Information Geometry (2022) 5:289–323
https://doi.org/10.1007/s41884-021-00052-8

RESEARCH PAPER

Entropy-regularized 2-Wasserstein distance between Gaussian measures

Anton Mallasto[1] · Augusto Gerolin[2] · Hà Quang Minh[3]

Received: 24 December 2019 / Revised: 16 July 2021 / Accepted: 17 July 2021 /
Published online: 16 August 2021

Abstract
Gaussian distributions are plentiful in applications dealing in uncertainty quantification and diffusivity. They furthermore stand as important special cases for frameworks providing geometries for probability measures, as the resulting geometry on Gaussians is often expressible in closed-form under the frameworks. In this work, we study the Gaussian geometry under the entropy-regularized 2-Wasserstein distance, by providing closed-form solutions for the distance and interpolations between elements. Furthermore, we provide a fixed-point characterization of a population barycenter when restricted to the manifold of Gaussians, which allows computations through the fixed-point iteration algorithm. As a consequence, the results yield closed-form expressions for the 2-Sinkhorn divergence. As the geometries change by varying the regularization magnitude, we study the limiting cases of vanishing and infinite magnitudes, reconfirming well-known results on the limits of the Sinkhorn divergence. Finally, we illustrate the resulting geometries with a numerical study.

Keywords Sinkhorn divergences · Multivariate Gaussian measures · Optimal transportation theory

✉ Anton Mallasto
anton.mallasto@aalto.fi

Augusto Gerolin
augustogerolin@gmail.com

Hà Quang Minh
minh.haquang@riken.jp

[1] Department of Computer Science, Aalto University, Helsinki, Finland

[2] Department of Theoretical Chemistry, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[3] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

1 Introduction

Optimal transport (OT) [82] studies the geometry of probability measures through the lifting of a cost function between samples. This is carried out by devising a coupling between two probability measures via a *transport plan*, so that one measure is transported to another with minimal total cost. The resulting geometry offers a favorable way of comparing probability measures one to another, which has lead to considerable success in machine learning, especially in generative modelling [6,24,28,55], where one aims at training a model distribution to sample from a given data distribution, and computer vision, where OT provides intuitive metrics between images [71]. Notably, OT can not only be used to derive divergences, but also metrics between probability distributions, referred to as the *p-Wasserstein metrics*.

To ease the computational aspects of OT, entropic relaxation was introduced, which transforms the constrained convex problem of transportation into an unconstrained strictly convex problem [20]. This is carried out via considering the sum of the total cost and the Kullbackk–Leibler (KL) divergence, between the transport plan and the independent joint distribution, scaled by some regularization magnitude. In addition to computational aspects, the entropic regularization also betters statistical properties [76], specifically, the complexity of estimating the OT quantity between measures through sampling [34,59,83]. Theoretical properties of the entropic regularization have been studied in e.g. metric geometry, machine learning and statistics [30,35,36,40,56,69,70]. It has also been applied in a variety of fields, including computer vision, density functional theory in chemistry, and inverse problems (e.g. [36,38,52,67]).

The resulting problem has close relations to the *Schrödinger problem* [75], which considers the most likely flow of a cloud of gas from an initial position to an observed position after a certain amount of time under a prior assumption on the evolution of the position, given by e.g. a Brownian motion. The resulting problem has found applications in fields such as mathematical physics, economics, optimization and probability [12,19,22,31,32,72,84]). Connections to OT have been considered in e.g. [20,33,50,73,74].

OT is not the only instance of a geometric framework for probability measures. Other popular choices include *information geometric divergences* [3,9] and *integral probability metrics* [64]. In contrast to these methods, OT and entropic OT has the advantage of metrizing the weak*-convergence of probability measures, which results in non-singular behavior when comparing measures of disjoint supports. On top of this, being able to decide the lifted cost function is important in applications, as the cost function can be used to incorporate modelling choices, determining which differences in samples are deemed most important. For example, the standard Euclidean metric is a poor choice for comparing images.

Gaussian distributions provide a meaningful testing ground for such frameworks since, in many cases, they result in closed-form expressions. In addition, the study of Gaussians under the OT framework result in useful divergences. In particular, divergences between centered Gaussians result in divergences between their corresponding covariance matrices. Both instances enjoy many applications in a plethora of fields, such as medical imaging [27], computer vision [79–81], brain computer interfaces [18], natural language processing [65], and assessing the quality of generative models [43]. Notably, the 2-Wasserstein metric between Gaussians is known as the *Bures metric* in quantum physics, where it is used to compare quantum states. Other popular divergences for Gaussians include the *affine-invariant Riemannian metric* [68], corresponding to the Fisher–Rao distance between centered Gaussians, the Alpha Log-Determinant divergences [14], corresponding to Rényi divergences between centered Gaussians, and the *log-Euclidean metric* [8]. A survey of some of the most common divergences and their resulting geometry on Gaussians can be found in [29]. More recently, applications have driven research into allowing determining optimal divergences for the task at hand, which has raised interest in studying interpolations between different divergences [4,17,78]. Generalizations of these divergences to the infinite-dimensional setting of Gaussian processes and covariance operators have also been considered [49,54,57,60,61].

The *Sinkhorn divergence* has been proposed in OT, applying the entropic regularization to define a parametric family of divergences, interpolating from the OT quantity to a *maximum mean discrepancy* (MMD), whose kernel is determined by the cost. In the present work, we provide a closed-form solution to the entropy-regularized 2-Wasserstein distance between multivariate Gaussians, which can then be applied in the computation of the corresponding Sinkhorn divergence between Gaussians. In addition, we study the task of interpolating between two Gaussians under the entropy-regularized 2-Wasserstein distance, and confirm known limiting properties of the divergences with respect to the regularization strength. Finally, we provide fixed-point expressions for the barycenter of population of Gaussians restricted to the Gaussian manifold, that can be employed in fixed-point iteration for computing the barycenter. The one-dimensional setting has been studied in [4,37]. The Schrödinger bridge between multivariate Gaussians has been considered in [15], including the study of the limiting case of bringing the noise of the driving Brownian motion to 0, resulting in the 2-Wasserstein case, in [16].

The paper is divided as follows: in Sect. 2, we briefly introduce the necessary background to develop the entropic OT theory of Gaussians, including the formulation of OT, entropic OT, and the corresponding dual and dynamical formulations. In Sect. 3, we compute explicit solutions to the entropy-relaxed 2-Wasserstein distance between Gaussians, including the dynamical formulation that allows for interpolation. As a consequence, we derive a closed-form solution for the corresponding Sinkhorn divergence. In Sect. 4, we study the *barycenters* of populations of Gaussians, restricted to the Gaussian manifold. We derive fixed-point expressions for the entropic 2-Wasserstein distance and the 2-Sinkhorn divergence. Finally, in Sect. 5, we illustrate the resulting interpolative and barycentric schemes. Especially, we consider varying the regularization magnitude, visualizing the interpolation between the OT and MMD problems in the Sinkhorn case [30,36,69].

Related work Several papers—all independently—have formulated the closed form solution of the Entropic regularized Optimal Transport for Gaussian measures [23,45] in any dimensions, including the case of unbalanced transport [45]. These results have been generalized for φ-exponential distributions [48], Gaussian measures on infinite-dimensional Hilbert spaces, including in particular Reproducing Kernel Hilbert Spaces, and Gaussian processes [62,63]. Both two and multi-marginal solution in the one-dimensional case first appeared in [38].

2 Background

In this section, we start by recalling the essential background for optimal transport (OT) and its entropy-relaxed version. More in-depth exposition for OT can be found in [82], and for computational aspects and entropic OT in [22].

2.1 Optimal transport

Let $(\mathcal{X}, d)$ be a metric space equipped with a lower semi-continuous *cost function* $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. Then, the optimal transport problem between two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is given by

$$\mathrm{OT}(\mu, \nu) = \min_{\gamma \in \mathrm{ADM}(\mu,\nu)} \mathbb{E}_\gamma[c], \tag{1}$$

where $\mathrm{ADM}(\mu, \nu)$ is the set of joint probabilities with marginals μ and ν, and $\mathbb{E}_\mu[f]$ denotes the expected value of f under μ

$$\mathbb{E}_\mu[f] = \int_{\mathcal{X}} f(x) \mathrm{d}\mu(x). \tag{2}$$

Additionally, by $\mathbb{E}[\mu]$ we denote the expectation of μ. A minimizer of (1) is denoted by γ_{opt} and called a *transport plan*.

The OT problem admits the following *Kantorovich (dual) formulation*

$$\mathrm{OT}(\mu, \nu) = \max_{\varphi, \psi \in \mathrm{ADM}(c)} \left\{ \mathbb{E}_\mu[\varphi] + \mathbb{E}_\nu[\psi] \right\}, \tag{3}$$

where $(\varphi, \psi) \in \mathrm{ADM}(c)$ is required to satisfy

$$\varphi(x) + \psi(y) \leq c(x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{X}. \tag{4}$$

Potentials $\varphi_{\mathrm{opt}}, \psi_{\mathrm{opt}}$ achieving the maximum in (3) are called *Kantorovich potentials*.

2.2 Wasserstein distances

The p-Wasserstein distance W_p between μ and ν is defined as

$$W_p(\mu, \nu) = \mathrm{OT}_{d^p}(\mu, \nu)^{\frac{1}{p}}, \tag{5}$$

where d is a metric on X and $p \geq 1$. The case $p = 2$ is particularly interesting, as the resulting metric is then induced by a pseudo-Riemannian metric structure [5,53].

2.3 2-Wasserstein distance between Gaussians

One of the rare cases where the 2-Wasserstein distance admits a closed form solution is between two multivariate Gaussian distributions $\mu_i = \mathcal{N}(m_i, K_i)$, $i = 0, 1$ with $d(x, y) = \|x - y\|$, which is given by [26,42,46,66]

$$W_2^2(\mu_0, \mu_1) = ||m_0 - m_1||^2 + \mathrm{Tr}(K_0) + \mathrm{Tr}(K_1) - 2\mathrm{Tr}\left(K_1^{\frac{1}{2}} K_0 K_1^{\frac{1}{2}}\right)^{\frac{1}{2}}. \quad (6)$$

It can be shown that (6) is induced by a *Riemannian metric* in the space of n-dimensional Gaussians $\mathcal{N}(\mathbb{R}^n)$, with the metric $g_K : T_K\mathcal{N}(\mathbb{R}^n) \times T_K\mathcal{N}(\mathbb{R}^n) \to \mathbb{R}$ given by [77]

$$g_K(U, V) = \mathrm{Tr}\left[v_{(K,U)} K v_{(K,V)}\right], \quad \forall\, K \in \mathcal{N}(\mathbb{R}^n),\ U, V \in T_K\mathcal{N}(\mathbb{R}^n), \quad (7)$$

where $v_{(K,V)}$ denotes the unique symmetric matrix solving the *Sylvester equation*

$$V = K v_{(K,V)} + v_{(K,V)} K. \quad (8)$$

Moreover, given $\mathcal{N}(m_0, K_0), \mathcal{N}(m_1, K_1) \in \mathcal{N}(\mathbb{R}^n)$, the geodesics under the metric (6) are given by $\mathcal{N}(m_t, K_t)$, with [58]

$$\begin{aligned} m_t &= (1-t)m_0 + tm_1, \\ K_t &= \left((1-t)I + tK_0^{-\frac{1}{2}}\left(K_0^{\frac{1}{2}} K_1 K_0^{\frac{1}{2}}\right)^{\frac{1}{2}} K_0^{-\frac{1}{2}}\right) K_0 \\ &\quad \times \left((1-t)I + tK_0^{-\frac{1}{2}}\left(K_0^{\frac{1}{2}} K_1 K_0^{\frac{1}{2}}\right)^{\frac{1}{2}} K_0^{-\frac{1}{2}}\right) \\ &= (1-t)^2 K_0 + t^2 K_1 + t(1-t)[(K_0K_1)^{1/2} + (K_1K_0)^{1/2}]. \end{aligned} \quad (9)$$

We remark that Eq. (6) is valid for all Gaussian distributions, including the case when K_0, K_1 are positive semi-definite. This is in contrast to the affine-invariant Riemannian distance $|| \log(K_0^{-1/2} K_1 K_0^{-1/2})||_F$, the Log-Euclidean distance $|| \log(K_0) - \log(K_1)||_F$, and the Kullback–Leibler divergence (see below), which require that K_0, K_1 be strictly positive definite.

Finally, the 2-Wasserstein barycenter $\bar{\mu}$ of a population of probability measures μ_i with weights $\lambda_i \geq 0$, $i = 1, 2, \ldots, N$ and $\sum_{i=1}^N \lambda_i = 1$, is defined as the minimizer

$$\bar{\mu} := \underset{\mu \in \mathcal{P}(\mathbb{R}^n)}{\arg\min} \sum_{i=1}^N \lambda_i W_2^2(\mu, \mu_i). \quad (10)$$

When the population consists of Gaussians $\mu_i = \mathcal{N}(m_i, K_i)$, one can show that the barycenter is Gaussian given by $\bar{\mu} = \mathcal{N}(\bar{m}, \bar{K})$, where $\bar{m}, \bar{K}$ satisfy [1, Thm. 6.1]

$$\bar{m} = \sum_{i=1}^{N} \lambda_i m_i, \quad \bar{K} = \sum_{i=1}^{N} \lambda_i \left(K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \right)^{\frac{1}{2}}. \tag{11}$$

2.4 Entropic relaxation

Let $\mu, \nu \in \mathcal{P}(X)$ with densities p_μ and p_ν. Then, we denote by

$$D_{\mathrm{KL}}(\mu || \nu) = -\mathbb{E}_\mu \left[\log \frac{p_\nu}{p_\mu} \right], \tag{12}$$

the *Kullback–Leibler divergence* (KL-divergence) between μ and ν. The *differential entropy* of μ is given by

$$H(\mu) = -\mathbb{E}_\mu [\log p_\mu]. \tag{13}$$

For a product measure, we have the identity

$$D_{\mathrm{KL}}(\gamma || \mu_0 \otimes \mu_1) = H(\mu_0) + H(\mu_1) - H(\gamma). \tag{14}$$

A special case that will be used later in this work is the KL-divergence between two non-degenerate multivariate Gaussian distributions $\mu_0 = \mathcal{N}(m_0, K_0)$ and $\mu_1 = \mathcal{N}(m_1, K_1)$ when $X = \mathbb{R}^n$, which is given by

$$\begin{aligned} D_{\mathrm{KL}}(\mu || \nu) =& \frac{1}{2} \left(\mathrm{Tr} \left(K_0^{-1} K_1 \right) + (m_1 - m_0)^T \, K_0^{-1} \, (m_1 - m_0) \right. \\ & \left. -n + \ln \left(\frac{\det K_1}{\det K_0} \right) \right), \end{aligned} \tag{15}$$

and for the entropy we have

$$H(\mu_0) = \frac{1}{2} \log \det \left(2\pi e K_0 \right). \tag{16}$$

Given $\epsilon > 0$, we relax (1) with a KL-divergence term between the transport plan and the independent joint distribution as, yielding the *entropic OT problem* [20]

$$\mathrm{OT}_c^\epsilon(\mu, \nu) = \min_{\gamma \in \mathrm{ADM}(\mu, \nu)} \left\{ \mathbb{E}_\gamma [c] + \epsilon D_{\mathrm{KL}}(\gamma || \mu \otimes \nu) \right\}, \tag{17}$$

which yields a strictly convex problem with respect to γ. Moreover, this problem is numerically more favorable to solve (1) compared, for instance, to the *Hungarian* and the *auction algorithm*, due to the Sinkhorn–Knopp algorithm. As shown, for instance in [12,19,25,40,72], the above problem has a unique minimizer given by

$$\gamma^\varepsilon = \alpha^\varepsilon(x) \beta^\varepsilon(y) k(x, y) \mu(x) \nu(y), \tag{18}$$

if and only if there exists functions α^ε and β^ε such that

$$\begin{aligned} \alpha^\varepsilon(x)\mathbb{E}_\nu\left[\beta^\varepsilon k(x,\cdot)\right] &= 1, \\ \beta^\varepsilon(y)\mathbb{E}_\mu\left[\alpha^\varepsilon k(\cdot,y)\right] &= 1, \end{aligned} \tag{19}$$

where $k(x,y) = \exp\left(-\frac{1}{\epsilon}c\right)$ denotes the *Gibbs kernel*. We call γ^ϵ an *entropic transport plan*. Moreover, when $\varepsilon \to 0$, γ^ε converges to γ_{opt}, a solution of the OT problem (1) [22,39,50]; while when $\varepsilon \to \infty$, γ^ε converges to the independent coupling $\gamma^\infty = \mu \otimes \nu$ [36,69]. The latter property shows in particular that, for large ε, the entropy-Regularized OT behaves like an inner product and not like a norm. In linear algebra, the polarization formula is the usual way of defining a norm from a inner product. That is the main idea of Sinkhorn divergence.

2.5 Sinkhorn divergence

The KL-divergence term in OT_c^ϵ acts as a bias, as discussed in [30]. This can be removed by defining the *p-Sinkhorn divergence* as

$$S_p^\epsilon(\mu,\nu) = \mathrm{OT}_{d^p}^\epsilon(\mu,\nu) - \frac{1}{2}(\mathrm{OT}_{d^p}^\epsilon(\mu,\mu) + \mathrm{OT}_{d^p}^\epsilon(\nu,\nu)). \tag{20}$$

As shown in [30] if, for example, $c = d^p$, $p \geq 1$ the Sinkhorn divergences metrizes the convergence in law in the space of probability measures.

2.6 Entropy-Kantorovich duality

In this subsection we summarize well-known results on the Entropy–Kantorich. For further details and proofs, we refer the reader to [25].

Given a probability measure μ, the class of Entropy-Kantorovich potentials is defined by the set of measurable functions φ on $\mathbb{R}^n$ satisfying

$$L_\varepsilon^{\exp}(\mathbb{R}^n,\mu) = \left\{\varphi : \mathbb{R}^n \to [-\infty,\infty[: \ 0 < \mathbb{E}_\mu\left[\exp\left(\frac{1}{\epsilon}\varphi\right)\right] < \infty\right\}. \tag{21}$$

Then, given $c = d^2$, where $d(x,y) = \|x-y\|$, $\varphi \in L_\varepsilon^{\exp}(\mathbb{R}^n,\mu_0)$ and $\psi \in L_\varepsilon^{\exp}(\mathbb{R}^n,\mu_1)$, the *entropic Kantorovich (dual) formulation* of $\mathrm{OT}_{d^2}^\epsilon(\mu,\nu)$ is given by [25,30,36,41,50],

$$\begin{aligned} \mathrm{OT}_{d^2}^\epsilon(\mu_0,\mu_1) = \sup_{\varphi,\psi} &\left\{\mathbb{E}_{\mu_0}[\varphi] + \mathbb{E}_{\mu_1}[\psi]\right. \\ &\left. -\varepsilon\left(\mathbb{E}_{\mu_0\otimes\mu_1}\left[\exp\left(\frac{(\varphi\oplus\psi)-d^2}{\varepsilon}\right)\right]-1\right)\right\}, \end{aligned} \tag{22}$$

where $(\varphi\oplus\psi)(x,y) = \varphi(x)+\psi(y)$, $\varphi \in L_\varepsilon^{\exp}(\mathbb{R}^n,\mu_0)$, and $\psi \in L_\varepsilon^{\exp}(\mathbb{R}^n,\mu_1)$.

Finally, the theorem below illustrate the relationship between the Entropy-Kantorovich potentials and the solution (19) of the Entropic regularized Optimal Transport problem (17), assuming the cost c is bounded.

Theorem 1 [25] *Let $\varepsilon > 0$ be a positive number, c is bounded cost, $\mu_0, \mu_1 \in \mathcal{P}(\mathbb{R}^n)$ be probability measures. Then, the supremum in* (22) *is attained for a unique couple $(\varphi^\epsilon, \psi^\epsilon)$ (up to the trivial transformation $(\varphi^\epsilon, \psi^\epsilon) \to (\varphi^\epsilon + \alpha, \psi^\epsilon - \alpha)$). Moreover, the following are equivalent:*

(a) (Maximizers) *φ^ϵ and ψ^ϵ are maximizing potentials for* (22).
(b) (Schrödinger system) *Let*

$$\gamma^\epsilon = \exp\left(\frac{1}{\epsilon}\left(\varphi^\epsilon \oplus \psi^\epsilon - d^2\right)\right)\mu_0 \otimes \mu_1, \tag{23}$$

then $\gamma^\epsilon \in$ ADM(μ_0, μ_1). *Furthermore, γ^ϵ is the (unique) minimizer of the problem* (17).

Elements of the pair $(\varphi^\epsilon, \psi^\epsilon)$ reaching a maximum in (22) are called *entropic Kantorovich potentials*. Finally, a relationship between α^ϵ, β^ϵ in (19), and the entropic Kantorovich potentials $\varphi^\epsilon, \psi^\epsilon$ above, is according to Theorem 1 given by

$$\varphi^\epsilon = \epsilon \log \alpha^\epsilon, \quad \psi^\epsilon = \epsilon \log \beta^\epsilon. \tag{24}$$

Using the dual formulation, we can show the following.

Proposition 1 *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ and c be a bounded cost. Then,* $\mathrm{OT}_c^\epsilon(\mu, \nu)$ *is strictly convex in both arguments.*

Proof Let $\mu_t = t\mu_0 + (1-t)\mu_1$ for $t \in (0, 1)$, and (φ_j, ψ_j) be the entropic Kantorovich potentials associated with $\mathrm{OT}_c^\epsilon(\mu_j, \nu)$ for $j = 0, 1$, and (φ, ψ) for $\mathrm{OT}_c^\epsilon(\mu_t, \nu)$. Then, using the dual formulation (22), we have

$$\begin{aligned}
\mathrm{OT}_c^\epsilon(\mu_t, \nu) &= t\left(\mathbb{E}_{\mu_0}[\varphi] + \mathbb{E}_\nu[\psi]\right) + (1-t)\left(\mathbb{E}_{\mu_1}[\varphi] + \mathbb{E}_\nu[\psi]\right)\\
&\quad - \epsilon t\left(\mathbb{E}_{\mu_0\otimes\nu}\left[\exp\left(\frac{(\varphi \otimes \psi) - c}{\epsilon}\right)\right] - 1\right)\\
&\quad - \epsilon(1-t)\left(\mathbb{E}_{\mu_1\otimes\nu}\left[\exp\left(\frac{(\varphi \otimes \psi) - c}{\epsilon}\right)\right] - 1\right)\\
&< t\left(\mathbb{E}_{\mu_0}[\varphi_0] + \mathbb{E}_\nu[\psi_0]\right) + (1-t)\left(\mathbb{E}_{\mu_1}[\varphi_1] + \mathbb{E}_\nu[\psi_1]\right) \qquad (25)\\
&\quad - \epsilon t\left(\mathbb{E}_{\mu_0\otimes\nu}\left[\exp\left(\frac{(\varphi_0 \otimes \psi_0) - c}{\epsilon}\right)\right] - 1\right)\\
&\quad - \epsilon(1-t)\left(\mathbb{E}_{\mu_1\otimes\nu}\left[\exp\left(\frac{(\varphi_1 \otimes \psi_1) - c}{\epsilon}\right)\right] - 1\right)\\
&= t\mathrm{OT}_c^\epsilon(\mu_0, \nu) + (1-t)\mathrm{OT}_c^\epsilon(\mu_1, \nu),
\end{aligned}$$

where the first equality results from linearity of expectations, and the inequality from noticing that the pair (φ, ψ) is a competitor for (φ_j, ψ_j), $j = 0, 1$, but due to uniqueness of the entropic Kantorovich potentials (up to scalar additives, Theorem 1), (φ, ψ) cannot be equal to (φ_0, ψ_0) and (φ_1, ψ_1) (unless $\mu_0 = \mu_1$), and will thus return lower values. □

2.7 Dynamical formulation of entropy relaxed optimal transport

Analogously to unregularized OT theory, the entropic-regularization of OT with distance cost admits a *dynamical* (aka *Benamou–Brenier*) formulation.

In the following, we again consider the particular case when the cost function is given by $c(x, y) = \|x - y\|^2$. Then, we can write (17) as [41,50]

$$\mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) = \min_{(\mu^{\epsilon}_t, v_t)} \int_0^1 \mathbb{E}_{\mu^{\epsilon}_t}\left[\|v_t\|^2\right] dt + H(\mu_0) + H(\mu_1), \tag{26}$$

where $t \in [0, 1]$, $\mu^{\epsilon}_0 = \mu_0$, $\mu^{\epsilon}_1 = \mu_1$, and

$$\partial_t \mu^{\epsilon}_t + \nabla \cdot (v_t \mu^{\epsilon}_t) = \frac{\varepsilon}{2} \Delta \mu^{\epsilon}_t. \tag{27}$$

where the minimum must be understood as taken among all couples (μ^{ϵ}_t, v_t) solving the continuity equation in the distributional sense (see appendix A); moreover, the minimum is attained if and only if $(\mu^{\epsilon}_t, v_t) = (\mu^{\epsilon}_t, \nabla\phi^{\varepsilon}_t)$, for a potential $\phi^{\varepsilon}_t : \mathbb{R}^d \to \mathbb{R}$, which is defined in the following via the entropic potentials. The resulting μ^{ϵ}_t is called the *entropic interpolation* between μ_0 and μ_1.

The solution can be characterized by (while abusing the notation and writing $\mu(x)$ for the density of μ, which will be done throughout this work)

$$\gamma^{\varepsilon}(x, y) = \alpha^{\varepsilon}(x)\beta^{\varepsilon}(y) \exp\left(-\frac{1}{\epsilon}\|x - y\|^2\right) \mu_0(x)\mu_1(y), \tag{28}$$

in (19) of the static problem (17) in conjunction with the heat flow allows us to compute the entropic interpolation from μ_0 to μ_1, which is given by [41,50,70]

$$\begin{aligned} \mu^{\epsilon}_t &= \mathcal{H}^{\mu_0}_{t\varepsilon}(\alpha^{\varepsilon})\, \mathcal{H}^{\mu_1}_{(1-t)\varepsilon}(\beta^{\varepsilon}), \\ \mathcal{H}^{\mu}_s[f] &= \int_{\mathbb{R}^n} \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{1}{s}\|x - z\|^2\right) f(z)\mu(z)\mathrm{d}z, \end{aligned} \tag{29}$$

and α^{ε},β^{ε} are the Entropy-Kantorovich potentials solving the system (19). In particular, we have that

$$\alpha^{\varepsilon}(x)\mathcal{H}^{\mu_1}_{\varepsilon}(\beta^{\varepsilon})(x) = 1, \quad \beta^{\varepsilon}(y)\, \mathcal{H}^{\mu_0}_{\varepsilon}(\alpha^{\varepsilon})(y) = 1. \tag{30}$$

In particular, when we send the regularization parameter $\varepsilon \to 0$, the curves of measures μ^{ϵ}_t converge to the 2-Wasserstein between μ_0 and μ_1 [40,50]. Moreover, we can also write the entropic interpolation μ_t^ϵ and the dynamic entropic Kantorovich potentials $(\varphi_t^\varepsilon, \psi_t^\varepsilon)$ via the relation $\varphi_t^\varepsilon + \psi_t^\varepsilon = \varepsilon \log \mu_t^\epsilon$.

Now, by defining $\phi_t^\varepsilon = (\varphi_t^\varepsilon - \psi_t^\varepsilon)/2$, it is easy to check that by imposing $v_t^\varepsilon = \nabla \phi_t^\varepsilon$ we have that $(\mu_t^\epsilon, v_t^\varepsilon)$ solves the Fokker–Planck equation

$$\partial_t \mu_t^\epsilon + \nabla \cdot (v_t^\varepsilon \mu_t^\epsilon) = \frac{\varepsilon}{2} \Delta \mu_t^\epsilon. \tag{31}$$

3 Entropy-regularized 2-Wasserstein distance between Gaussians

In this section we consider the special case of (17) and (20) when $c(x, y) = d^2(x, y) = |x - y|^2$ is the Euclidian distance in $\mathbb{R}^n$ and $\mu_0 \sim \mathcal{N}(m_0, K_0)$, $\nu \sim \mathcal{N}(m_1, K_1)$ are multivariate Gaussian distributions. We are interested in obtain explicity formulas for the optimal coupling γ^ε solving (17), the Entropy-Kantorovich maximizers $(\varphi^\epsilon, \psi^\epsilon)$ in (22) and the entropic displacement interpolation μ_t^ε in (29).

We start by showing that we can assume, without loss of generality, that μ_0 and μ_1 are centered Gaussian distributions. The general case is obtain just by a shift depending on the L^2-distance of the center of both Gaussians.

Proposition 2 *Let $c(x, y) = \|x - y\|^2$, $X_i \sim \mu_i \in \mathcal{P}(\mathbb{R}^n)$ for $i = 0, 1$ and $m_i = \mathbb{E}[\mu_i]$. Denote by $\hat{X}_i = X_i - m_i \sim \hat{\mu}_i$ the corresponding centered distributions. Then*

$$\mathrm{OT}_{d^2}^\epsilon(\mu_0, \mu_1) = \|m_0 - m_1\|^2 + \mathrm{OT}_{d^2}^\epsilon\left(\hat{\mu}_0, \hat{\mu}_1\right). \tag{32}$$

Proof Recall the definition given in (17)

$$\mathrm{OT}_c^\epsilon(\mu_0, \mu_1) = \min_{\gamma \in \mathrm{ADM}(\mu_0, \mu_1)} \left\{ \mathbb{E}_\gamma[c] + \epsilon D_{\mathrm{KL}}(\gamma || \mu_0 \otimes \mu_1) \right\}. \tag{33}$$

Then, as $c = d^2$, for the first term we can write

$$\begin{aligned} \mathbb{E}_\gamma\left[d^2\right] &= \int_{\mathbb{R}^n} \|x - y\|^2 d\gamma(x, y) \\ &= \int_{\mathbb{R}^n} \Big(\|(x - m_0) - (y - m_1)\|^2 + \|m_0 - m_1\|^2 \\ &\quad +2\left((x - m_0) - (y - m_1)\right)^T (m_0 - m_1) \Big) d\gamma(x, y) \\ &= \|m_0 - m_1\|^2 + \int_{\mathbb{R}^n} \|x - y\|^2 d\gamma(x + m_0, y + m_1). \end{aligned} \tag{34}$$

We now verify that the requirement $\gamma \in \mathrm{ADM}(\mu_0, \mu_1)$ is equivalent with $\gamma(\cdot + m_0, \cdot + m_1) \in \mathrm{ADM}(\hat{\mu}_0, \hat{\mu}_1)$, which results from

$$\int_{\mathbb{R}^n} \gamma(x + m_0, y + m_1) \mathrm{dy} = \mu_0(x + m_0) = \hat{\mu}_0(x), \tag{35}$$

and similarly for the other margin. Finally, for the entropy term, we use the identity (14). Now, as the entropy of a distribution does not depend on the expected value, we have $H(\mu_i) = H(\hat{\mu}_i)$, and therefore

$$D_{\mathrm{KL}}(\gamma || \hat{\mu}_0 \otimes \hat{\mu}_1) = H(\mu_0) + H(\mu_1) - H(\gamma). \tag{36}$$

Putting everything together, we get

$$\begin{aligned} \mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) &= \|m_0 - m_1\|^2 \\ &\quad + \min_{\gamma \in \mathrm{ADM}(\hat{\mu}_0, \hat{\mu}_1)} \left\{ \mathbb{E}_\gamma[d^2] + \epsilon D_{D_{\mathrm{KL}}}(\gamma || \hat{\mu}_0 \otimes \hat{\mu}_1) \right\}, \\ &= \|m_0 - m_1\|^2 + \mathrm{OT}^{\epsilon}_{d^2}(\hat{\mu}_0, \hat{\mu}_1). \end{aligned} \tag{37}$$

□

Proposition 3 *Let $\mu_i = \mathcal{N}(0, K_i) \in \mathcal{N}(\mathbb{R}^n)$ for $i = 0, 1$. Then, the unique optimal plan γ^ϵ in $\mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1)$ is a centered Gaussian distribution.*

Proof Note that $\mathbb{E}_\gamma[d^2]$ depends only on the mean and covariance of γ, and therefore remains constant, if γ is replaced with a Gaussian with the corresponding mean and covariance (which we can do, as the marginals are Gaussians). Then, for the other term, using the identity (14)

$$D_{\mathrm{KL}}(\gamma || \mu_0 \otimes \mu_1) = H(\mu) + H(\nu) - H(\gamma). \tag{38}$$

It is readily seen that the γ with a fixed covariance matrix minimizing this expression is Gaussian, as Gaussians achieve maximal entropy over distributions sharing a fixed covariance matrix. Therefore, we can deduce that γ^ϵ is Gaussian. Finally, as both of the marginals μ_0 and μ_1 are centered, so is γ^ϵ. □

We now arrive at the main theorem of this work, detailing the entropic 2-Wasserstein geometry between multivariate Gaussians. The proof is based on studying the Schrödinger system given in (19). We give an alternative proof for the statement **a.** in Theorem 2 in Appendix B, by finding the minimizer of the OT problem. Recall, that a noteworthy property of the entropic interpolant, is that even if we are interpolating from μ to itself, the trajectory does not constantly stay at μ.

Theorem 2 *Let $\mu_i = \mathcal{N}(0, K_i)$, for $i = 0, 1$, be two centered multivariate Gaussian distributions in $\mathbb{R}^n$, write $N^{\epsilon}_{ij} = \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K_j K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}$ and $M^\epsilon = I + \left(I + \frac{16}{\epsilon^2} K_0 K_1\right)^{\frac{1}{2}}$. Then,*

(a) *The density of the optimal entropy relaxed plan γ^ϵ is given by*

$$\gamma^\epsilon(x,y) = \alpha^\epsilon(x)\beta^\epsilon(y)\exp\left(-\frac{\|x-y\|^2}{\epsilon}\right)\mu_0(x)\mu_1(y), \tag{39}$$

where $\alpha^\epsilon(x) = \exp\left(x^T Ax + a\right)$, $\beta^\epsilon(y) = \exp\left(y^T By + b\right)$, and

$$\begin{aligned} A &= \frac{1}{4}K_0^{-\frac{1}{2}}\left(I + \frac{4}{\epsilon}K_0 - N_{01}^\epsilon\right)K_0^{-\frac{1}{2}} \\ B &= \frac{1}{4}K_1^{-\frac{1}{2}}\left(I + \frac{4}{\epsilon}K_1 - N_{10}^\epsilon\right)K_1^{-\frac{1}{2}} \\ \exp(a+b) &= \sqrt{\frac{1}{2^n}\det\left(M^\epsilon\right)}. \end{aligned} \tag{40}$$

(b) *The entropic optimal transport quantity is given by*

$$\begin{aligned} \mathrm{OT}_{d^2}^\epsilon(\mu_0,\mu_1) &= \mathrm{Tr}(K_0) + \mathrm{Tr}(K_1) \\ &\quad - \frac{\epsilon}{2}\left(\mathrm{Tr}(M^\epsilon) - \log\det(M^\epsilon) + n\log 2 - 2n\right) \end{aligned} \tag{41}$$

(c) *The entropic displacement interpolation μ_t^ϵ, $t \in [0,1]$, between μ_0 and μ_1 is given by $\mu_t^\epsilon = \mathcal{N}\left(0, K_t^\epsilon\right)$, where*

$$\begin{aligned} K_t^\epsilon &= \frac{(1-t)^2\epsilon^2}{16}K_1^{-\frac{1}{2}}\left(-I + \left(\frac{4t}{(1-t)\epsilon}K_1 + N_{10}^\epsilon\right)^2\right)K_1^{-\frac{1}{2}} \\ &= \frac{t^2\epsilon^2}{16}K_0^{-\frac{1}{2}}\left(-I + \left(\frac{4(1-t)}{t\epsilon}K_0 + N_{01}^\epsilon\right)^2\right)K_0^{-\frac{1}{2}} \\ &= (1-t)^2K_0 + t^2K_1 + t(1-t)\left[\left(\frac{\epsilon^2}{16}I + K_0K_1\right)^{1/2}\right. \\ &\quad \left. + \left(\frac{\epsilon^2}{16}I + K_1K_0\right)^{1/2}\right]. \end{aligned} \tag{42}$$

Proof **Part a.** Recall that α^ε, β^ε are the unique functions that give the density of the optimal plan γ^ϵ

$$\gamma^\epsilon(x,y) = \alpha^\varepsilon(x)\beta^\varepsilon(y)\exp\left(-\frac{\|x-y\|^2}{\epsilon}\right)\mu_0(x)\mu_1(y). \tag{43}$$

The optimal plan is required to have the right marginals (19), that is,

$$\begin{aligned}\mu_0(x) &= \int_{\mathbb{R}^n} \gamma^\epsilon(x, y)\mathrm{d}y \\ &= \alpha^\varepsilon(x) \int_{\mathbb{R}^n} \beta^\varepsilon(y) \exp\left(-\frac{\|x-y\|^2}{\epsilon}\right) \mu_0(x)\mu_1(y)\mathrm{d}y, \\ \mu_1(y) &= \int_{\mathbb{R}^n} \gamma^\epsilon(x, y)\mathrm{d}x \\ &= \beta^\varepsilon(y) \int_{\mathbb{R}^n} \alpha^\varepsilon(x) \exp\left(-\frac{\|x-y\|^2}{\epsilon}\right) \mu_0(x)\mu_1(y)\mathrm{d}x.\end{aligned} \tag{44}$$

Assuming $\alpha^\varepsilon(x) = \exp(x^T Ax + a)$ and $\beta^\varepsilon(y) = \exp(y^T By + b)$, substituting in μ_0 and μ_1, and after some simplifications, the system reads

$$\begin{aligned}1 &= \frac{\exp(a+b)}{\sqrt{\det(2\pi K_1)}} \exp\left(x^T\left(A - \frac{1}{\epsilon}I\right)x\right) \\ &\quad \times \int_X \exp\left(y^T\left(B - \frac{1}{\epsilon}I - \frac{1}{2}K_1^{-1}\right)y + \frac{2}{\epsilon}x^T y\right)\mathrm{d}y, \\ 1 &= \frac{\exp(a+b)}{\sqrt{\det(2\pi K_0)}} \exp\left(y^T\left(B - \frac{1}{\epsilon}I\right)y\right) \\ &\quad \times \int_Y \exp\left(x^T\left(A - \frac{1}{\epsilon}I - \frac{1}{2}K_0^{-1}\right)x + \frac{2}{\epsilon}y^T x\right)\mathrm{d}x.\end{aligned} \tag{45}$$

Using the identity

$$\int_X \exp\left(-x^T Cx + b^T x\right)\mathrm{d}x = \sqrt{\frac{\pi^n}{\det(C)}} \exp\left(\frac{1}{4}b^T C^{-1} b\right), \tag{46}$$

the system (45) results in

$$\begin{aligned}A &= \frac{1}{\epsilon}I + \frac{1}{\epsilon^2}\left(B - \frac{1}{\epsilon}I - \frac{1}{2}K_1^{-1}\right)^{-1}, \\ B &= \frac{1}{\epsilon}I + \frac{1}{\epsilon^2}\left(A - \frac{1}{\epsilon}I - \frac{1}{2}K_0^{-1}\right)^{-1}, \\ \exp(a+b) &= \sqrt{\det(2K_1)\det\left(\frac{1}{\epsilon}I + \frac{1}{2}K_1^{-1} - B\right)} \\ \exp(a+b) &= \sqrt{\det(2K_0)\det\left(\frac{1}{\epsilon}I + \frac{1}{2}K_0^{-1} - A\right)}\end{aligned} \tag{47}$$

Let us solve for A and B first. From system (47), we get that A and B can be written as

$$\begin{aligned} A &= \frac{1}{\epsilon} I + \frac{1}{\epsilon^2}\left(\frac{1}{\epsilon^2}\left(A - \frac{1}{\epsilon} I - \frac{1}{2} K_0^{-1}\right)^{-1} - \frac{1}{2} K_1^{-1}\right)^{-1}, \\ B &= \frac{1}{\epsilon} I + \frac{1}{\epsilon^2}\left(\frac{1}{\epsilon^2}\left(B - \frac{1}{\epsilon} I - \frac{1}{2} K_1^{-1}\right)^{-1} - \frac{1}{2} K_0^{-1}\right)^{-1}. \end{aligned} \tag{48}$$

Then, one can show, that the A, B given in (40) solves this system. Plugging A, B in the expressions for $\exp(a + b)$ in (47), we get

$$\exp(a + b) = \sqrt{\frac{1}{2^n} \det\left(I + \left(I + \frac{16}{\epsilon^2} K_0 K_1\right)^{\frac{1}{2}}\right)}, \tag{49}$$

for which a possible solution is given by

$$a = b = \frac{1}{4}\left(-n \log 2 + \log\det\left(I + \left(I + \frac{16}{\epsilon^2} K_0 K_1\right)^{\frac{1}{2}}\right)\right). \tag{50}$$

Now, we show that A solves the equation given in (48). Manipulating (48) we see that it suffices to show the equality

$$\left(A - \frac{1}{\epsilon} I\right)^{-1} = \left(A - \frac{1}{\epsilon} I - \frac{1}{2} K_0^{-1}\right)^{-1} - \frac{1}{2} K_1^{-1}. \tag{51}$$

Substituting in A given in (40), the left-hand side reads

$$\left(A - \frac{1}{\epsilon} I\right)^{-1} = 4 K_0^{\frac{1}{2}}\left(I - \left(I + \frac{16}{\epsilon^2} K_0^{\frac{1}{2}} K_1 K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} K_0^{\frac{1}{2}}, \tag{52}$$

whereas the right-hand side is given by

$$\begin{aligned} &\left(A - \frac{1}{\epsilon} I - \frac{1}{2} K_0^{-1}\right)^{-1} - \frac{1}{2} K_1^{-1} \\ &= -4 K_0^{\frac{1}{2}}\left(\frac{\epsilon^2}{8}\left(K_0^{\frac{1}{2}} K_1 K_0^{\frac{1}{2}}\right)^{-1} + \left(I + \left(I + \frac{16}{\epsilon^2} K_0^{\frac{1}{2}} K_1 K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1}\right) K_0^{\frac{1}{2}}. \end{aligned} \tag{53}$$

Therefore, we need to show the equality

$$\begin{aligned}\left(I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} &= \left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1}, \\ &\quad -\left(\frac{8}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{-1}\end{aligned} \tag{54}$$

which can be derived as follows

$$\begin{aligned}&-\left(\frac{8}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{-1}+\left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} \\ &\quad = -2\left(I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1}+\left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} \\ &\qquad \times\left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} \\ &\quad = \left(I-2\left(I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1}\right)\left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} \\ &\quad = \left(I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}-2I\right)\left(I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} \\ &\qquad \times\left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} \\ &\quad = \left(I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1},\end{aligned} \tag{55}$$

where the first step results from writing

$$-\left(\frac{8}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{-1} = -2\left(-I+\left(I+\frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)\right)^{-1}, \tag{56}$$

and using $M-I=(M^{\frac{1}{2}}+1)(M^{\frac{1}{2}}-1)$ on the right-hand side.

Part b. Let $\varphi_\epsilon(x)=\epsilon\log\alpha_\epsilon(x)$ and $\psi_\epsilon(y)=\epsilon\log\beta_\epsilon(y)$, and as previously,

$$M^\epsilon = I+\left(I+\frac{16}{\epsilon^2}K_0K_1\right)^{\frac{1}{2}}, \tag{57}$$

then plugging φ^ϵ and ψ^ϵ into (22) yields

$$\begin{aligned}
\mathrm{OT}^\epsilon_{d^2}(\mu_0, \mu_1) &= \mathbb{E}_{\mu_0}[\varphi_\epsilon] + \mathbb{E}_{\mu_1}[\psi_\epsilon] \\
&\quad - \epsilon\left(\mathbb{E}_{\mu_0\otimes\mu_1}\left[\exp\left(\frac{1}{\epsilon}\left((\varphi\oplus\psi) - d^2\right)\right)\right] - 1\right) \\
&= \epsilon\left(\mathbb{E}_{\mu_0}[\log\alpha_\epsilon] + \mathbb{E}_{\mu_1}[\log\beta_\epsilon]\right) \\
&\quad - \epsilon\left(\mathbb{E}_{\mu_0\otimes\mu_1}\left[\alpha^\epsilon\beta^\epsilon\exp\left(-\frac{1}{\epsilon}d^2\right)\right] - 1\right) \\
&= \epsilon\left(\mathbb{E}_{X\sim\mu_0}\left[X^T A X + a\right] + \mathbb{E}_{Y\sim\mu_1}\left[Y^T B Y + b\right]\right) \\
&= \epsilon\left(\mathrm{Tr}\left[K_0 A\right] + \mathrm{Tr}\left[K_1 B\right] + a + b\right) \qquad (58) \\
&= \frac{\epsilon}{4}\mathrm{Tr}\left[I + \frac{4}{\epsilon}K_0 - \left(I + \frac{16}{\epsilon^2}K_0^{\frac{1}{2}}K_1K_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \\
&\quad + \frac{\epsilon}{4}\mathrm{Tr}\left[I + \frac{4}{\epsilon}K_1 - \left(I + \frac{16}{\epsilon^2}K_1^{\frac{1}{2}}K_0K_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \\
&\quad + \epsilon(a+b) \\
&= \mathrm{Tr}K_0 + \mathrm{Tr}K_1 - \frac{\epsilon}{2}\left(\mathrm{Tr}M^\epsilon - \log\det M^\epsilon + n\log 2 - 2n\right),
\end{aligned}$$

where we used the fact that $C^{\frac{1}{2}}DC^{\frac{1}{2}}$ has same eigenvalues as CD, and so $\mathrm{Tr}\left[(I + C^{\frac{1}{2}}DC^{\frac{1}{2}})^{\frac{1}{2}}\right] = \mathrm{Tr}\left[(I + CD)^{\frac{1}{2}}\right]$ for any square and positive-definite matrices C and D.

Part c. As we have solved for α^ϵ and β^ϵ for the optimal plan, the entropic interpolant μ_t^ϵ between μ_0 and μ_1 is given by (29), which we rewrite here

$$\mu_t^\epsilon(x) = \left(\mathcal{H}^{\mu_0}_{t\epsilon}[\alpha^\epsilon](x)\right)\left(\mathcal{H}^{\mu_1}_{(1-t)\epsilon}[\beta^\epsilon](x)\right). \qquad (59)$$

Then, we can compute

$$\begin{aligned}
\mathcal{H}^{\mu_0}_{t\epsilon}[\alpha^\epsilon](x) &= \frac{1}{\sqrt{\det((2\pi)^2 t\epsilon K_0)}}\int_{\mathbb{R}^n}\exp\left(z^T Az + a - \frac{1}{t\epsilon}\|x - z\|^2 - \frac{1}{2}z^T K_0^{-1}z\right)\mathrm{d}z \\
&= \frac{\exp\left(a - \frac{1}{t\epsilon}x^T x\right)}{\sqrt{\det(2\pi t\epsilon K_0)}}\int_{\mathbb{R}^n}\exp\left(z^T\left(A - \frac{1}{t\epsilon}I - \frac{1}{2}K_0^{-1}\right)z + \frac{2}{t\epsilon}x^T z\right)\mathrm{d}z \\
&= \frac{\exp(a)}{\sqrt{\det\left(2\pi t\epsilon K_0\right)\det\left(\frac{1}{t\epsilon}I + \frac{1}{2}K_0^{-1} - A\right)}} \qquad (60) \\
&\quad \times\exp\left(\frac{1}{t^2\epsilon^2}x^T\left(\left(\frac{1}{t\epsilon}I + \frac{1}{2}K_0^{-1} - A\right)^{-1} - I\right)x\right),
\end{aligned}$$

similar computation yields

$$\begin{aligned}&\mathcal{H}^{\mu_1}_{(1-t)\epsilon}[\beta^\epsilon](x)\\&\quad=\frac{\exp(b)}{\sqrt{\det\left(2\pi(1-t)\epsilon K_1\right)\det\left(\frac{1}{(1-t)\epsilon}I+\frac{1}{2}K_1^{-1}-B\right)}}\\&\qquad\times\exp\left(\frac{1}{(1-t)^2\epsilon^2}x^T\left(\left(\frac{1}{(1-t)\epsilon}I+\frac{1}{2}K_1^{-1}-B\right)^{-1}-I\right)x\right).\end{aligned}\tag{61}$$

Putting these together, we get

$$\begin{aligned}\mu_t^\epsilon(x)&=\left(\mathcal{H}^{\mu_0}_{t\epsilon}[\alpha^\epsilon](x)\right)\left(\mathcal{H}^{\mu_1}_{(1-t)\epsilon}[\beta^\epsilon](x)\right)\\&=N\exp\left(x^T\left[\frac{1}{t^2\epsilon^2}\left(\left(\frac{1}{t\epsilon}I+\frac{1}{2}K_0^{-1}-A\right)^{-1}-I\right)\right.\right.\\&\qquad\left.\left.+\frac{1}{(1-t)^2\epsilon^2}\left(\left(\frac{1}{(1-t)\epsilon}I+\frac{1}{2}K_1^{-1}-B\right)^{-1}-I\right)\right]x\right)\\&:=N\exp\left(x^T\left(T_0(A)+T_1(B)\right)x\right),\end{aligned}\tag{62}$$

where N is a normalizing constant. We can simplify the matrix $T_0(A)+T_1(B)$ in (62). Write

$$N_{10}^\epsilon=\left(I+\frac{16}{\epsilon^2}K_1^{\frac{1}{2}}K_0K_1^{\frac{1}{2}}\right)^{\frac{1}{2}},\tag{63}$$

and consider the first term

$$\begin{aligned}T_0(A)&=\frac{1}{t^2\epsilon^2}\left(\left(\frac{1}{t\epsilon}I+\frac{1}{2}K_0^{-1}-A\right)^{-1}-I\right)\\&=\frac{1}{t^2\epsilon^2}\left(\left(\frac{(1-t)}{t\epsilon}I+\frac{1}{2}K_0^{-1}-\frac{1}{\epsilon^2}\left(B-\frac{1}{\epsilon}I-\frac{1}{2}K_1^{-1}\right)^{-1}\right)^{-1}-t\epsilon I\right)\\&=\frac{1}{t^2\epsilon}\left(\left(\frac{(1-t)}{t}I-(I-\epsilon B)^{-1}\right)^{-1}-t\epsilon I\right)\\&=\frac{1}{t^2\epsilon}\left(\left(\frac{t}{(1-t)}I-\frac{t^2}{(1-t)^2}\left(\frac{t}{(1-t)}I+(I-\epsilon B)\right)^{-1}\right)-t\epsilon I\right)\\&=\frac{1}{(1-t)^2\epsilon^2}\left(I-\left(\frac{1}{(1-t)\epsilon}I-B\right)^{-1}\right)\\&=\frac{4}{(1-t)^2\epsilon^2}K_1^{\frac{1}{2}}\left(-I+\frac{4t}{\epsilon(1-t)}K_1+N_{10}^\epsilon\right)^{-1}K_1^{\frac{1}{2}},\end{aligned}\tag{64}$$

where second equality follows from (47), third from (48), fourth from the Woodbury matrix inverse identity

$$(C+D)^{-1} = C^{-1} - C^{-1}\left(C^{-1} + D^{-1}\right)^{-1} C^{-1}, \tag{65}$$

and the last one from substituting in B given in (40).

Likewise, we can substitute B in the second term $T_1(B)$, which yields

$$\begin{aligned} T_1(B) &= \frac{1}{(1-t)^2\epsilon^2}\left(\left(\frac{1}{(1-t)\epsilon}I + \frac{1}{2}K_1^{-1} - B\right)^{-1} - I\right) \\ &= \frac{4}{(1-t)^2\epsilon^2}K_1^{\frac{1}{2}}\left(I + \frac{4t}{\epsilon(1-t)}K_1 + N_{10}^{\epsilon}\right)^{-1} K_1^{\frac{1}{2}}. \end{aligned} \tag{66}$$

Putting the two terms together, we get

$$\begin{aligned} T_0(A) + T_1(B) &= \frac{4}{(1-t)^2\epsilon^2}K_1^{\frac{1}{2}}\left(\left(-I + \frac{4t}{\epsilon(1-t)}K_1 + N_{10}^{\epsilon}\right)^{-1}\right. \\ &\quad \left. + \left(I + \frac{4t}{\epsilon(1-t)}K_1 + N_{10}^{\epsilon}\right)^{-1}\right) K_1^{\frac{1}{2}} \\ &= \frac{8}{(1-t)^2\epsilon^2}K_1^{\frac{1}{2}}\left(I - \left(\frac{4t}{(1-t)\epsilon}K_1 + N_{10}^{\epsilon}\right)^2\right)^{-1} K_1^{\frac{1}{2}}. \end{aligned} \tag{67}$$

Note, that we can write (62) as a Gaussian with covariance matrix K_t

$$\begin{aligned} \mu_t^{\epsilon}(x) &= N\exp\left(x^T\left(T_0(A) + T_1(B)\right)x\right) \\ &= N\exp\left(-\frac{1}{2}x^T\left(K_t^{\epsilon}\right)^{-1}x,\right) \end{aligned} \tag{68}$$

and so

$$\begin{aligned} K_t^{\epsilon} &= -\frac{1}{2}\left(T_0(A) + T_1(B)\right)^{-1} \\ &= \frac{(1-t)^2\epsilon^2}{16}K_1^{-\frac{1}{2}}\left(-I + \left(\frac{4t}{(1-t)\epsilon}K_1 + N_{10}^{\epsilon}\right)^2\right)K_1^{-\frac{1}{2}} \\ &= (1-t)^2K_0 + t^2K_1 + t(1-t)\left[\left(\frac{\epsilon^2}{16}I + K_0K_1\right)^{1/2} + \left(\frac{\epsilon^2}{16}I + K_1K_0\right)^{1/2}\right]. \end{aligned} \tag{69}$$

Where for the last step we use the formula

$$\left(I+\frac{16}{\epsilon^2}K_0K_1\right)^{1/2}=K_0^{1/2}\left(I+\frac{16}{\epsilon^2}K_0^{1/2}K_1K_0^{1/2}\right)^{1/2}K_0^{-1/2}. \tag{70}$$

□

Above we only considered centered Gaussians. Now we combine the results obtained in Proposition 2 and Theorem 2 to deduce the general case. As a consequence, we also derive the corresponding formulas for the Sinkhorn divergence between two Gaussians

Corollary 1 *Let $\mu_i=\mathcal{N}(m_i,K_i)$, for $i=0,1$, be two multivariate Gaussian distributions in $\mathbb{R}^n$. Then,*

(a)
$$\begin{aligned}\mathrm{OT}^{\epsilon}_{d^2}(\mu_0,\mu_1)&=\|m_0-m_1\|^2+\mathrm{Tr}(K_0)+\mathrm{Tr}(K_1)\\&\quad-\frac{\epsilon}{2}\left(\mathrm{Tr}(M^{\epsilon})-\log\det(M^{\epsilon})+n\log 2-2n\right)\end{aligned}\tag{71}$$

(b) *The entropic interpolant between μ_0 and μ_1 is $\mu_t^{\epsilon}=\mathcal{N}\left(m_t,K_t\right)$, $t\in[0,1]$, where $m_t=(t-1)m_0-tm_1$, and K_t is given in (42).*

(c) *Write $M_{ij}^{\epsilon}=I+\left(I+\frac{16}{\epsilon^2}K_iK_j\right)^{\frac{1}{2}}$, then*

$$\begin{aligned}S_2^{\epsilon}(\mu_0,\mu_1)&=\|m_0-m_1\|_2^2+\frac{\epsilon}{4}\left(\mathrm{Tr}\left(M_{00}^{\epsilon}-2M_{01}^{\epsilon}+M_{11}^{\epsilon}\right)\right.\\&\quad\left.+\log\left(\frac{\det^2(M_{01}^{\epsilon})}{\det(M_{00}^{\epsilon})\det(M_{11}^{\epsilon})}\right)\right).\end{aligned}\tag{72}$$

We will now emphasize an identity that can be derived from the calculations of Theorem 2, which we find useful.

Lemma 1 *Let C, D be symmetric positive-definite matrices. Then,*

$$\begin{aligned}&\frac{4}{\epsilon}D^{\frac{1}{2}}\left(I+\left(I+\frac{16}{\epsilon^2}D^{\frac{1}{2}}CD^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1}D^{\frac{1}{2}}\\&=I-\frac{\epsilon}{4}C^{-\frac{1}{2}}\left(I+\frac{4}{\epsilon}C-\left(I+\frac{16}{\epsilon^2}C^{\frac{1}{2}}DC^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)C^{-\frac{1}{2}}.\end{aligned}\tag{73}$$

Proof Similarly to (40), let

$$\begin{aligned}A&=\frac{1}{4}C^{-\frac{1}{2}}\left(I+\frac{4}{\epsilon}C-\left(I+\frac{16}{\epsilon^2}C^{\frac{1}{2}}DC^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)C^{-\frac{1}{2}}\\B&=\frac{1}{4}D^{-\frac{1}{2}}\left(I+\frac{4}{\epsilon}D-\left(I+\frac{16}{\epsilon^2}D^{\frac{1}{2}}CD^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)D^{-\frac{1}{2}}.\end{aligned}\tag{74}$$

Then, substituting B into the first equation of (47) (while remembering to replace $K_0 \mapsfrom C$, $K_1 \mapsfrom D$) results in

$$\begin{aligned} A &= \frac{1}{\epsilon} I + \frac{1}{\epsilon^2}\left(B - \frac{1}{\epsilon} I - \frac{1}{2} D^{-1}\right)^{-1} \\ &= \frac{1}{\epsilon} I + \frac{1}{\epsilon^2}\left(\frac{1}{4} D^{-\frac{1}{2}}\left(I + \frac{4}{\epsilon} D - \left(I + \frac{16}{\epsilon^2} D^{\frac{1}{2}} C D^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) D^{-\frac{1}{2}} \right. \\ &\quad \left. - \frac{1}{\epsilon} I - \frac{1}{2} D^{-1}\right)^{-1} \\ &= \frac{1}{\epsilon} I - \frac{4}{\epsilon^2} D^{\frac{1}{2}}\left(I + \left(I + \frac{16}{\epsilon^2} D^{\frac{1}{2}} C D^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} D^{\frac{1}{2}}, \end{aligned} \tag{75}$$

and so the result follows from substituting in A, multiplying both sides by $-\epsilon$, and moving $-I$ from right-hand side to left-hand side. □

Next, we study the limiting cases of ϵ going to 0 and ∞, reconfirming that the Sinkhorn divergence interpolates between 2-Wasserstein and MMD [30,36,69].

Proposition 4 *Let $\mu_i = \mathcal{N}(m_i, K_i)$, for $i = 0, 1$, be two multivariate Gaussian distributions in $\mathbb{R}^n$. Then,*

(a)

$$\begin{aligned} \lim_{\epsilon \to 0} \mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) &= W_2^2(\mu_0, \mu_1) \\ \lim_{\epsilon \to \infty} \mathrm{OT}^{\epsilon}_{d^2}(\mu_0, \mu_1) &= \|m_0 - m_1\|^2 + \mathrm{Tr}(K_0) + \mathrm{Tr}(K_1) \end{aligned} \tag{76}$$

(b)

$$\begin{aligned} \lim_{\epsilon \to 0} S_2^{\epsilon}(\mu_0, \mu_1) &= W_2^2(\mu_0, \mu_1) \\ \lim_{\epsilon \to \infty} S_2^{\epsilon}(\mu_0, \mu_1) &= \|m_0 - m_1\|^2 \end{aligned} \tag{77}$$

(c) *For $t \in [0, 1]$, denote by μ_t the 2-Wasserstein geodesic given in (9), and by μ_t^{ϵ} the entropic 2-Wasserstein interpolant between μ_0 and μ_1 given in (42). Then,*

$$\lim_{\epsilon \to 0} \mu_t^{\epsilon} = \mu_t. \tag{78}$$

Proof **Part a.** The $\epsilon \to 0$ case is a straight-forward computation

$$\begin{aligned}\mathrm{OT}^{\epsilon}_{d^2}(\mu_0,\mu_1) &= \|m_0 - m_1\|^2 + \mathrm{Tr}(K_0) + \mathrm{Tr}(K_1)\\ &\quad - \frac{\epsilon}{2}\big(\mathrm{Tr}(M^{\epsilon}) - \log\det(M^{\epsilon}) + n\log 2 - 2n\big)\\ &= \|m_0 - m_1\|^2 + \mathrm{Tr}(K_0) + \mathrm{Tr}(K_1)\\ &\quad - 2\mathrm{Tr}\left(\frac{\epsilon}{4}I + \left(\frac{\epsilon^2}{16}I + K_0K_1\right)^{\frac{1}{2}}\right)\\ &\quad + \frac{\epsilon}{2}\log\left(\det\left(\frac{\epsilon}{4}I + \left(\frac{\epsilon^2}{16}I + K_0K_1\right)^{\frac{1}{2}}\right)\right)\\ &\quad + \frac{\epsilon n}{2}(\log 2 - \log\epsilon + 2).\end{aligned} \tag{79}$$

Therefore, since $\epsilon\log\epsilon \to 0$ when $\epsilon \to 0$,

$$\begin{aligned}\lim_{\varepsilon\to 0}\mathrm{OT}^{\epsilon}_{d^2}(\mu_0,\mu_1) &= \|m_0 - m_1\|^2 + \mathrm{Tr}(K_0) + \mathrm{Tr}(K_1) - 2\mathrm{Tr}\,(K_0K_1)^{\frac{1}{2}}\\ &= W_2^2(\mu_0,\mu_1).\end{aligned} \tag{80}$$

We now compute the limit when $\varepsilon \to \infty$. It is enough to show that the term

$$\frac{\epsilon}{2}\left(\mathrm{Tr}(M^{\epsilon}) - \log\det\left(M^{\epsilon}\right) + n\log 2 - 2n\right), \tag{81}$$

goes to 0 when $\varepsilon \to \infty$. In fact, denote by $\{\lambda_i\}_{i=1}^{n}$ the eigenvalues of K_1K_2. Then,

$$\begin{aligned}&\frac{\epsilon}{2}\left(\mathrm{Tr}(M^{\epsilon}) - \log\det\left(M^{\epsilon}\right) + n\log 2 - 2n\right)\\ &\quad= \frac{\epsilon}{2}\sum_{i=1}^{n}\left(-1 + \left(1 + \frac{16}{\epsilon^2}\lambda_i\right)^{\frac{1}{2}} - \log\left(\frac{1}{2}\left(1 + \left(1 + \frac{16}{\epsilon^2}\lambda_i\right)^{\frac{1}{2}}\right)\right)\right).\end{aligned} \tag{82}$$

So, first notice that for any $\lambda > 0$,

$$\epsilon\left(-1 + \left(1 + \frac{16}{\epsilon^2}\lambda\right)^{\frac{1}{2}}\right) = \frac{16\lambda}{\epsilon + (\epsilon + 16\lambda)^{\frac{1}{2}}} \overset{\epsilon\to\infty}{=} 0. \tag{83}$$

Second, we have

$$\begin{aligned}
&\lim_{\epsilon\to\infty} \epsilon \log\left(\frac{1}{2}\left(1+\left(1+\frac{16}{\epsilon^2}\lambda\right)^{\frac{1}{2}}\right)\right)\\
&\overset{\text{L}'\text{Hospital}}{=} \lim_{\epsilon\to\infty} \frac{16\lambda}{\epsilon^3\left(1+\left(1+\frac{16}{\epsilon^2}\lambda\right)^{\frac{1}{2}}\right)\left(1+\frac{16}{\epsilon^2}\lambda\right)^{\frac{1}{2}}\log^2\left(\frac{1}{2}\left(1+\left(1+\frac{16}{\epsilon^2}\lambda\right)^{\frac{1}{2}}\right)\right)}\\
&= 0,
\end{aligned} \tag{84}$$

and so the result follows.

Part b. Straight-forward application of the above result to (72).

Part c. By a straight-forward computation on (42),

$$\begin{aligned}
K_t^\epsilon &= (1-t)^2K_0 + t^2K_1 + t(1-t)\left[\left(\frac{\epsilon^2}{16}I + K_0K_1\right)^{1/2}\right.\\
&\quad\left. + \left(\frac{\epsilon^2}{16}I + K_1K_0\right)^{1/2}\right] \\
&\overset{\epsilon\to 0}{=} (1-t)^2K_0 + t^2K_1 + t(1-t)[(K_0K_1)^{1/2} + (K_1K_0)^{1/2}]\\
&= K_t.
\end{aligned} \tag{85}$$

□

4 Entropic and Sinkhorn barycenters

In this section, we compute barycenters under the entropic regularization of the 2-Wasserstein distance (e.g. [10,11,13,21,25,47,51]) and the 2-Sinkhorn divergence of a population of multivariate Gaussians, restricted to the manifold of Gaussians.

4.1 Entropic 2-Wasserstein barycenter

Given N probability measures $\mu_i \in \mathcal{P}(\mathbb{R}^n)$, $i = 1, 2, \ldots, N$, the entropic barycenter $\bar{\mu}$ with weights $\lambda_i \geq 0$ is defined in the vein of *Karcher* and *Fréchet means*, given as

$$\bar{\mu} := \underset{\mu\in\mathcal{P}(\mathbb{R}^n)}{\arg\min} \sum_{i=1}^N \lambda_i \mathrm{OT}_{d^2}^\epsilon(\mu, \mu_i), \quad \sum_{i=1}^N \lambda_i = 1. \tag{86}$$

Then, (86) is strictly convex, as $\mathrm{OT}_c^\epsilon(\mu, \nu)$ is strictly convex in both μ and ν as stated by Prop. 1.

Next, let us focus on the Gaussian case. We lack the proof that such a barycenter will indeed be a Gaussian, so do note, that the following statement requires the restriction to Gaussians for the candidate barycenters.

Theorem 3 (Entropic Barycenter of Gaussians) *Let $\mu_i = \mathcal{N}(m_i, K_i)$, $i = 1, 2, \dots, N$ be a population of multivariate Gaussians. Then, their entropic barycenter* (86) *with weights $\lambda_i \geq 0$ such that $\sum_{i=1}^N \lambda_i = 1$, restricted to the manifold of Gaussians $\mathcal{N}(\mathbb{R}^n)$, is given by $\bar\mu = \mathcal{N}(\bar m, \bar K)$, where*

$$\bar m = \sum_{i=1}^N \lambda_i m_i, \quad \bar K = \frac{\epsilon}{4} \sum_{i=1}^N \lambda_i \left(-I + \left(I + \frac{16}{\epsilon^2} \bar K^{\frac{1}{2}} K_i \bar K^{\frac{1}{2}} \right)^{\frac{1}{2}} \right). \tag{87}$$

Proof Proposition 2 allows us to split the geometry into the L^2-geometry between the means and the entropic 2-Wasserstein geometry between the centered Gaussians (or their covariances). Then, it immediately follows that

$$\bar m = \sum_{i=1}^N \lambda_i m_i. \tag{88}$$

Therefore, we restrict our analysis to the case of centered distributions. Remark again, that in general, the minimizer of (86) might not be Gaussian, even when the population consists of Gaussians. However, here we will look for the barycenter on the manifold of Gaussian measures.

We begin with a straight-forward computation of the gradient of the objective given in (86)

$$\begin{aligned}
&\nabla_K \sum_{i=1}^N \lambda_i \mathrm{OT}^{\epsilon}_{d^2} \left(\mathcal{N}(0, K), \mathcal{N}(0, K_i)\right) \\
&= \nabla_K \sum_{i=1}^N \lambda_i \left(\mathrm{Tr} K + \mathrm{Tr} K_i - \frac{\epsilon}{2} \mathrm{Tr} \left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \right. \\
&\quad + \frac{\epsilon}{2} \log\det \left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \\
&\quad \left. - \frac{\epsilon}{2} \left(n \log 2 - 2n \right) \right), \\
&= \sum_{i=1}^N \lambda_i \left(\nabla_K \mathrm{Tr} K - \frac{\epsilon}{2} \nabla_K \mathrm{Tr} \left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \right. \\
&\quad \left. + \frac{\epsilon}{2} \nabla_K \log\det \left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \right).
\end{aligned} \tag{89}$$

where we used the closed-form solution obtained in the part **b.** of Theorem 2. Now, recall that $\nabla_K \mathrm{Tr} K = I$. For the second term, it holds

$$\begin{aligned} &\nabla_K \mathrm{Tr}\left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \\ &\quad = \frac{8}{\epsilon^2} K_i^{\frac{1}{2}} \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{-\frac{1}{2}} K_i^{\frac{1}{2}}. \end{aligned} \tag{90}$$

Finally, for the third term, we have

$$\begin{aligned} &\nabla_K \log\det\left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \\ &= \nabla_K \mathrm{Tr}\left(\mathrm{Log}\left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)\right) \\ &= \frac{8}{\epsilon^2} K_i^{\frac{1}{2}} \left(\left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right) + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} K_i^{\frac{1}{2}}, \end{aligned} \tag{91}$$

where $\mathrm{Log}(M)$ denotes the matrix square-root, and we use the results

$$\log\det(M) = \mathrm{Tr}\,(\mathrm{Log}(M))\,, \quad \nabla_M \mathrm{Tr} f(M) = f'(M), \tag{92}$$

when f is a matrix function given by a Taylor series, such as the matrix square-root or the matrix logarithm.

Using the Woodbury matrix identity (65), one gets

$$(I + A)^{-1} = -A^{-1} + (A^2 + A)^{-1}, \tag{93}$$

for an invertible A. Substituting (90) and (91) in (89), and using (93) with $A = \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}$, we get

$$\begin{aligned} &\nabla_K \sum_{i=1}^{N} \lambda_i \mathrm{OT}_{d^2}^{\epsilon}\left(\mathcal{N}(0, K), \mathcal{N}(0, K_i)\right) \\ &\quad = \sum_{i=1}^{N} \lambda_i \left(I - \frac{4}{\epsilon} K_i^{\frac{1}{2}} \left(I + \left(I + \frac{16}{\epsilon^2} K_i^{\frac{1}{2}} K K_i^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^{-1} K_i^{\frac{1}{2}}\right) \\ &\quad = \frac{\epsilon}{4} \sum_{i=1}^{N} \lambda_i K^{-\frac{1}{2}} \left(I + \frac{4}{\epsilon} K - \left(I + \frac{16}{\epsilon^2} K^{\frac{1}{2}} K_i K^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) K^{-\frac{1}{2}}. \end{aligned} \tag{94}$$

The last equality follows from Lemma 1 with the substitutions $C \mapsfrom K$ and $D \mapsfrom K_i$. Finally, setting (94) to zero, we get that the optimal $\bar{K}$ satisfies the expression given in (87). □

4.2 Sinkhorn barycenter

Now, we compute the barycenter of a population of Gaussians under the Sinkhorn divergence, defined by

$$\bar{\mu} := \underset{\mu \in \mathcal{P}(\mathbb{R}^n)}{\arg\min} \sum_{i=1}^{N} \lambda_i S_2^{\epsilon}(\mu, \mu_i), \quad \lambda_i \geq 0 \text{ and } \sum_{i=1}^{N} \lambda_i = 1. \tag{95}$$

Note that as $S_{\epsilon}^2(\mu, \nu)$ is convex in both μ and ν [30, Thm. 1], and so (95) is convex in μ. Now, similarly to the entropic barycenter case, we look for the barycenter of a population of Gaussians in the space of Gaussians $\mathcal{N}(\mathbb{R}^n)$.

Theorem 4 (Sinkhorn Barycenter of Gaussians) *Let $\mu_i = \mathcal{N}(m_i, K_i)$, $i = 1, 2, \ldots, N$ be a population of multivariate Gaussians. Then, their Sinkhorn barycenter (95) with weights $\lambda_i \geq 0$ such that $\sum_{i=1}^{N} \lambda_i = 1$, restricted to the manifold of Gaussians $\mathcal{N}(\mathbb{R}^n)$, is given by $\bar{\mu} = \mathcal{N}(\bar{m}, \bar{K})$, where*

$$\bar{m} = \sum_{i=1}^{N} \lambda_i m_i, \quad \bar{K} = \frac{\epsilon}{4}\left(-I + \left(\sum_{i=1}^{N} \lambda_i \left(I + \frac{16}{\epsilon^2} \bar{K}^{\frac{1}{2}} K_i \bar{K}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)^2\right)^{\frac{1}{2}}. \tag{96}$$

Proof As in the entropic 2-Wasserstein case, we take $\mu = \mathcal{N}(0, K)$ to be of Gaussian form. Then, we can compute the gradient

$$\begin{aligned} &\nabla_K \sum_{i=1}^{N} \lambda_i S_2^{\epsilon}\left(\mathcal{N}(0, K), \mathcal{N}(0, K_i)\right) \\ &\quad = \nabla_K \sum_{i=1}^{N} \lambda_i \Big(\mathrm{OT}_{d^2}^{\epsilon}\left(\mathcal{N}(0, K), \mathcal{N}(0, K_i)\right) \\ &\qquad - \frac{1}{2}\mathrm{OT}_{d^2}^{\epsilon}\left(\mathcal{N}(0, K), \mathcal{N}(0, K)\right) \\ &\qquad - \frac{1}{2}\mathrm{OT}_{d^2}^{\epsilon}\left(\mathcal{N}(0, K_i), \mathcal{N}(0, K_i)\right) \Big), \end{aligned} \tag{97}$$

where the last term disappears. Then, we can use the gradient of the first term, which we computed in (94). A very similar computation yields

$$\nabla_K \mathrm{OT}_{d^2}^{\epsilon}(K, K) = \frac{\epsilon}{2} K^{-\frac{1}{2}} \left(I + \frac{4}{\epsilon} K - \left(I + \frac{16}{\epsilon^2} K^2 \right)^{\frac{1}{2}} \right) K^{-\frac{1}{2}}. \tag{98}$$

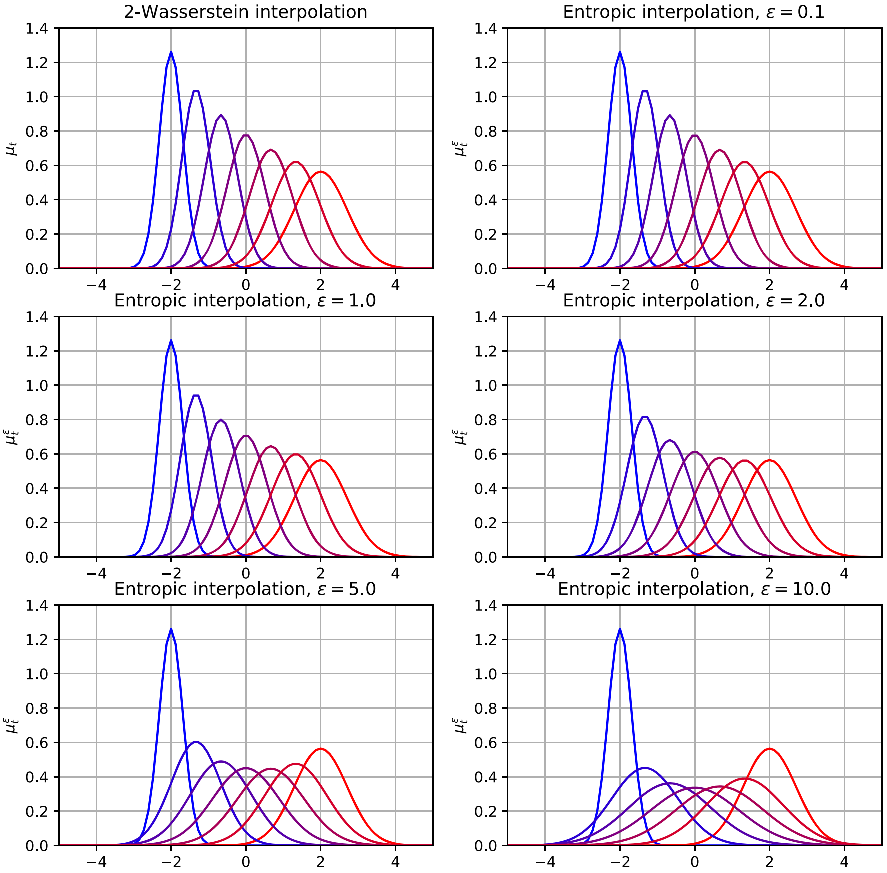


Fig. 1 Entropic interpolants μ_t^ϵ between two one-dimensional Gaussians given by $\mu_0 = \mathcal{N}(-2, 0.1)$ (blue) and $\mu_1 = \mathcal{N}(2, 0.5)$ (red), with varying regularization strengths ϵ, accompanied by the 2-Wasserstein interpolant in the top-left corner (corresponding to $\epsilon = 0$)

Substituting (94) and (98) into (97) yields

$$\begin{aligned}&\nabla_K \sum_{i=1}^{N} \lambda_i S_2^\epsilon \left(\mathcal{N}(0, K), \mathcal{N}(0, K_i)\right)\\&= \frac{\epsilon}{4} \sum_{i=1}^{N} \lambda_i K^{-\frac{1}{2}} \left(\left(I + \frac{16}{\epsilon^2} K^2 \right)^{\frac{1}{2}} - \left(I + \frac{16}{\epsilon^2} K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) K^{-\frac{1}{2}}.\end{aligned} \tag{99}$$

When (99) is set to zero, we find, that the optimal $\bar{K}$ satisfies the relation given in (96). □

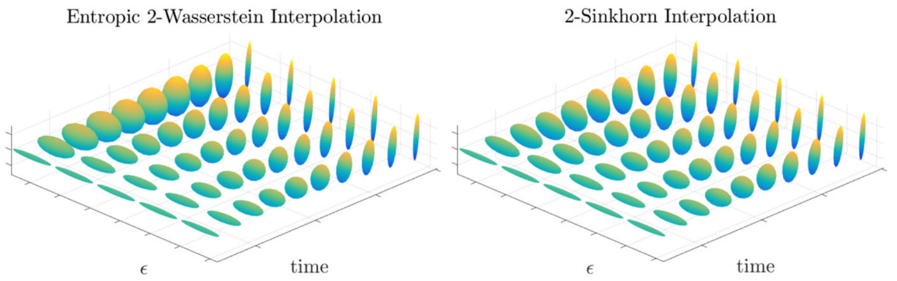


Fig. 2 Interpolants between two three-dimensional Gaussians with varying regularization strengths ϵ, accompanied by the 2-Wasserstein interpolant, given by the first row (parallel to the time axis). The following rows visualize the interpolation for $\epsilon \in \{0.01, 1, 2, 5, 20\}$ in increasing order

4.3 Existence and uniqueness of solution

Theorems 3 and 4 derive the fixed point equations, namely Eqs. (87) and (96), respectively, that the corresponding barycenter must satisfy, under the assumption that it is strictly positive. For the Sinkhorn barycenter in Theorem 4, existence and uniqueness of solution was shown in [45] via the Brouwer Fixed Point Theorem, under the assumption that all K_i's are strictly positive. For the entropic barycenter in Theorem 3, a non-trivial solution exists, in which case it is unique, only when ε is sufficiently small, otherwise it is the Dirac δ-measure. This was shown in one-dimension by [44] and for any finite dimension by [62]. The more general setting, where the barycenter can be singular, is treated in [62].

4.4 Fixed-point iteration

The fixed-point iteration algorithm is defined by

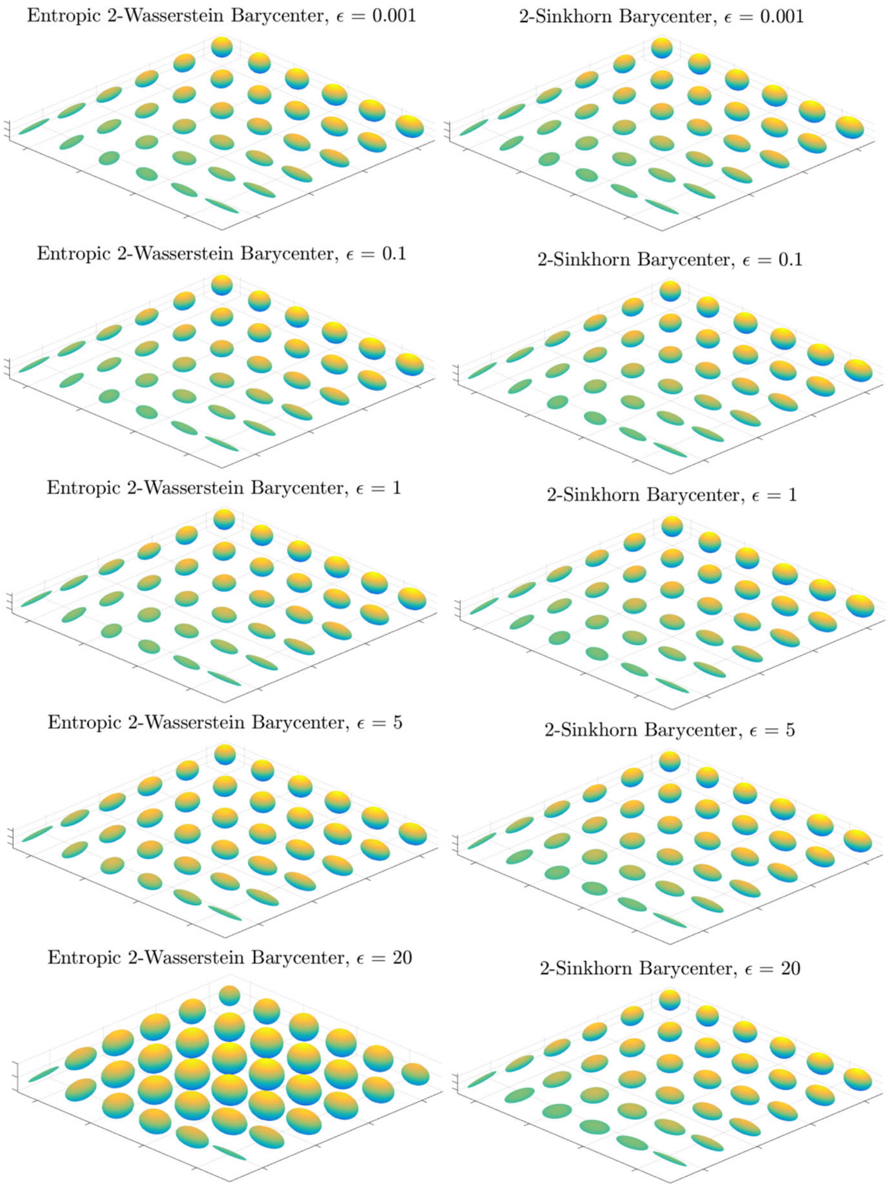
$$x_{k+1} = F(x_k), \tag{100}$$

where the initial case x_0 is handpicked by the user. The *Banach fixed-point theorem* is a well-known result stating that such an iteration converges to a fixed-point, i.e. an element x satisfying $x = F(x)$, if F is a *contraction mapping*.

In the case of the 2-Wasserstein barycenter given in (11), the fixed-point iteration can be shown to converge [2] to the unique barycenter. In the entropic 2-Wasserstein and the 2-Sinkhorn cases we leave such a proof as future work. However, while computing the numerical results in Sect. 5, the fixed-point iteration always succeeded to converge.

5 Numerical illustrations

We will now illustrate the resulting entropic 2-Wasserstein distance and 2-Sinkhorn divergence for Gaussians by employing the closed-form solutions to visualize entropic interpolations between end point Gaussians. Furthermore, we employ the fixed-point



Entropic 2-Wasserstein Barycenter, $\epsilon = 0.001$

2-Sinkhorn Barycenter, $\epsilon = 0.001$

Entropic 2-Wasserstein Barycenter, $\epsilon = 0.1$

2-Sinkhorn Barycenter, $\epsilon = 0.1$

Entropic 2-Wasserstein Barycenter, $\epsilon = 1$

2-Sinkhorn Barycenter, $\epsilon = 1$

Entropic 2-Wasserstein Barycenter, $\epsilon = 5$

2-Sinkhorn Barycenter, $\epsilon = 5$

Entropic 2-Wasserstein Barycenter, $\epsilon = 20$

2-Sinkhorn Barycenter, $\epsilon = 20$

Fig. 3 Barycentric spans of the four corner tensors under the entropic 2-Wasserstein metric and the 2-Sinkhorn divergence for varying ϵ

iteration (100) in conjunction with the fixed-point expressions of the barycenters for their visualization.

First, we consider the interpolant between one-dimensional Gaussians given in Fig. 1, where the densities of the interpolants are plotted. As one can see, increasing ϵ causes the middle of the interpolation to flatten out. This results from the Fokker–

Planck equation (31), which governs the diffusion of the evolution of processes that are objected to Brownian noise. In the limit $\epsilon \to \infty$, we would witness a heat death of the distribution.

The same can be seen in the three-dimensional case, depicted in Fig. 2, visualized using the code accompanying [29]. Here, the ellipsoids are determined by the eigenvectors and -values of the covariance matrix of the corresponding Gaussian, and the colors visualize the level sets of the ellipsoids. Note that a large ellipsoid corresponds to high variance in each direction, and does not actually increase the mass of the distribution. Such visualizations are common in *diffusion tensor imaging* (DTI), where the tensors (covariance matrices) define Gaussian diffusion of water at voxels images produced by magnetic resonance imaging (MRI) [7].

Finally, we consider the entropic 2-Wasserstein and Sinkhorn barycenters in Fig. 3. We consider four different Gaussians, placed in the corners of the square fields in the figure, and plot the barycenters for varying weights, resulting in the *barycentric span* of the four Gaussians. As the results show, the barycenters are very similar under the two frameworks with small ϵ. However, as ϵ is increased, the Sinkhorn barycenter seems to be more resiliant against the fattening of the barycenters, which can be seen in the 2-Wasserstein case.

Acknowledgements This work was initiated during the authors' stay at the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation (Grant No. DMS-1440415). AM was supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation, and by the Academy of Finland (Flagship programme: 328400), acknowledging the computational resources provided by Aalto Science-IT project. AG acknowledges funding by the European Research Council under H2020/MSCA-IF OTmeetsDFT (grant ID: 795942).

Funding Open access funding provided by Aalto University.

Appendix A: Distributional solutions of Fokker–Planck equation

We just recall the definition of distributional solution of the Fokker-Planck equation.

Definition 1 We say that a family of pairs measures/vector fields (η_t, v_t) with $v_t \in L^1(\eta_t; \mathbb{R}^n)$ and $\int_0^1 \|v_t\|_{L^1(\eta_t)} dt = \int_0^1 \int_{\mathbb{R}^n} |v_t| d\eta_t dt$ solves the continuity equation on $]0, T[$ in the distributional sense if for any bounded and Lipschitz test function $f \in C_c^1(]0, T[\times \mathbb{R}^n)$

$$\int_0^1 \int_{\mathbb{R}^n} (\partial_t f) d\eta_t dt + \int_0^1 \int_{\mathbb{R}^n} \left(\nabla f \cdot v_t - \frac{\varepsilon}{2} \Delta f \right) d\eta_t dt = 0.$$

Appendix B: Alternative Proof of Theorem 2b

Recall, that by Propositions 2 and 3, we can restrict to plans that are centered Gaussians, that is,

$$\gamma = \mathcal{N}(0, \Gamma), \quad \Gamma = \begin{bmatrix} K_1 & C^T \\ C & K_2 \end{bmatrix}. \tag{101}$$

Substituting (101) into (17) yields

$$\begin{aligned} \mathrm{OT}^{\epsilon}_{d^2}(\mu_1, \mu_2) &= \min_{C \in \mathbb{R}^{n\times n}} F(C) \\ &:= \min_{C \in \mathbb{R}^{n\times n}} \Bigg\{ \mathrm{Tr}(K_1) + \mathrm{Tr}(K_2) \\ &\qquad - 2\,\mathrm{Tr}(C) + \frac{\epsilon}{2} \log\left(\frac{\det(K_1K_2)}{\det(\Gamma)}\right)\Bigg\}. \end{aligned} \tag{102}$$

The covariance matrix Γ should be a symmetric positive-definite matrix, which is equivalent to its Schur complement $S(C)$ being positive definite, that is,

$$S(C) := K_1 - C^T K_2^{-1} C \succeq 0. \tag{103}$$

If $S(C)$ fails to be strictly positive definite, $F(C)$ explodes to infinity, and so it suffices to consider C so that

$$S(C) \succ 0. \tag{104}$$

Now recall the Schur block matrix determinant formula

$$\det(\Gamma) = \det(S(C))\det(K_2). \tag{105}$$

Then, following the argumentation in the proof of [42, Prop. 7], when the value of $S(C) = S$ is fixed, we can write

$$\max_{C\,:\,S(C)=S} \mathrm{Tr}(C) = \mathrm{Tr}\left(K_2^{\frac{1}{2}}(K_1 - S)K_2^{\frac{1}{2}}\right)^{\frac{1}{2}}, \tag{106}$$

and so applying (105) and (106) to (102), we get

$$\begin{aligned} \min_{C\,:\,S(C)=S} F(C) = \mathrm{Tr}(K_1) + \mathrm{Tr}(K_2) - 2\mathrm{Tr}\left(K_2^{\frac{1}{2}}(K_1 - S)K_2^{\frac{1}{2}}\right)^{\frac{1}{2}} \\ + \frac{\epsilon}{2}\left(\log\det(K_1) - \log\det(S)\right), \end{aligned} \tag{107}$$

leaving us with the task of minimizing (107) with respect to S. Note that we could maximize (106) independently with respect to C, as $\det(\Gamma)$ is constant over the fiber $\{C : S(C) = S\}$.

As F is strictly convex with respect to S, a solution to (102) can be found when the gradient of the expression with respect to S is zero, leading to

$$\nabla_S F(S) = K_2^{\frac{1}{2}} \left(K_2^{\frac{1}{2}} \left(K_1 - S\right) K_2^{\frac{1}{2}} \right)^{-\frac{1}{2}} K_2^{\frac{1}{2}} - \frac{\epsilon}{2} S^{-1} = 0. \tag{108}$$

Moving the second term to RHS, multiplying (108) by $(K_1 - S)^{\frac{1}{2}}$ from right, multiplying each side by their corresponding transposes, and some elementary manipulations of the equation, we arrive at a *continuous algebraic Riccati equation* (CARE)

$$\epsilon^2 K_1 - \epsilon^2 S - 4SK_2S = 0. \tag{109}$$

In general, CAREs do not admit an analytical solution. However, we are in luck, as one can check that (109) is solved by

$$\hat{S} = \frac{\epsilon}{8} K_2^{-\frac{1}{2}} \left(-\epsilon I + \left(\epsilon^2 I + 16 K_2^{\frac{1}{2}} K_1 K_2^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) K_2^{-\frac{1}{2}}. \tag{110}$$

Finally, it is straight-forward to check that the solution $\hat{S}$ is indeed symmetric and positive-definite, and therefore satisfies (104). Plugging $\hat{S}$ in (107), noticing that $K_2^{\frac{1}{2}} K_1 K_2^{\frac{1}{2}}$ has same eigenvalues as $K_1 K_2$, and some simplifications concludes the proof.

Now, we compute the OT quantity given $\hat{S}$. We first compute the trace term (107), which gives

$$\begin{aligned}
\operatorname{Tr}\left(K_2^{\frac{1}{2}} (K_1 - \hat{S}) K_2^{\frac{1}{2}} \right)^{\frac{1}{2}} &= \operatorname{Tr}\left(K_1 K_2 - \frac{\epsilon}{8} \left(-\epsilon I + \left(\epsilon^2 I + 16 K_1 K_2 \right)^{\frac{1}{2}} \right) \right)^{\frac{1}{2}} \\
&= \operatorname{Tr}\left(\frac{\epsilon^2}{16} I + K_1 K_2 + \frac{\epsilon^2}{16} I - \frac{\epsilon^2}{8} \left(I + \frac{16}{\epsilon^2} K_1 K_2 \right)^{\frac{1}{2}} \right)^{\frac{1}{2}} \\
&= \frac{\epsilon}{4} \operatorname{Tr}\left(\left(-I + \left(I + \frac{16}{\epsilon^2} K_1 K_2 \right)^{\frac{1}{2}} \right)^2 \right)^{\frac{1}{2}} \\
&= \frac{\epsilon}{4} \operatorname{Tr}\left(-I + \left(I + \frac{16}{\epsilon^2} K_1 K_2 \right)^{\frac{1}{2}} \right) \\
&= \frac{\epsilon}{4} \left(\operatorname{Tr}\left(M_\epsilon \right) - 2n \right)
\end{aligned} \tag{111}$$

For the other term, write $\{\lambda_i\}_{i=1}^n$ for the eigenvalues of $K_1 K_2$ and $m_i = 1 + \frac{16}{\epsilon^2} \lambda_i$

$$\log\det(K_1) - \log\det(\hat{S}) = \log\det(K_1 K_2)$$

$$
\begin{aligned}
&- \log\det\left(\frac{\epsilon^2}{8}\left(-I + \left(I + \frac{16}{\epsilon^2}K_1K_2\right)^{\frac{1}{2}}\right)\right)\\
&= \sum_{i=1}^{n} \log\left(\frac{\epsilon^2(m_i - 1)}{16}\right) - \sum_{i=1}^{n} \log\left(\frac{\epsilon^2}{8}(m_i^{\frac{1}{2}} - 1)\right)\\
&= \sum_{i=1}^{n} \log\left(\frac{1}{2}(1 + m_i^{\frac{1}{2}})\right)\\
&= \sum_{i=1}^{n} \log\left(1 + \left(1 + \frac{16}{\epsilon^2}\lambda_i\right)^{\frac{1}{2}}\right) - n\log 2\\
&= \log\det(M_\epsilon) - n\log 2. \qquad (112)
\end{aligned}
$$

References

1. Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM J. Math. Anal. **43**(2), 904–924 (2011)
2. Álvarez-Esteban, P.C., Barrio, E., Cuesta-Albertos, J.A., Matrán, C.: A fixed-point approach to barycenters in Wasserstein space. J. Math. Anal. Appl. **441**(2), 744–762 (2016)
3. Amari, S.: Information Geometry and its Applications, vol. 194. Springer, Berlin (2016)
4. Amari, S., Karakida, R., Oizumi, M.: Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. Inf. Geom. **1**(1), 13–37 (2018)
5. Ambrosio, L., Gigli, N., Savaré, G.: Gradient Flows: In Metric Spaces and in the Space of Probability Measures. Springer Science & Business Media, New York (2008)
6. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, 7–9 August, 2017 (2017)
7. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-Euclidean metrics for fast and simple calculus on diffusion tensors. Magn. Reson. Med. **56**(2), 411–421 (2006)
8. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. **29**(1), 328–347 (2007)
9. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry, vol. 64. Springer, Berlin (2017)
10. Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative Bregman projections for regularized transportation problems. SIAM J. Sci. Comput. **37**(2), A1111–A1138 (2015)
11. Bigot, J., Cazelles, E., Papadakis, N.: Penalization of barycenters in the Wasserstein space. SIAM J. Math. Anal. **51**(3), 2261–2285 (2019)
12. Borwein, J.M., Lewis, A.S., Nussbaum, R.D.: Entropy minimization, DAD problems, and doubly stochastic kernels. J. Funct. Anal. **123**(2), 264–307 (1994)
13. Cazelles, E., Bigot, J., Papadakis, N.: Regularized barycenters in the Wasserstein space. In: International Conference on Geometric Science of Information, pp. 83–90. Springer (2017)
14. Chebbi, Z., Moakher, M.: Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function. Linear Algebra Appl. **436**(7), 1872–1889 (2012)
15. Chen, Y., Georgiou, T.T., Pavon, M.: Optimal steering of a linear stochastic system to a final probability distribution, part I. IEEE Trans. Autom. Control **61**(5), 1158–1169 (2015)
16. Chen, Y., Georgiou, T.T., Pavon, M.: On the relation between optimal transport and Schrödinger bridges: a stochastic control viewpoint. J. Optim. Theory Appl. **169**(2), 671–691 (2016)
17. Cichocki, A., Cruces, S., Amari, S.: Log-determinant divergences revisited: alpha-beta and gamma log-det divergences. Entropy **17**(5), 2988–3034 (2015)
18. Congedo, M., Barachant, A., Bhatia, R.: Riemannian geometry for EEG-based brain-computer interfaces; a primer and a review. Brain-Comput. Interfaces **4**(3), 155–174 (2017)

19. Csiszár, I.: I-divergence geometry of probability distributions and minimization problems. Ann. Probab. **3**:146–158 (1975)
20. Cuturi, M.: Sinkhorn distances: lightspeed computation of optimal transport. Adv. Neural Inf. Process. Syst. **26**, 2292–2300 (2013)
21. Cuturi, M., Doucet, A.: Fast computation of Wasserstein barycenters. In: International Conference on Machine Learning, pp. 685–693 (2014)
22. Cuturi, M., Peyré, G.: Computational optimal transport. Found. Trends® Mach. Learn. **11**(5–6), 355–607 (2019)
23. del Barrio, E., Loubes, J.-M.: The statistical effect of entropic regularization in optimal transportation. arXiv preprint arXiv:2006.05199 (2020)
24. Deshpande, I., Zhang, Z., Schwing, A.: Generative modeling using the sliced Wasserstein distance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3483–3491 (2018)
25. Di Marino, S., Gerolin, A.: An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. J. Sci. Comput. **85**(2), 1–28 (2020)
26. Dowson, D.C., Landau, B.V.: The Fréchet distance between multivariate normal distributions. J. Multivar. Anal. **12**(3), 450–455 (1982)
27. Dryden, I.L., Koloydenko, A., Zhou, D.: Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. **3**, 1102–1123 (2009)
28. Dukler, Y., Li, W., Lin, A., Montúfar, G.: Wasserstein of Wasserstein loss for learning generative models. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 97, pp. 1716–1725 (2019)
29. Feragen, A., Fuster, A.: Geometries and interpolations for symmetric positive definite matrices. In: Modeling, Analysis, and Visualization of Anisotropy, pp. 85–113. Springer (2017)
30. Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S., Trouve, A., Peyré, G.: Interpolating between optimal transport and MMD using Sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 2681–2690 (2019)
31. Franklin, J., Lorenz, J.: On the scaling of multidimensional matrices. Linear Algebra Appl. **114**, 717–735 (1989)
32. Galichon, A.: Optimal Transport Methods in Economics. Princeton University Press, Princeton (2018)
33. Galichon, A., Salanié, B.: Matching with Trade-Offs: Revealed Preferences Over Competing Characteristics. Sciences po publications, Sciences Po (2010)
34. Genevay, A., Chizat, L., Bach, F., Cuturi, M., Peyré, G.: Sample Complexity of Sinkhorn Divergences. In: Chaudhuri, K., Sugiyama, M. (eds.) Proceedings of Machine Learning Research, Proceedings of Machine Learning Research, vol. 89, pp. 1574–1583 (2019)
35. Genevay, A., Cuturi, M., Peyré, G., Bach, F.: Stochastic optimization for large-scale optimal transport. Adv. Neural Inf. Process. Syst. **29**, 3440–3448 (2016)
36. Genevay, A., Peyre, G., Cuturi, M.: Learning Generative Models with Sinkhorn Divergences. In: Storkey, A., Perez-Cruz, F. (eds.) Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, vol. 84, pp. 1608–1617 (2018)
37. Gentil, I., Léonard, C., Ripani, L.: About the analogy between optimal transport and minimal entropy. In: Annales de la Faculté des Sciences de Toulouse. Mathématiques, vol. 3, pp. 569–600 (2017)
38. Gerolin, A., Grossi, J., Gori-Giorgi, P.: Kinetic correlation functionals from the entropic regularisation of the strictly-correlated electrons problem. J. Chem. Theory Comput. **16**(1), 488–498 (2019)
39. Gerolin, A., Kausamo, A., Rajala, T.: Multi-marginal entropy-transport with repulsive cost. Calc. Var. Partial Differ. Equ. **59**(3), 90 (2020)
40. Gigli, N., Tamanini, L.: Second order differentiation formula on $RCD^*(K, N)$ spaces. J. Eur. Math. Soc. (JEMS) (2018)
41. Gigli, N., Tamanini, L.: Benamou–Brenier and duality formulas for the entropic cost on $RCD^*(K, N)$ spaces. Probab. Theory Relat. Fields (2018)
42. Givens, C.R., Shortt, R.M.: A class of Wasserstein metrics for probability distributions. Mich. Math. J. **31**(2), 231–240 (1984)
43. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
44. Janati, H., Cuturi, M., Gramfort, A.: Debiased Sinkhorn barycenters. In: Proceedings of the 37th International Conference on Machine Learning, pp. 4692–4701 (2020)

45. Janati, H., Muzellec, B., Peyré, G., Cuturi, M.: Entropic optimal transport between unbalanced Gaussian measures has a closed form. Adv. Neural Inf. Process. Syst. **33** (2020)
46. Knott, M., Smith, C.S.: On the optimal mapping of distributions. J. Optim. Theory Appl. **43**(1), 39–49 (1984)
47. Kroshnin, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A., Tupitsa, N., Uribe, C.: On the complexity of approximating Wasserstein barycenter. In: International Conference on Machine Learning, pp. 3530–3540 (2019)
48. Kum, S., Duong, M.H., Lim, Y., Yun, S.: Penalization of barycenters for φ-exponential distributions. arXiv preprint arXiv:2006.08743 (2020)
49. Larotonda, G.: Nonpositive curvature: a geometrical approach to Hilbert–Schmidt operators. Differ. Geom. Appl. **25**, 679–700 (2007)
50. Léonard, C.: A survey of the Schrödinger problem and some of its connections with optimal transport. Discrete Contin. Dyn. Syst. A **34**(4), 1533–1574 (2014)
51. Lin, T., Ho, N., Cuturi, M., Jordan, M.I.: On the complexity of approximating multimarginal optimal transport. arXiv preprint arXiv:1910.00152 (2019)
52. Lunz, S., Öktem, O., Schönlieb, C.-B.: Adversarial regularizers in inverse problems. In: Advances in Neural Information Processing Systems, pp. 8507–8516 (2018)
53. Malagò, L., Montrucchio, L., Pistone, G.: Wasserstein Riemannian geometry of Gaussian densities. Inf. Geom. **1**(2), 137–179 (2018)
54. Mallasto, A., Feragen, A.: Learning from uncertain curves: the 2-Wasserstein metric for Gaussian processes. Adv. Neural Inf. Process. Syst. **30**, 5660–5670 (2017)
55. Mallasto, A., Frellsen, J., Boomsma, W., Feragen, A.: (q, p)-Wasserstein GANs: comparing ground metrics for Wasserstein GANs. arXiv preprint arXiv:1902.03642 (2019)
56. Mallasto, A., Montúfar, G., Gerolin, A.: How well do WGANs estimate the Wasserstein metric? arXiv:1910.03875 (2019)
57. Masarotto, V., Panaretos, V.M., Zemel, Y.: Procrustes metrics on covariance operators and optimal transportation of Gaussian processes. Sankhya A, pp. 1–42 (2018)
58. McCann, R.J.: A convexity principle for interacting gases. Adv. Math. **128**(1), 153–179 (1997)
59. Mena, G., Weed, J.: Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In: Advances in Neural Information Processing Systems (2019)
60. Minh, H.Q.: Infinite-dimensional Log-Determinant divergences between positive definite trace class operators. Linear Algebra Appl. **528**, 331–383 (2017)
61. Minh, H.Q., San Biagio, M., Murino, V.: Log-Hilbert–Schmidt metric between positive definite operators on Hilbert spaces. Adv. Neural Inf. Process. Syst. **27**, 388–396 (2014)
62. Minh, H.Q.: Entropic regularization of Wasserstein distance between infinite-dimensional Gaussian measures and Gaussian processes. preprint arXiv:2011.07489 (2020)
63. Minh, H.Q.: Convergence and finite sample approximations of entropic regularized Wasserstein distances in Gaussian and RKHS settings. arXiv preprint arXiv:2101.01429 (2021)
64. Müller, A.: Integral probability metrics and their generating classes of functions. Adv. Appl. Probab. **29**(2), 429–443 (1997)
65. Muzellec, B., Cuturi, M.: Generalizing point embeddings using the Wasserstein space of elliptical distributions. Adv. Neural Inf. Process. Syst. **31**, 10237–10248 (2018)
66. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra Appl. **48**, 257–263 (1982)
67. Patrini, G., van den Berg, R., Forre, P., Carioni, M., Bhargav, S., Welling, M., Genewein, T., Nielsen, F.: Sinkhorn Autoencoders. In: Uncertainty in Artificial Intelligence, pp. 733–743 (2020)
68. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. Int. J. Comput. Vis. **66**(1), 41–66 (2006)
69. Ramdas, A., Trillos, N., Cuturi, M.: On Wasserstein two-sample testing and related families of non-parametric tests. Entropy **19**(2), 47 (2017)
70. Ripani, L.: The Schrödinger problem and its links to optimal transport and functional inequalities. Ph.D. thesis, University Lyon 1 (2017)
71. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. Int. J. Comput. Vis. **40**(2), 99–121 (2000)
72. Ruschendorf, L.: Convergence of the iterative proportional fitting procedure. Ann. Stat. **23**(4), 1160–1174 (1995)

73. Rüschendorf, L., Thomsen, W.: Note on the Schrödinger equation and I-projections. Stat. Probab. Lett. **17**(5), 369–375 (1993)
74. Rüschendorf, L., Thomsen, W.: Closedness of sum spaces and the generalized Schrödinger problem. Theory Probab. Appl. **42**(3), 483–494 (1998)
75. Schrödinger, E.: Über die umkehrung der naturgesetze. Verlag Akademie der wissenschaften in kommission bei Walter de Gruyter u Company (1931)
76. Sommerfeld, M.: Wasserstein distance on finite spaces: Statistical inference and algorithms. PhD thesis, Georg-August-Universität Göttingen (2017)
77. Takatsu, A.: Wasserstein geometry of Gaussian measures. Osaka J. Math. **48**(4), 1005–1026 (2011)
78. Thanwerdas, Y., Pennec, X.: Exploration of balanced metrics on symmetric positive definite matrices. In: International Conference on Geometric Science of Information, pp. 484–493. Springer (2019)
79. Tuzel, O., Porikli, F., Meer, P.: Region covariance: a fast descriptor for detection and classification. In: European Conference on Computer Vision, pp. 589–600. Springer (2006)
80. Tuzel, O., Porikli, F., Meer, P.: Human detection via classification on Riemannian manifolds. CVPR **1**, 4 (2007)
81. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on Riemannian manifolds. IEEE Trans. Pattern Anal. Mach. Intell. **30**(10), 1713–1727 (2008)
82. Villani, C.: Optimal transport: Old and New, Grundlehren der mathematischen Wissenschaften, vol. 338. Springer Science & Business Media (2008)
83. Weed, J., Bach, F.: Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli **25**(4A), 2620–2648 (2019)
84. Zambrini, J.-C.: The research program of stochastic deformation (with a view toward geometric mechanics). In Stochastic Analysis: A Series of Lectures, pp. 359–393. Springer (2015)