \mu_{2r}) = 1 - \log\Big(1 + \frac{1}{e^8}\Big) + \frac{1}{2}\log\Big(1 + \frac{1}{e^4}\Big) + \frac{1}{2}\log\Big(1 + \frac{1}{e^{16}}\Big),$$
$$\frac{1}{\varepsilon} S_\varepsilon(\mu_0, \mu_{2r}) = 4 - \frac{1}{2}\log 2 + \frac{1}{2}\log\Big(1 + \frac{1}{e^{16}}\Big).$$

This yields $\frac{1}{\sqrt{\varepsilon}} \cdot \big(\sqrt{S_\varepsilon(\mu_0, \mu_{2r})} - \sqrt{S_\varepsilon(\mu_0, \mu_r)} - \sqrt{S_\varepsilon(\mu_r, \mu_{2r})}\big) \approx 0.093 > 0$, which breaks the triangle inequality.

Remark 7.3. We observe that in the proof of Theorem 7.1 and in Figure 2, the triangle inequality for $\sqrt{S_\varepsilon}$ breaks when the initial and intermediate measures are concentrated below the blur scale and the target measure is more spread out.

7.2. The Sinkhorn divergence is not jointly convex

The second negative result about the Sinkhorn divergence deals with joint convexity. The Sinkhorn divergence S_ε interpolates between squared MMD and Wasserstein distances which are both jointly convex functions of their inputs, and moreover S_ε is coordinate-wise convex as recalled in Theorem 1.1. Thus it is natural to investigate the convexity properties of S_ε as a function on $\mathcal{P}(X) \times \mathcal{P}(X)$, and we show that the answer is negative.

Throughout this section let $c(x, y) = |x - y|^2$ be the quadratic cost.

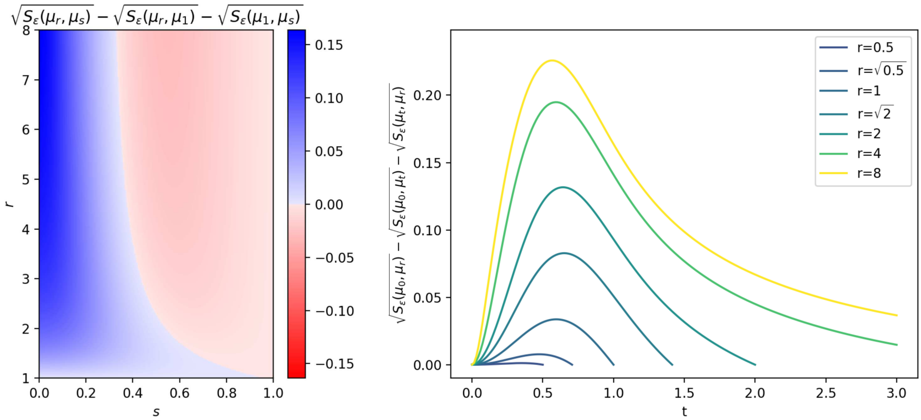

Figure 2. Gap in the triangle inequality for $\sqrt{S_\varepsilon}$ on the measures $\mu_r = \frac{1}{2}\delta_r + \frac{1}{2}\delta_{-r}$, discussed in Remark 7.2, for $\varepsilon = 1$. Left: The gap in the triangle inequality when going from μ_s to μ_r via μ_1. The triangle inequality breaks in the blue region, where $s \ll 1 = \sqrt{\varepsilon} \ll r$. Right: Gap for $s = 0$ and varying intermediate position $0 \le t \le r$ for different values of r.

Theorem 7.4. *Let $X \subseteq \mathbb{R}$ be the closure of an open set with* $\operatorname{diam}(X) > \sqrt{2\varepsilon}$. *Then the Sinkhorn divergence S_ε is not jointly convex over $\mathcal{P}(X) \times \mathcal{P}(X)$.*

We first examine the measures that will form the counterexample.

Lemma 7.5. *For $r \in [0,\infty)$ set $\mu_r := \frac{1}{2}\delta_r + \frac{1}{2}\delta_{-r}$. Then for $r, s \in [0,\infty)$ it holds that*

$$\begin{aligned}\frac{1}{\varepsilon}\mathrm{OT}_\varepsilon(\mu_r,\mu_s) &= \log 2 - \log(k_c(r,s) + k_c(r,-s)),\\ \frac{1}{\varepsilon}S_\varepsilon(\mu_r,\mu_s) &= -\log(k_c(r,s) + k_c(r,-s))\\ &\qquad + \frac{1}{2}\big(\log(1+k_c(r,-r)) + \log(1+k_c(s,-s))\big).\end{aligned}$$

Proof. We first compute the Schrödinger potentials for (μ_r,μ_s). As μ_r and μ_s are symmetric around 0, we have $f_{\mu_r,\mu_s}(r) = f_{\mu_r,\mu_s}(-r) =: f$, $g_{\mu_r,\mu_s}(s) = g_{\mu_r,\mu_s}(-s) =: g$. The Schrödinger system (2.3) then reduces to

$$\exp(-f/\varepsilon) = \frac{1}{2}k_c(r,s)\exp(g/\varepsilon) + \frac{1}{2}k_c(r,-s)\exp(g/\varepsilon),$$

i.e. $\frac{1}{\varepsilon}(f+g) = \log 2 - \log(k_c(r,s) + k_c(r,-s))$. Now the claim follows from (2.2). ∎

With these explicit expressions at hand, the proof of Theorem 7.4 comes down to an appropriate choice of parameters.

Proof of Theorem 7.4. Use $\mu_r := \frac{1}{2}\delta_r + \frac{1}{2}\delta_{-r}$ as in Lemma 7.5. For lighter notation, we assume that $\pm r \in X$ and set $\kappa = k_c(r, -r)$. We compute the Taylor expansion of $S_\varepsilon(\mu_r, \mu_{r+t})$ for $r \in [0, \infty)$. With the explicit formula for the Sinkhorn divergence from Lemma 7.5 and the second-order Taylor expansions

$$\begin{aligned} k_c(r, r+x) &= 1 - \frac{1}{\varepsilon}x^2 + \mathcal{O}(x^3), \\ k_c(r, -(r+x)) &= \kappa - \kappa\frac{4r}{\varepsilon}x - \kappa\Big(\frac{1}{\varepsilon} - \frac{8r^2}{\varepsilon^2}\Big)x^2 + \mathcal{O}(x^3), \\ \log(1+\kappa+x) &= \log(1+\kappa) + \frac{1}{1+\kappa}x - \frac{1}{2(1+\kappa)^2}x^2 + \mathcal{O}(x^3), \end{aligned}$$

we obtain

$$\begin{aligned} \frac{1}{\varepsilon}S_\varepsilon(\mu_r, \mu_{r+t}) = &-\log\Big(1 + \kappa - \frac{1}{\varepsilon}t^2 - \kappa\Big(\frac{4r}{\varepsilon}t + \Big(\frac{1}{\varepsilon} - \frac{8r^2}{\varepsilon^2}\Big)t^2\Big) + \mathcal{O}(t^3)\Big) \\ &+ \frac{1}{2}\log(1+\kappa) \\ &+ \frac{1}{2}\log\Big(1 + \kappa - \kappa\Big(\frac{4r}{\varepsilon}\cdot 2t + \Big(\frac{1}{\varepsilon} - \frac{8r^2}{\varepsilon^2}\Big)\cdot 4t^2\Big) + \mathcal{O}(t^3)\Big) \\ = &\frac{1}{\varepsilon}t^2 + t^2\frac{\kappa}{1+\kappa}\Big(\frac{1}{1+\kappa}\frac{8r^2}{\varepsilon^2} - \frac{2}{\varepsilon}\Big) + \mathcal{O}(t^3). \end{aligned} \tag{7.2}$$

As $\kappa \le 1$, when $r > \sqrt{\varepsilon/2}$ this is greater than $\frac{1}{\varepsilon}S_\varepsilon(\delta_{\pm r}, \delta_{\pm(r+t)}) = \frac{1}{\varepsilon}t^2$ for small t. Thus in this case $S_\varepsilon(\mu_r, \mu_{r+t}) > \frac{1}{2}S_\varepsilon(\delta_r, \delta_{r+t}) + \frac{1}{2}S_\varepsilon(\delta_{-r}, \delta_{-(r+t)})$, disproving convexity.

When $\operatorname{diam}(X) > \sqrt{2\varepsilon}$ we can always translate X to contain points $\pm r$ with $r > \sqrt{\varepsilon/2}$, hence disproving joint convexity of S_ε. ■

The non-convexity of S_ε descends to the level of the metric tensor.

Corollary 7.6. *Let $X \subseteq \mathbb{R}$ be the closure of an open set with* $\operatorname{diam}(X) > \sqrt{2\varepsilon}$*. Then the map $(\mu, \dot\mu) \mapsto \mathbf{g}_\mu(\dot\mu, \dot\mu)$ is not jointly convex on $\mathcal{P}(X) \times \mathcal{C}^1(X)^*_0$.*

Proof. We follow the proof of Theorem 7.4. Denote $\mu = \mu_r$, $\dot\mu = \frac{1}{2}\delta'_r - \frac{1}{2}\delta'_{-r}$. Theorem 3.4 and (7.2) give

$$\mathbf{g}_\mu(\dot\mu, \dot\mu) = \lim_{t\to 0}\frac{S_\varepsilon(\mu_r, \mu_{r+t})}{t^2} = 1 + \frac{\kappa}{1+\kappa}\Big(\frac{1}{1+\kappa}\frac{8r^2}{\varepsilon} - 2\Big), \tag{7.3}$$

which is greater than $\mathbf{g}_{\delta_{\pm r}}(\pm\delta'_{\pm r}, \pm\delta'_{\pm r}) = 1$ when $r > \sqrt{\varepsilon/2}$. ■

The counterexample from the proof of Corollary 7.6 is explored further in Figure 3.

As discussed in the introduction, if the metric tensor were jointly convex, then existence of geodesic (Theorem 5.4) could be obtained with arguments of convex analysis. Corollary 7.6 shows that some more involved arguments such as the ones given in Section 5.2 are necessary. Joint convexity of the metric tensor $\tilde{\mathbf{g}}_\mu(\dot\beta, \dot\beta)$ in μ and $\dot\beta$ is ill defined, because the compatibility assumption $\langle B(\mu), \dot\beta\rangle = 0$ can fail at a convex combination.

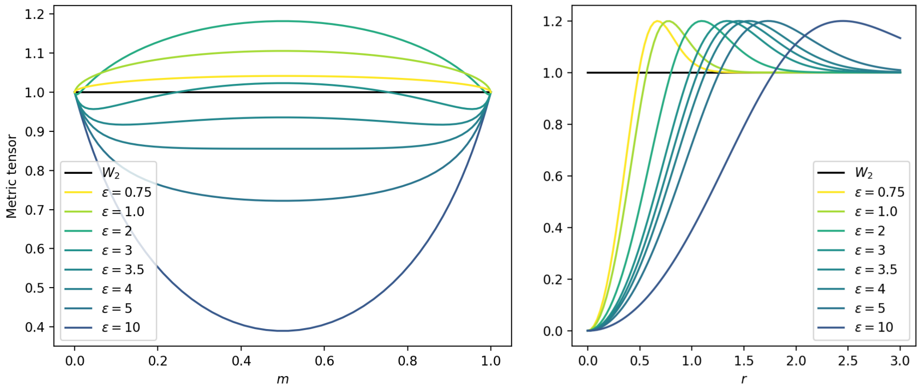

Figure 3. Numerically computed values of the metric tensor of two point masses moving apart with $\mu = m\delta_{-r} + (1-m)\delta_r$, $\dot\mu = -m\delta'_{-r} + (1-m)\delta'_r$ for different values of ε. This is the linear interpolation between a point mass at $-r$ moving to the left and a point mass at r moving to the right ($m = 1$ and $m = 0$ respectively). Left: Fixed $r = 1$, varying mass m. Joint convexity of the metric tensor fails, as the graphs sometimes exceed the value 1. Right: Fixed $m = \frac{1}{2}$, varying position r. This is the term (7.3) computed in the proof of Corollary 7.6, which only depends on $r/\sqrt{\varepsilon}$. In both plots, the black line shows $\|v\|^2_{L^2(X,\mu)} = \mathbf{g}^0_\mu(\dot\mu, \dot\mu) = 1$ for the metric tensor of classical optimal transport (1.3).

Remark 7.7. Although very explicit, the previous computations do not shed light on the reason why S_ε fails to be jointly convex. Instead, we present in addition a more abstract argument, which is originally how we proved Theorem 7.4. It proceeds by contradiction, by combining three results already present in the literature:

(a) A very insightful remark by Savaré and Sodini [70], which says that $\mathrm{OT} := \mathrm{OT}_0$ (i.e. the unregularized transport cost induced by c from (1.1)) is the largest convex and weak-$*$-lower semicontinuous function which coincides with c on Dirac masses. Since the set of couplings between two Dirac masses contains only one element, with $c(x, y) = |x - y|^2$ we have $S_\varepsilon(\delta_x, \delta_y) = c(x, y)$ for any $x, y \in X$. As a consequence, if S_ε were jointly convex, as it is in addition weak-$*$-continuous, then $S_\varepsilon(\mu_0, \mu_1) \le \mathrm{OT}(\mu_0, \mu_1)$ for all $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R})$. So rather than negating the convexity of S_ε, it is enough to find measures μ_0, μ_1 such that $S_\varepsilon(\mu_0, \mu_1) > \mathrm{OT}(\mu_0, \mu_1)$.

(b) We then use a small-ε asymptotics of S_ε [21, 22]. Introducing the Fisher information of a measure μ with respect to the Lebesgue measure as

$$I(\mu) := 4\int_{\mathbb{R}} |\nabla\sqrt{\rho}|^2 \,\mathrm{d}x, \quad \text{if } \mu = \rho\,\mathrm{d}x \text{ and } \sqrt{\rho} \in H^1(\mathbb{R}, \mathrm{d}x),$$

and $+\infty$ otherwise, this result reads as follows: Let $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R})$ have bounded densities and supports and denote by $(\mu_t)_{t\in[0,1]}$ the Wasserstein-2 geodesic connecting them.

If $I(\mu_0)$, $I(\mu_1)$, and $\int_0^1 I(\mu_t)\,\mathrm{d}t$ are finite, then it holds that

$$S_\varepsilon(\mu_0,\mu_1) - \mathrm{OT}_0(\mu_0,\mu_1) = \frac{\varepsilon^2}{8}\left(\int_0^1 I(\mu_t)\,\mathrm{d}t - \frac{I(\mu_0)+I(\mu_1)}{2}\right) + o(\varepsilon^2). \tag{7.4}$$

Thus, given (a), it is enough to find measures μ_0, μ_1 such that

$$\int_0^1 I(\mu_t)\,\mathrm{d}t > \frac{I(\mu_0)+I(\mu_1)}{2}, \tag{7.5}$$

and this will disprove joint convexity of S_ε for ε small enough.

(c) Finally, we use the property that the Fisher information is not geodesically convex [19]. Specifically, by [19, Exm. 2 in Sect. 3.2] there exists a Wasserstein geodesic $(\tilde{\mu}_t)_{t\in[0,1]}$ on $\mathbb{R}$ such that the function $t \mapsto I(\tilde{\mu}_t)$ is not convex. This implies that we can find t_0, t_1 such that

$$\frac{1}{t_1 - t_0}\int_{t_0}^{t_1} I(\tilde{\mu}_t)\,\mathrm{d}t > \frac{I(\tilde{\mu}_{t_0}) + I(\tilde{\mu}_{t_1})}{2},$$

because if the equality were in the other direction for any t_0, t_1, it would imply convexity of $t \mapsto I(\tilde{\mu}_t)$. Defining $\mu_t := \tilde{\mu}_{t_0+t(t_1-t_0)}$ for $t \in [0,1]$, we obtain (7.5) and conclude.

A scaling argument then extends the lack of convexity to any $\varepsilon > 0$. Indeed, it suffices to define $L^\varepsilon(x) := \varepsilon^{-1/2}x$ and observe that $S_\varepsilon(\mu,\nu) = \varepsilon S_1(L^\varepsilon_\#\mu, L^\varepsilon_\#\nu)$, as already mentioned in (4.11), so that non-convexity of S_ε for one $\varepsilon > 0$ is equivalent to non-convexity for any $\varepsilon > 0$.

At the technical level, however, there is a slight gap in the argument we have just outlined, since (7.4) requires μ_0, μ_1 to have bounded support, while the counterexample built by Carrillo and Slepčev [19, Exm. 2 in Sect. 3.2] has unbounded support. This difficulty can be overcome by a standard truncation procedure, but the resulting argument would be much heavier to read than the straightforward computation of Proposition 7.5.

8. Outlook

In this article we have introduced a Riemannian distance associated with entropic optimal transport and established its most fundamental properties, such as its connection to RKHS spaces, the existence of geodesics, its metrization of the weak-$*$ topology, as well as some instructive examples, like translations, Gaussians, and the two-point space. This Riemannian perspective provides further theoretical underpinning for the widespread application of the Sinkhorn divergence in machine learning. After this foundation has been established, there are various potential applications and more advanced intriguing directions for further study that we comment on now.

Convergence when $\varepsilon \to 0$. One immediate question is, as one sends ε to 0, whether the limit of d_S is the quadratic Monge–Kantorovich distance, and whether there is a corresponding metric convergence in the sense of Gromov–Hausdorff. In Section 4.4 we claim that we should have at least convergence of the metric tensor, but a rigorous proof is missing.

Homogenization when the metric space X is refined. The new Sinkhorn distance is a geodesic distance on any compact metric space (X, d) when $c = d^2$ induces a positive definite universal kernel, in particular also in cases where (X, d) is not a length space itself, such as the two-point space. This is in stark contrast to the quadratic Monge–Kantorovich distance, for which there are no geodesics when X is discrete. A distance on probability measures over discrete graphs that resembles the Monge–Kantorovich distance has been introduced in [48] with non-trivial homogenization results given, for instance, in [34, 35]. For the distance d_S for measures on the two-point space, a consequence of Proposition 6.3 is that $\mathsf{d}_S(\delta_{x_1}, \delta_{x_2}) \approx d(x_1, x_2)$ when $d(x_1, x_2) \ll \sqrt{\varepsilon}$. Based on this one may wonder, whether the metric space $(\mathcal{P}(X), \mathsf{d}_S)$ can be approximated by $(\mathcal{P}(X_n), \mathsf{d}_S)$ where $(X_n)_{n\in\mathbb{N}}$ is a sequence of increasingly fine discrete approximations of X with respect to the ground metric d. In other words, does d_S over $\mathcal{P}(X)$ arise as the homogenization of itself? In combination with the previous question one might even ask whether it is possible to decrease ε to zero at a suitable rate, while X_n is refined, such that the joint limit becomes the quadratic Monge–Kantorovich distance over $\mathcal{P}(X)$. This provides a novel way to define geodesic distances on probability measures over discrete spaces with a potentially consistent continuum limit.

Numerical approximation. A more practical concern is the efficient numerical approximation of d_S and its geodesics. As the Riemannian tensor of d_S was chosen to be the Hessian of the Sinkhorn divergence S_ε, one might conjecture that

$$\mathsf{d}_S(\mu, \nu)^2 = \lim_{N\to\infty} \inf\left\{N \cdot \textstyle\sum_{n=0}^{N-1} S_\varepsilon(\rho_n, \rho_{n+1}) \,\middle|\, \rho_0, \ldots, \rho_N \in \mathcal{P}(X),\ \rho_0 = \mu,\ \rho_N = \nu\right\}$$

in the spirit of time-discrete variational geodesic calculus [66]. For this it seems that one requires a more uniform version of the expansion of Theorem 3.4.

Sample complexity. From a statistical perspective, an interesting question is whether d_S inherits the favorable sample complexity of the Sinkhorn divergence [30, 52].

Linearization. For the quadratic Monge–Kantorovich distance, local approximation via linearization has been established as a useful tool for data analysis applications [75] which reduces the computational complexity and allows one to combine the geometry of optimal transport with standard data analysis tools on Hilbert spaces. If this can be adapted to the new metric d_S it may provide a practically relevant application for geometric data analysis.

Gradient flows. Gradient flow with respect to the Monge–Kantorovich distance is a well-studied topic with important consequences at the theoretical and practical levels. It would be important to understand how the gradient flows with respect to the metric d_S behave.

They should be recovered as a scaling limit if ones tries to use the minimizing movement scheme for gradient flows, but with the Sinkhorn divergence S_ε in place of the more classical squared Monge–Kantorovich distance, for a fixed $\varepsilon > 0$. This may lead to a new class of statistically robust interacting particle methods in high dimensions with potential applications in machine learning, similar to [2, 67]; see also [1].

Extension to non-compact base spaces. In this article we have restricted the analysis to compact base spaces (X, d) and the extension to the non-compact setting will be an interesting question. For this, we expect several challenges. The uniform contraction estimate of Proposition 3.8 will degenerate and so bounded invertibility of $\mathrm{id} - K_\mu^2$ is no longer implied. The self-transport potentials $f_{\mu,\mu}$ may become unbounded and thus the measures $A(\mu)$ may no longer be finite. This poses challenges in the RKHS analysis, including the regularity of the maps A, B, the relation between the tensors $\mathbf{g}_\mu$ and $\tilde{\mathbf{g}}_\mu$, and their equivalence with the norm.

Of course this list is far from complete and in particular more questions will emerge over time, as our understanding of the new distance deepens.

A. Some results of functional analysis

Lemma A.1. *Let Y be a Banach space, X a compact topological space. For $n \in \mathbb{N}$ let $\omega_n \in Y^*$ with $\omega_n \to \omega \in Y^*$ in the weak-$*$ topology. For $x \in X$ let $k_x \in Y$, such that $x \mapsto k_x$ is continuous with respect to the norm topology on Y. Then*

$$\sup_{x \in X} |\langle \omega_n, k_x \rangle - \langle \omega, k_x \rangle| \to 0,$$

so $x \mapsto \langle \omega_n, k_x \rangle$ converges uniformly on X as $n \to \infty$.

We usually apply this with $k_x = k(x, \cdot)$ for some suitable $k : X \times X \to \mathbb{R}$ and $Y = \mathcal{M}(X)$.

Proof of Lemma A.1. Let $\phi_n(x) = \langle \omega_n, k_x \rangle$, $\phi(x) = \langle \omega, k_x \rangle$. By assumption, $\phi, \phi_n \in \mathcal{C}(X)$, $\phi_n \to \phi$ pointwise. The functions ϕ_n are equi-continuous, as for any $n \in \mathbb{N}$,

$$|\phi_n(x) - \phi_n(y)| \leq \|\omega_n\|_{Y^*} \cdot \|k_x - k_y\|_Y,$$

the weak-$*$-convergent sequence $(\omega_n)_n$ is bounded, and the map $x \mapsto k_x$ is uniformly continuous. Uniform convergence now follows from the Arzelà–Ascoli theorem. ∎

Lemma A.2 ([14, Prop. 3.5 (iv)]). *Let Y be a Banach space. For $n \in \mathbb{N}$ let $\omega_n \in Y^*$ with $\omega_n \to \omega \in Y^*$ in the weak-$*$ topology. Let $y_n \in Y$ with $y_n \to y \in Y$ in norm. Then*

$$\lim_{n \to \infty} \langle \omega_n, y_n \rangle = \langle \omega, y \rangle.$$

We also state the following easy lemma which is used in Section 3, while refinements with stronger assumptions on k can be found in Appendix B below.

Lemma A.3. *If $k \in \mathcal{C}^{(m,m)}(X \times X)$ and $H_k[\nu](y) = \langle \nu, k(\cdot, y)\rangle$, then H_k is defined on $\mathcal{C}^m(X)^*$ and valued in $\mathcal{C}^m(X)$. It is weak-$*$-to-norm continuous, and norm-to-norm continuous and compact.*

Proof. We can easily prove $\partial_y^\alpha H_k[\nu](y) = \langle \nu, \partial_y^\alpha k(\cdot, y)\rangle$ for any multi-index α with $|\alpha| \leq m$ as the function $y \mapsto k(\cdot, y)$ is of class $\mathcal{C}^m$ seen as valued in the Banach space $\mathcal{C}^m(X)$. From this, if $\nu_n \to \nu$ weak-$*$ and $y_n \to y$ then $\partial_y^\alpha H_k[\nu_n](y_n) \to \partial_y^\alpha H_k[\nu](y)$, and the convergence can be made uniform with Lemma A.1. Thus H_k is weak-$*$-to-norm continuous, and norm-to-norm compactness follows as the unit ball of $\mathcal{C}^m(X)$ is weak-$*$ compact. ∎

B. Additional results on reproducing kernel Hilbert spaces

Recall the construction of the RKHS $\mathcal{H}_k$ corresponding to a positive definite universal kernel $k \in \mathcal{C}(X \times X)$ from Section 4.1 on a compact Hausdorff space X. A detailed analysis of universal kernels can be found in [72] and [53]. For all $\phi \in \mathcal{H}_k$ and all $x \in X$ it holds that $\phi(x) = \langle \phi, k(x, \cdot)\rangle_{\mathcal{H}_k}$. Using the Cauchy–Schwarz inequality we find

$$\|\phi\|_\infty = \sup_{x\in X} |\langle \phi, k(x,\cdot)\rangle_{\mathcal{H}_k}| \leq \|\phi\|_{\mathcal{H}_k} \cdot \sup_{x\in X} k(x,x) \tag{B.1}$$

for any $\phi \in \mathcal{H}_k$.

We now fix $m \geq 0$ an integer. If $m > 0$, we further impose the analogous version of Assumption 1 for k. That is, we assume that X is the closure of a bounded open set in $\mathbb{R}^d$ and $k \in \mathcal{C}^{(m,m)}(X \times X)$. The case $m = 0$ stands for X a compact metric space and $k \in \mathcal{C}^{(0,0)}(X \times X) = \mathcal{C}(X \times X)$ as above. In this case we have $\mathcal{H}_k \hookrightarrow \mathcal{C}^m(X)$ and $\mathcal{C}^m(X)^* \hookrightarrow \mathcal{H}_k^*$ and the embeddings are continuous [71, Cor. 4]. In this work we need a quantitative estimate on these embeddings, which we present now.

Lemma B.1. *We have $\partial_x^\alpha k(x, \cdot) \in \mathcal{H}_k$ for any $x \in X$ and any $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \leq m$. Moreover, if $\phi \in \mathcal{H}_k$ and $\nu \in \mathcal{C}^m(X)^*$, then*

$$\|\phi\|_{\mathcal{C}^m} \leq \|\phi\|_{\mathcal{H}_k} \cdot \|k\|_{\mathcal{C}^{(m,m)}}^{\frac{1}{2}}, \quad \|\nu\|_{\mathcal{H}_k^*} \leq \|\nu\|_{\mathcal{C}^{m},*} \cdot \|k\|_{\mathcal{C}^{(m,m)}}^{\frac{1}{2}}.$$

Proof. Using that H_k is defined on $\mathcal{H}_k^*$ and valued in $\mathcal{H}_k$, as well as $\mathcal{C}^m(X)^* \hookrightarrow \mathcal{H}_k^*$, we have $H_k[\partial^\alpha|_x](y) = \langle \partial^\alpha|_x, k(\cdot, y)\rangle = \partial_x^\alpha k(x, y) \in \mathcal{H}_k$ with $\|H_k[\partial^\alpha|_x]\|_{\mathcal{H}_k}^2 = \partial^{(\alpha,\alpha)} k(x,x) \leq \|k\|_{\mathcal{C}^{(m,m)}}$. We get the bound on $\|\phi\|_{\mathcal{C}^m}$ from

$$|\partial^\alpha \phi(x)| = |\langle \partial^\alpha|_x, \phi\rangle| = |\langle H_k[\partial^\alpha|_x], \phi\rangle_{\mathcal{H}_k}| \leq \|H_k[\partial^\alpha|_x]\|_{\mathcal{H}_k} \|\phi\|_{\mathcal{H}_k},$$

and the one on $\|\nu\|_{\mathcal{H}_k^*}$ by duality. ∎

Lemma B.2. *We have the following weak-to-norm injections:*

(1) *The injection of $\mathcal{H}_k \hookrightarrow \mathcal{C}^m(X)$ is weak-to-norm continuous, hence norm-to-norm compact.*

(2) *The operator $H_k : \mathcal{C}^m(X)^* \to \mathcal{H}_k$ is weak-$*$-to-norm continuous, hence norm-to-norm compact.*

Proof. For the first point, let $\phi_n \in \mathcal{H}_k$ converge weakly to $\phi \in \mathcal{H}_k$. Then from the proof of Lemma B.1, $\partial^\alpha \phi_n(x) = \langle \phi_n, \partial_x^\alpha k(x,\cdot)\rangle_{\mathcal{H}_k} \to \langle \phi, \partial_x^\alpha k(x,\cdot)\rangle_{\mathcal{H}_k} = \partial^\alpha \phi(x)$ for all $x \in X$. Uniform convergence for any allowed α follows from Lemma A.1.

For the second point let $\nu_n \to \nu$ weak-$*$ in $\mathcal{C}^m(X)^*$. From the continuity of the injection $\mathcal{C}^m(X)^* \hookrightarrow \mathcal{H}_k^*$ we obtain $H_k[\nu_n] \to H_k[\nu]$ weakly in $\mathcal{H}_k$, which by the first point implies norm convergence of $H_k[\nu_n]$ to $H_k[\nu]$ in $\mathcal{C}^m(X)$. Thus Lemma A.2 applied to the duality pairing between $\mathcal{C}^m(X)$ and $\mathcal{C}^m(X)^*$ shows that $\|H_k[\nu_n] - H_k[\nu]\|_{\mathcal{H}_k}^2 = \langle \nu_n - \nu, H_k[\nu_n] - H_k[\nu]\rangle \to 0$.

Finally, the claims of compactness follow from the Banach–Alaoglu theorem. ■

To finish this section, we briefly show that the universality of k_c implies universality of k_μ for all $\mu \in \mathcal{P}(X)$: Note that $\exp(f_{\mu,\mu}/\varepsilon)$ is bounded from above and away from zero. Now for any $\phi \in \mathcal{C}(X)$ an approximation $\|\sum_{i=1}^n a_i k_c(x_i,.) - \phi/\exp(f_{\mu,\mu}/\varepsilon)\|_\infty \le \delta$ yields

$$\left\| \sum_{i=1}^n \frac{a_i}{\exp(f_{\mu,\mu}(x_i)/\varepsilon)} k_\mu(x_i,.) - \phi \right\|_\infty \le \delta \cdot \|\exp(f_{\mu,\mu}/\varepsilon)\|_\infty .$$

C. The Sobolev space of Hilbert space-valued functions

We briefly summarize a few results from [40, Sect. 3.2]. Let $\mathcal{H}$ be a Hilbert space. We write $L^2((0,1);\mathcal{H})$ for the space of measurable functions $u : (0,1) \to \mathcal{H}$ for which $\|u(\cdot)\|_{\mathcal{H}}$ is 2-integrable, up to equality almost everywhere. It carries the obvious inner product. We say that a function $u \in L^2((0,1);\mathcal{H})$ has the distributional derivative $v \in L^2((0,1);\mathcal{H})$ and write $v = u'$ if for all $\phi \in \mathcal{C}_c^\infty((0,1);\mathbb{R})$ it holds that

$$\int_0^1 u(t)\phi'(t)\,\mathrm{d}t = -\int_0^1 v(t)\phi(t)\,\mathrm{d}t \tag{C.1}$$

as Bochner integrals. The first Sobolev space is given by

$$\mathcal{H}^1((0,1);\mathcal{H}) = \{u : (0,1) \to \mathcal{H} \mid u \text{ has a distr. derivative } u' \text{ and } u, u' \in L^2((0,1);\mathcal{H})\}.$$

It is a Hilbert space with the inner product $\langle u_1, u_2\rangle_{\mathcal{H}^1} = \langle u_1, u_2\rangle_{L^2} + \langle u_1', u_2'\rangle_{L^2}$.

Theorem C.1 ([40, Thm. 2.5]). *A measurable function $u : (0,1) \to \mathcal{H}$ is Bochner integrable if and only if $\|u(\cdot)\|_{\mathcal{H}}$ is Lebesgue integrable. In this case $\int_0^1 u(t)\,\mathrm{d}t \in \mathcal{H}$ and it holds that*

$$\left\| \int_0^1 u(t)\,\mathrm{d}t \right\|_{\mathcal{H}} \le \int_0^1 \|u(t)\|_{\mathcal{H}}\,\mathrm{d}t.$$

Theorem C.2. *Let $u \in L^2((0,1);\mathcal{H})$. Then the following are equivalent:*

(1) $u \in \mathcal{H}^1((0,1);\mathcal{H})$.

(2) $u \in \mathrm{AC}^2((0,1);\mathcal{H})$ *in the sense of Definition 5.7, i.e. there exist a representation of u and a function $m \in L^2((0,1);\mathbb{R})$, such that $\|u(t)-u(s)\|_{\mathcal{H}} \le |\int_s^t m(\tau)\,\mathrm{d}\tau|$ for all $t, s \in (0,1)$.*

(3) *u is differentiable a.e., $u' \in L^2((0,1);\mathcal{H})$, and there is a $t_0 \in (0,1)$, such that for almost all t it holds that*

$$u(t) = u(t_0) + \int_{t_0}^t u'(\tau)\,\mathrm{d}\tau.$$

If $u \in \mathcal{H}^1((0,1);\mathcal{H})$ then $m(\tau) = \|u'(\tau)\|_{\mathcal{H}}$ is a possible choice for point (2).

Proof. (1) $\Rightarrow$ (3): [40, Thm. 3.7, Prop. 3.8]. (3) $\Rightarrow$ (2): Theorem C.1. (2) $\Rightarrow$ (1): [40, Thm. 3.13]. ∎

As a consequence of absolute continuity, every $u \in \mathcal{H}^1((0,1);\mathcal{H})$ continuously extends to the boundary points $\{0,1\}$ of the interval $(0,1)$.

Lemma C.3. *Let $\mathcal{H}$ be separable, $u, u' \in L^2((0,1);\mathcal{H})$. Then $u \in \mathcal{H}^1((0,1);\mathcal{H})$ with distributional derivative u' in the sense of* (C.1) *if and only if*

$$\int_0^1 \langle u(t), \phi'(t)\rangle_{\mathcal{H}}\,\mathrm{d}t = -\int_0^1 \langle u'(t), \phi(t)\rangle_{\mathcal{H}}\,\mathrm{d}t \tag{C.2}$$

for all $\phi \in \mathcal{C}_c^\infty((0,1);\mathcal{H})$.

Proof. Let $\psi_1, \psi_2, \ldots \in \mathcal{H}$ be an orthonormal basis. By [40, Prop. 3.8], we have $u \in \mathcal{H}^1((0,1);\mathcal{H})$ with derivative u' if and only if $\langle u, \psi_i\rangle_{\mathcal{H}} \in \mathrm{AC}^2((0,1);\mathbb{R}) = \mathcal{H}^1((0,1);\mathbb{R})$ with $\partial_t \langle u(t), \psi_i\rangle_{\mathcal{H}} = \langle u'(t), \psi_i\rangle_{\mathcal{H}}$ for all $i \in \mathbb{N}$. We first show that (C.2) holds under the original definition. For any $\omega_i \in \mathcal{C}_c^\infty((0,1);\mathbb{R})$ we get

$$\begin{aligned}\int_0^1 \langle u(t), \partial_t(\omega_i(t)\psi_i)\rangle_{\mathcal{H}}\,\mathrm{d}t &= \int_0^1 \omega_i'(t)\langle u(t), \psi_i\rangle_{\mathcal{H}}\,\mathrm{d}t\\ &= -\int_0^1 \omega_i(t)\langle u'(t), \psi_i\rangle_{\mathcal{H}}\,\mathrm{d}t\\ &= -\int_0^1 \langle u'(t), \omega_i(t)\psi_i\rangle_{\mathcal{H}}\,\mathrm{d}t.\end{aligned} \tag{C.3}$$

Thus (C.2) holds for functions of the form $\phi(t) = \sum_{i=1}^N \omega_i(t)\psi_i$. Such functions are dense in $\mathcal{C}_c^\infty((0,1);\mathcal{H})$ with respect to the norm on the larger space $\mathcal{H}^1((0,1);\mathcal{H})$. As the left and right-hand sides of (C.2) respectively are bounded by $\|u\|_{L^2((0,1);\mathcal{H})} \cdot \|\phi\|_{\mathcal{H}^1((0,1);\mathcal{H})}$ and $\|u'\|_{L^2((0,1);\mathcal{H})} \cdot \|\phi\|_{\mathcal{C}((0,1);\mathcal{H})}$, the claim follows from an approximation argument.

The converse statement is obtained by the reverse argument, where we obtain equality of the middle terms in (C.3) from equality of the outer terms. ∎

Acknowledgments. The research leading to this work started during a visit of BS to HL and LT in the Department of Decision Sciences of Bocconi University, which is warmly acknowledged for the hospitality. We also thank the two anonymous referees for their many suggestions which improved the quality and scope of this article.

Funding. HL and LT acknowledge the support of MUR-Prin 2022-202244A7YL "Gradient Flows and Non-Smooth Geometric Structures with Applications to Optimization and Machine Learning", funded by the European Union – Next Generation EU. LT is also a member of INdAM and the GNAMPA group. GM gratefully acknowledges funding of the DFG within CRC1456-A04. BS gratefully acknowledges funding of the DFG within the Emmy Noether programme and in project CRC1456-A03.

References

[1] M. Agarwal, Z. Harchaoui, G. Mulcahy, and S. Pal, Iterated Schrödinger bridge approximation to Wasserstein gradient flows. 2024, arXiv:2406.10823v1

[2] M. Agarwal, Z. Harchaoui, G. Mulcahy, and S. Pal, Langevin diffusion approximation to same marginal Schrödinger bridge. 2025, arXiv:2505.07647v1

[3] M. Agueh and G. Carlier, Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** (2011), no. 2, 904–924 Zbl 1223.49045 MR 2801182

[4] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows in metric spaces and in the space of probability measures*. 2nd edn., Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2008 Zbl 1145.35001 MR 2401600

[5] V. Arnold, Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Ann. Inst. Fourier (Grenoble)* **16** (1966), no. 1, 319–361 Zbl 0148.45301 MR 0202082

[6] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numer. Math.* **84** (2000), no. 3, 375–393 Zbl 0968.76069 MR 1738163

[7] J.-D. Benamou, G. Carlier, S. Di Marino, and L. Nenna, An entropy minimization approach to second-order variational mean-field games. *Math. Models Methods Appl. Sci.* **29** (2019), no. 8, 1553–1583 Zbl 07139235 MR 3986799

[8] J.-D. Benamou, G. Carlier, and F. Santambrogio, Variational mean field games. In *Active particles. Vol. 1. Advances in theory, models, and applications*, pp. 141–171, Model. Simul. Sci. Eng. Technol., Birkhäuser/Springer, Cham, 2017 MR 3644590

[9] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004 Zbl 1145.62002 MR 2239907

[10] P. J. Bickel and D. A. Freedman, Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** (1981), no. 6, 1196–1217 Zbl 0449.62034 MR 0630103

[11] J. Bigot, R. Gouet, T. Klein, and A. López, Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** (2017), no. 1, 1–26 Zbl 1362.62065 MR 3606732

[12] Y. Brenier, The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. Amer. Math. Soc.* **2** (1989), no. 2, 225–255 Zbl 0697.76030 MR 0969419

[13] Y. Brenier and D. Vorotnikov, On optimal transport of matrix-valued measures. *SIAM J. Math. Anal.* **52** (2020), no. 3, 2849–2873 Zbl 1460.49036 MR 4111672

[14] H. Brezis, *Functional analysis, Sobolev spaces and partial differential equations*. Universitext, Springer, New York, 2011 Zbl 1220.46002 MR 2759829

[15] C. Bunne, Y.-P. Hsieh, M. Cuturi, and A. Krause, The Schrödinger bridge between Gaussian measures has a closed form. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 5802–5833, Proceedings of Machine Learning Research 206, PMLR, 2023

[16] G. Carlier, L. Chizat, and M. Laborde, Displacement smoothness of entropic optimal transport. *ESAIM Control Optim. Calc. Var.* **30** (2024), article no. 25 Zbl 1537.49038 MR 4729266

[17] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer, Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.* **49** (2017), no. 2, 1385–1418 Zbl 1365.90197 MR 3635459

[18] G. Carlier and M. Laborde, A differential approach to the multi-marginal Schrödinger system. *SIAM J. Math. Anal.* **52** (2020), no. 1, 709–717 Zbl 1436.49058 MR 4062805

[19] J. A. Carrillo and D. Slepčev, Example of a displacement convex functional of first order. *Calc. Var. Partial Differential Equations* **36** (2009), no. 4, 547–564 Zbl 1180.35293 MR 2558330

[20] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.* **274** (2018), no. 11, 3090–3123 Zbl 1387.49066 MR 3782988

[21] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré, Faster Wasserstein distance estimation with the Sinkhorn divergence. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 2257–2269, Curran Associates, 2020

[22] G. Conforti and L. Tamanini, A formula for the time derivative of the entropic cost and applications. *J. Funct. Anal.* **280** (2021), no. 11, article no. 108964 Zbl 1462.35476 MR 4232667

[23] S. I. R. Costa, S. A. Santos, and J. E. Strapasson, Fisher information distance: A geometrical reading. *Discrete Appl. Math.* **197** (2015), 59–69 Zbl 1329.94027 MR 3398961

[24] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transportation distances. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 2292–2300, Curran Associates, 2013

[25] A. Delalande and Q. Mérigot, Quantitative stability of optimal transport maps under variations of the target measure. *Duke Math. J.* **172** (2023), no. 17, 3321–3357 Zbl 1531.49027 MR 4688680

[26] S. Di Marino and A. Gerolin, An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *J. Sci. Comput.* **85** (2020), no. 2, article no. 27 Zbl 1478.49043 MR 4164428

[27] J. Dieudonné, *Foundations of modern analysis*. Pure and Applied Mathematics 10-I, Academic Press, New York-London, 1969 Zbl 0176.00502 MR 0349288

[28] 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 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, Proceedings of Machine Learning Research 89, PMLR, 2019

[29] M. Gelbrich, On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Math. Nachr.* **147** (1990), 185–203 Zbl 0711.60003 MR 1127323

[30] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré, Sample complexity of Sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1574–1583, Proceedings of Machine Learning Research 89, PMLR, 2019

[31] A. Genevay, G. Peyre, and M. Cuturi, Learning generative models with Sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, Proceedings of Machine Learning Research 84, PMLR, 2018

[32] I. Gentil, C. Léonard, and L. Ripani, About the analogy between optimal transport and minimal entropy. *Ann. Fac. Sci. Toulouse Math. (6)* **26** (2017), no. 3, 569–601 Zbl 1380.49067 MR 3669966

[33] N. Gigli, Second order analysis on $(\mathscr{P}_2(M), W_2)$. *Mem. Amer. Math. Soc.* **216** (2012), no. 1018, xii+154 Zbl 1253.58008 MR 2920736

[34] N. Gigli and J. Maas, Gromov–Hausdorff convergence of discrete transportation metrics. *SIAM J. Math. Anal.* **45** (2013), no. 2, 879–899 Zbl 1268.49054 MR 3045651

[35] P. Gladbach, E. Kopfer, and J. Maas, Scaling limits of discrete optimal transport. *SIAM J. Math. Anal.* **52** (2020), no. 3, 2759–2802 Zbl 1447.49062 MR 4110822

[36] Z. Goldfeld, K. Kato, G. Rioux, and R. Sadhu, Limit theorems for entropic optimal transport maps and Sinkhorn divergence. *Electron. J. Stat.* **18** (2024), no. 1, 980–1041 Zbl 07823233 MR 4718466

[37] A. Gonzalez-Sanz, J.-M. Loubes, and J. Niles-Weed, Weak limits of entropy regularized optimal transport; potentials, plans and divergences. [v1] 2022, [v2] 2024, arXiv:2207.07427v2

[38] H. Hotelling, Spaces of statistical parameters. *Bull. Am. Math. Soc.* **36** (1930), 191 Zbl 56.0463.24

[39] S. Kondratyev, L. Monsaingeon, and D. Vorotnikov, A new optimal transport distance on the space of finite Radon measures. *Adv. Differential Equations* **21** (2016), no. 11-12, 1117–1164 Zbl 1375.49062 MR 3556762

[40] M. Kreuter, *Sobolev spaces of vector-valued functions*. Master's thesis, Ulm University, 2015

[41] S. Lang, *Fundamentals of differential geometry*. Grad. Texts in Math. 191, Springer, New York, 1999 Zbl 0932.53001 MR 1666820

[42] V. Laschos and A. Mielke, Geometric properties of cones with applications on the Hellinger-Kantorovich space, and a new distance on the space of probability measures. *J. Funct. Anal.* **276** (2019), no. 11, 3529–3576 Zbl 1422.60009 MR 3944304

[43] H. Lavenant, Harmonic mappings valued in the Wasserstein space. *J. Funct. Anal.* **277** (2019), no. 3, 688–785 Zbl 1419.31006 MR 3954993

[44] C. Léonard, From the Schrödinger problem to the Monge–Kantorovich problem. *J. Funct. Anal.* **262** (2012), no. 4, 1879–1920 Zbl 1236.49088 MR 2873864

[45] C. Léonard, A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst.* **34** (2014), no. 4, 1533–1574 Zbl 1277.49052 MR 3121631

[46] C. Léonard, On the convexity of the entropy along entropic interpolations. In *Measure theory in non-smooth spaces*, pp. 194–242, Partial Differ. Equ. Meas. Theory, De Gruyter Open, Warsaw, 2017 Zbl 1485.28018 MR 3701740

[47] M. Liero, A. Mielke, and G. Savaré, Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Invent. Math.* **211** (2018), no. 3, 969–1117 Zbl 1412.49089 MR 3763404

[48] J. Maas, Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.* **261** (2011), no. 8, 2250–2292 Zbl 1237.60058 MR 2824578

[49] A. Mallasto, A. Gerolin, and H. Q. Minh, Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Inf. Geom.* **5** (2022), no. 1, 289–323 MR 4452151

[50] N. F. Marshall and R. R. Coifman, Manifold learning with bi-stochastic kernels. *IMA J. Appl. Math.* **84** (2019), no. 3, 455–482 Zbl 1472.68148 MR 3954914

[51] R. J. McCann, A convexity principle for interacting gases. *Adv. Math.* **128** (1997), no. 1, 153–179 Zbl 0901.49012 MR 1451422

[52] G. Mena and J. Niles-Weed, Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Curran Associates, 2019

[53] C. A. Micchelli, Y. Xu, and H. Zhang, Universal kernels. *J. Mach. Learn. Res.* **7** (2006), 2651–2667 Zbl 1222.68266 MR 2274454

[54] T. Mikami, Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probab. Theory Related Fields* **129** (2004), no. 2, 245–260 Zbl 1061.58034 MR 2063377

[55] G. Mordant, The entropic optimal (self-) transport problem: Limit distributions for decreasing regularization with application to score function estimation. 2024, arXiv:2412.12007

[56] M. Nutz, Introduction to entropic optimal transport. Lecture notes, Columbia University, https://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf visited on 26 September 2025

[57] M. Nutz and J. Wiesel, Entropic optimal transport: Convergence of potentials. *Probab. Theory Related Fields* **184** (2022), no. 1-2, 401–424 Zbl 1504.90095 MR 4498514

[58] M. Nutz and J. Wiesel, Stability of Schrödinger potentials and convergence of Sinkhorn's algorithm. *Ann. Probab.* **51** (2023), no. 2, 699–722 Zbl 1518.90069 MR 4546630

[59] F. Otto, The geometry of dissipative evolution equations: The porous medium equation. *Comm. Partial Differential Equations* **26** (2001), no. 1-2, 101–174 Zbl 0984.35089 MR 1842429

[60] V. I. Paulsen and M. Raghupathi, *An introduction to the theory of reproducing kernel Hilbert spaces*. Cambridge Stud. Adv. Math. 152, Cambridge University Press, Cambridge, 2016 Zbl 1364.46004 MR 3526117

[61] G. Peyré and M. Cuturi, Computational optimal transport. *Found. Trends Mach. Learn.* **11** (2018), no. 5-6, 355–607

[62] R. Peyre, Comparison between W_2 distance and $\dot{\mathrm{H}}^{-1}$ norm, and localization of Wasserstein distance. *ESAIM Control Optim. Calc. Var.* **24** (2018), no. 4, 1489–1501 Zbl 1415.49031 MR 3922440

[63] C. Radhakrishna Rao, Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37** (1945), 81–91 Zbl 0063.06420 MR 0015748

[64] A. Ramdas, N. García Trillos, and M. Cuturi, On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** (2017), no. 2, article no. 47 MR 3608466

[65] W. Rudin, *Real and complex analysis*. 3rd edn., McGraw-Hill, New York, 1987 Zbl 0925.00005 MR 0924157

[66] M. Rumpf and B. Wirth, Variational time discretization of geodesic calculus. *IMA J. Numer. Anal.* **35** (2015), no. 3, 1011–1046 Zbl 1329.65133 MR 3407252

[67] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré, Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence And Statistics*, pp. 3515–3530, Proceedings of Machine Learning Research 151, PMLR, 2022

[68] F. Santambrogio, *Optimal transport for applied mathematicians*. Progr. Nonlinear Differential Equations Appl. 87, Birkhäuser/Springer, Cham, 2015 Zbl 1401.49002 MR 3409718

[69] C. Sarrazin and B. Schmitzer, Linearized optimal transport on manifolds. *SIAM J. Math. Anal.* **56** (2024), no. 4, 4970–5016 Zbl 1556.49031 MR 4770386

[70] G. Savaré and G. E. Sodini, A simple relaxation approach to duality for optimal transport problems in completely regular spaces. *J. Convex Anal.* **29** (2022), no. 1, 1–12 Zbl 1486.49046 MR 4403781

[71] C.-J. Simon-Gabriel and B. Schölkopf, Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *J. Mach. Learn. Res.* **19** (2018), article no. 44 Zbl 1467.62057 MR 3874152

[72] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.* **12** (2011), 2389–2410 Zbl 1280.68198 MR 2825431

[73] A. Takatsu, Wasserstein geometry of Gaussian measures. *Osaka J. Math.* **48** (2011), no. 4, 1005–1026 Zbl 1245.60013 MR 2871291

[74] C. Villani, *Optimal transport*. Grundlehren Math. Wiss. 338, Springer, Berlin, 2009 Zbl 1156.53003 MR 2459454

[75] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *Int. J. Comput. Vis.* **101** (2013), no. 2, 254–269 Zbl 1259.68222 MR 3021062

Received 3 September 2024; revised 1 July 2025; accepted 29 August 2025.

Hugo Lavenant
Department of Decision Sciences and BIDSA, Bocconi University, Via Roberto Sarfatti 25, 20136 Milan, Italy; hugo.lavenant@unibocconi.it

Jonas Luckhardt
Institute for Computer Science, Georg-August-Universität Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany; jonas.luckhardt@uni-goettingen.de

Gilles Mordant
Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany; Department of Applied and Computational Mathematics, Yale University, 12 Hillhouse Avenue, New Haven, CT 06511, USA;
gilles.mordant@uni-goettingen.de, gilles.mordant@yale.edu

Bernhard Schmitzer
Institute for Computer Science, Georg-August-Universität Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany; schmitzer@cs.uni-goettingen.de

Luca Tamanini
Dipartimento di Matematica e Fisica "Niccolò Tartaglia", Università Cattolica del Sacro Cuore, Via della Garzetta 48, 25133 Brescia, Italy; luca.tamanini@unicatt.it