Article

Entropy-Regularized Optimal Transport on Multivariate Normal and q-Normal Distributions

Qijun Tong * and Kei Kobayashi

Department of Mathematics, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan; kei@math.keio.ac.jp
* Correspondence: tong.qujin@gmail.com

Abstract: The distance and divergence of the probability measures play a central role in statistics, machine learning, and many other related fields. The Wasserstein distance has received much attention in recent years because of its distinctions from other distances or divergences. Although computing the Wasserstein distance is costly, entropy-regularized optimal transport was proposed to computationally efficiently approximate the Wasserstein distance. The purpose of this study is to understand the theoretical aspect of entropy-regularized optimal transport. In this paper, we focus on entropy-regularized optimal transport on multivariate normal distributions and q-normal distributions. We obtain the explicit form of the entropy-regularized optimal transport cost on multivariate normal and q-normal distributions; this provides a perspective to understand the effect of entropy regularization, which was previously known only experimentally. Furthermore, we obtain the entropy-regularized Kantorovich estimator for the probability measure that satisfies certain conditions. We also demonstrate how the Wasserstein distance, optimal coupling, geometric structure, and statistical efficiency are affected by entropy regularization in some experiments. In particular, our results about the explicit form of the optimal coupling of the Tsallis entropy-regularized optimal transport on multivariate q-normal distributions and the entropy-regularized Kantorovich estimator are novel and will become the first step towards the understanding of a more general setting.

Keywords: optimal transportation; entropy regularization; Wasserstein distance; Tsallis entropy; q-normal distribution

1. Introduction

Comparing probability measures is a fundamental problem in statistics and machine learning. A classical way to compare probability measures is the Kullback–Leibler divergence. Let M be a measurable space and μ, ν be the probability measure on M; then, the Kullback–Leibler divergence is defined as:

$$\mathrm{KL}(\mu|\nu) = \int_M d\mu \log \frac{d\mu}{d\nu}. \tag{1}$$

The Wasserstein distance [1], also known as the earth mover distance [2], is another way of comparing probability measures. It is a metric on the space of probability measures derived by the mass transportation theory of two probability measures. Informally, optimal transport theory considers an optimal transport plan between two probability measures under a cost function, and the Wasserstein distance is defined by the minimum total transport cost. A significant difference between the Wasserstein distance and the Kullback–Leibler divergence is that the former can reflect the metric structure, whereas the latter cannot. The Wasserstein distance can be written as:

$$W_p(\mu, \nu) := \left\{ \inf_{\pi \in \Pi(\mu,\nu)} \int_{M \times M} d(x,y)^p d\pi(x,y) \right\}^{\frac{1}{p}}, \tag{2}$$

where $d(\cdot, \cdot)$ is a distance function on a measurable metric space M and $\Pi(\mu, \nu)$ denotes the set of probability measures on $M \times M$, whose marginal measures correspond to μ and ν. In recent years, the application of optimal transport and the Wasserstein distance has been studied in many fields such as statistics, machine learning, and image processing. For example, Reference [3] generated the interpolation of various three-dimensional (3D) objects using the Wasserstein barycenter. In the field of word embedding in natural language processing, Reference [4] embedded each word as an elliptical distribution, and the Wasserstein distance was applied between the elliptical distributions. There are many studies on the applications of optimal transport to deep learning, including [5–7]. Moreover, Reference [8] analyzed the denoising autoencoder [9] with gradient flow in the Wasserstein space.

In the application of the Wasserstein distance, it is often considered in a discrete setting where μ and ν are discrete probability measures. Then, obtaining the Wasserstein distance between μ and ν can be formulated as a linear programming problem. In general, however, it is computationally intensive to solve such linear problems and obtain the optimal coupling of two probability measures. For such a situation, a novel numerical method, entropy regularization, was proposed by [10],

$$C_\lambda(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} c(x, y)\pi(x, y)dxdy - \lambda \mathrm{Ent}(\pi). \tag{3}$$

This is a relaxed formulation of the original optimal transport of a cost function $c(\cdot, \cdot)$, in which the negative Shannon entropy $-\mathrm{Ent}(\cdot)$ is used as a regularizer. For a small λ, $C_\lambda(\mu, \nu)$ can approximate the p-th power of the Wasserstein distance between two discrete probability measures, and it can be computed efficiently by using Sinkhorn's algorithm [11].

More recently, many studies have been published on improving the computational efficiency. According to [12], the most computationally efficient algorithm at this moment to solve the linear problem for the Wasserstein distance is Lee–Sidford linear solver [13], which runs in $O(n^{2.5})$. Reference [14] proved that a complexity bound for the Sinkhorn algorithm is $\tilde{O}(n^2 \varepsilon^{-2})$, where ε is the desired absolute performance guarantee. After [10] appeared, various algorithms have been proposed. For example, Reference [15] adopted stochastic optimization schemes for solving the optimal transport. The Greenkhorn algorithm [16] is the greedy variant of the Sinkhorn algorithm, and Reference [12] proposed its acceleration. Many other approaches such as adapting a variety of standard optimization algorithms to approximate the optimal transport problem can be found in [12,17–19]. Several specialized Newton-type algorithms [20,21] achieve complexity bound $\tilde{O}(n^2 \varepsilon^{-1})$ [22,23], which are the best ones in terms of computational complexity at the present moment.

Moreover, entropy-regularized optimal transport has another advantage. Because of the differentiability of the entropy-regularized optimal transport and the simple structure of Sinkhorn's algorithm, we can easily compute the gradient of the entropy-regularized optimal transport cost and optimize the parameter of a parametrized probability distribution by using numerical differentiation or automatic differentiation. Then, we can define a differentiable loss function that can be applied to various supervised learning methods [24]. Entropy-regularized optimal transport can be used to approximate not only the Wasserstein distance, but also its optimal coupling as a mapping function. Reference [25] adopted the optimal coupling of the entropy-regularized optimal transport as a mapping function from one domain to another.

Despite the empirical success of the entropy-regularized optimal transport, its theoretical aspect is less understood. Reference [26] studied the expected Wasserstein distance between a probability measure and its empirical version. Similarly, Reference [27] showed the consistency of the entropy-regularized optimal transport cost between two empirical distributions. Reference [28] showed that minimizing the entropy-regularized optimal transport cost between empirical distributions is equivalent to a type of maximum likelihood estimator. Reference [29] considered Wasserstein generative adversarial networks

with an entropy regularization. Reference [30] constructed information geometry from the convexity of the entropy-regularized optimal transport cost.

Our intrinsic motivation of this study is to produce an analytical solution about the entropy-regularized optimal transport problem between continuous probability measures so that we can gain insight into the effects of entropy regularization in a theoretical, as well as an experimental way. In our study, we generalized the Wasserstein distance between two multivariate normal distributions by entropy regularization. We derived the explicit form of the entropy-regularized optimal transport cost and its optimal coupling, which can be used to analyze the effect of entropy regularization directly. In general, the nonregularized Wasserstein distance between two probability measures and its optimal coupling cannot be expressed in a closed form; however, Reference [31] proved the explicit formula for multivariate normal distributions. Theorem 1 is a generalized form of [31]. We obtain an explicit form of the entropy-regularized optimal transport between two multivariate normal distributions. Furthermore, by adopting the Tsallis entropy [32] as the entropy regularization instead of the Shannon entropy, our theorem can be generalized to multivariate q-normal distributions.

Some readers may find it strange to study the entropy-regularized optimal transport for multivariate normal distributions, where the exact (nonregularized) optimal transport has been obtained explicitly. However, we think it is worth studying from several perspectives:

- Normal distributions are the simplest and best-studied probability distributions, and thus, it is useful to examine the regularization theoretically in order to infer results for other distributions. In particular, we will partly answer the questions "How much do entropy constraints affect the results?" and "What does it mean to constrain by the entropy?" for the simplest cases. Furthermore, as a first step in constructing a theory for more general probability distributions, in Section 4, we propose a generalization to multivariate q-normal distributions.

- Because normal distributions are the limit distributions in asymptotic theories using the central limit theorem, studying normal distributions is necessary for the asymptotic theory of regularized Wasserstein distances and estimators computed by them. Moreover, it was proposed to use the entropy-regularized Wasserstein distance to compute a lower bound of the generalization error for a variational autoencoder [29]. The study of the asymptotic behavior of such bounds is one of the expected applications of our results.

- Though this has not yet been proven theoretically, we suspect that entropy regularization is efficient not only for computational reasons, such as the use of the Sinkhorn algorithm, but also in the sense of efficiency in statistical inference. Such a phenomenon can be found in some existing studies, including [33]. Such statistical efficiency is confirmed by some experiments in Section 6.

The remainder of this paper is organized as follows. First, we review some definitions of optimal transport and entropy regularization in Section 2. Then, in Section 3, we provide an explicit form of the entropy-regularized optimal transport cost and its optimal coupling between two multivariate normal distributions. We also extend this result to q-normal distributions for Tsallis entropy regularization in Section 4. In Section 5, we obtain the entropy-regularized Kantorovich estimator of probability measures on $\mathbb{R}^n$ with a finite second moment that are absolutely continuous with respect to the Lebesgue measure in Theorem 3. We emphasize that Theorem 3 is not limited to the case of multivariate normal distribution, but can handle a wider range of probability measures. We analyze how entropy regularization affects the optimal result experimentally in certain sections.

We note that after publishing the preprint version of the paper, we found closely related results [34,35] reported within half a year. In Janati et al. [34], they proved the same result as Theorem 1 based on solving the fixed-point equation behind Sinkhorn's algorithm. Their results include the unbalanced optimal transport between unbalanced multivariate normal distributions. They also studied the convexity and differentiability of the objective

function of the entropy-regularized optimal transport. In [35], the same closed-form as Theorem 1 was proven by ingeniously using the Schrödinger system. Although there are some overlaps, our paper has significant novelty in the following respects. Our proof is more direct than theirs and can be extended directly to the proof for the Tsallis entropy-regularized optimal transport between multivariate q-normal distributions provided in Section 4. Furthermore, Corollaries 1 and 2 are novel and important results to evaluate how much the entropy regularization affects the estimation results or not at all. We also obtain the entropy-regularized Kantorovich estimator in Theorem 3.

2. Preliminary

In this section, we review some definitions of optimal transport and entropy-regularized optimal transport. These definitions were referred to in [1,36]. In this section, we use a tuple (M, Σ) for a set M and σ-algebra on M and $\mathcal{P}(X)$ for the set of all probability measures on a measurable space X.

Definition 1 (Pushforward measure). *Given measurable spaces (M_1, Σ_1) and (M_2, Σ_2), a measure $\mu : \Sigma_1 \to [0, +\infty]$, and a measurable mapping $\varphi : M_1 \to M_2$, the pushforward measure of μ by φ is defined by:*

$$\forall B \in \Sigma_2, \ \varphi_\# \mu(B) := \mu\left(\varphi^{-1}(B)\right). \tag{4}$$

Definition 2 (Optimal transport map). *Consider a measurable space (M, Σ), and let $c : M \times M \to \mathbb{R}_+$ denote a cost function. Given $\mu, \nu \in \mathcal{P}(M)$, we call $\varphi : M \to M$ the optimal transport map if φ realizes the infimum of:*

$$\inf_{\varphi_\# \mu = \nu} \int_M c(x, \varphi(x)) d\mu(x). \tag{5}$$

This problem was originally formalized by [37]. However, the optimal transport map does not always exist. Then, Kantorovich considered a relaxation of this problem in [38].

Definition 3 (Coupling). *Given $\mu, \nu \in \mathcal{P}(M)$, the coupling of μ and ν is a probability measure on $M \times M$ that satisfies:*

$$\forall A \in \Sigma, \ \pi(A \times M) = \mu(A), \quad \pi(M \times A) = \nu(A). \tag{6}$$

Definition 4 (Kantorovich problem). *The Kantorovich problem is defined as finding a coupling π of μ and ν that realizes the infimum of:*

$$\int_{M \times M} c(x, y) d\pi(x, y). \tag{7}$$

Hereafter, let $\Pi(\mu, \nu)$ be the set of all couplings of μ and ν. When we adopt a distance function as the cost function, we can define the Wasserstein distance.

Definition 5 (Wasserstein distance). *Given $p \geq 1$, a measurable metric space (M, Σ, d), and $\mu, \nu \in \mathcal{P}(M)$ with a finite p-th moment, the p-Wasserstein distance between μ and ν is defined as:*

$$W_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left(\int_{M \times M} d(x, y)^p d\pi(x, y)\right)^{\frac{1}{p}}. \tag{8}$$

Now, we review the definition of entropy-regularized optimal transport on $\mathbb{R}^n$.

Definition 6 (Entropy-regularized optimal transport). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$, $\lambda > 0$, and let $\pi(x,y)$ be the density function of the coupling of μ and ν, whose reference measure is the Lebesgue measure. We define the entropy-regularized optimal transport cost as:*

$$C_\lambda(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} c(x,y)\pi(x,y)dxdy - \lambda \mathrm{Ent}(\pi), \tag{9}$$

where $\mathrm{Ent}(\cdot)$ *denotes the Shannon entropy of a probability measure:*

$$\mathrm{Ent}(\pi) = - \int_{\mathbb{R}^n \times \mathbb{R}^n} \pi(x,y) \log \pi(x,y)dxdy. \tag{10}$$

There is another variation in entropy-regularized optimal transport defined by the relative entropy instead of the Shannon entropy:

$$\tilde{C}_\lambda(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^n \times \mathbb{R}^n} c(x,y)\pi(x,y)dxdy + \lambda \mathrm{KL}(\pi|d\mu \otimes d\nu). \tag{11}$$

This is definable even when $\Pi(\mu,\nu)$ includes a coupling that is not absolutely continuous with respect to the Lebesgue measure. We note that when both μ and ν are absolutely continuous, the infimum is attained by the same π for C_λ and $\tilde{C}_\lambda$, and it depends only on μ and ν. In the following part of the paper, we assume the absolute continuity of μ, ν, and π with respect to the Lebesgue measure for well-defined entropy regularization.

3. Entropy-Regularized Optimal Transport between Multivariate Normal Distributions

In this section, we provide a rigorous solution of entropy-regularized optimal transport between two multivariate normal distributions. Throughout this section, we adopt the squared Euclidean distance $\|x - y\|^2$ as the cost function. To prove our theorem, we start by expressing C_λ using mean vectors and covariance matrices. The following lemma is a known result; for example, see [31].

Lemma 1. *Let $X \sim P, Y \sim Q$ be two random variables on $\mathbb{R}^n$ with means μ_1, μ_2 and covariance matrices Σ_1, Σ_2, respectively. If $\pi(x,y)$ is a coupling of P and Q, we have:*

$$\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2 \pi(x,y)dxdy = \|\mu_1 - \mu_2\|^2 + \mathrm{tr}\{\Sigma_1 + \Sigma_2 - 2\mathrm{Cov}(X,Y)\}. \tag{12}$$

Proof. Without loss of generality, we can assume X and Y are centralized, because:

$$\int \|(x - \mu_1) - (y - \mu_2)\|^2 \pi(x,y)dxdy = \int \|x-y\|^2 \pi(x,y)dxdy - \|\mu_1 - \mu_2\|^2. \tag{13}$$

Therefore, we have:

$$\int \|x-y\|^2 \pi(x,y)dxdy = E[\|X-Y\|^2] = E[\mathrm{tr}\{(X-Y)(X-Y)^{\mathrm{T}}\}]$$
$$= \mathrm{tr}\{\Sigma_1 + \Sigma_2 - 2\mathrm{Cov}(X,Y)\}. \tag{14}$$

By adding $\|\mu_1 - \mu_2\|^2$, we obtain (12). $\square$

Lemma 1 shows that $\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x-y\|^2 \pi(x,y)dxdy$ can be parameterized by the covariance matrices $\Sigma_1, \Sigma_2, \mathrm{Cov}(X,Y)$. Because Σ_1 and Σ_2 are fixed, the infinite-dimensional optimization of the coupling π is a finite-dimensional optimization of covariance matrix $\mathrm{Cov}(X,Y)$.

We prepare the following lemma to prove Theorem 1.

Lemma 2. *Under a fixed mean and covariance matrix, the probability measure that maximizes the entropy is a multivariate normal distribution.*

Lemma 2 is a particular case of the principle of maximum entropy [39], and the proof can be found in [40] Theorem 3.1.

Theorem 1. *Let $P \sim \mathcal{N}(\mu_1, \Sigma_1), Q \sim \mathcal{N}(\mu_2, \Sigma_2)$ be two multivariate normal distributions. The optimal coupling π of P and Q of the entropy-regularized optimal transport:*

$$C_\lambda(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \pi(x,y) dx dy - 4\lambda \text{Ent}(\pi). \qquad (*) \qquad (15)$$

is expressed as:

$$\pi \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_\lambda \\ \Sigma_\lambda^{\mathsf{T}} & \Sigma_2 \end{pmatrix} \right) \qquad (16)$$

where:

$$\Sigma_\lambda := \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \lambda^2 I)^{1/2}\Sigma_1^{-1/2} - \lambda I. \qquad (17)$$

Furthermore, $C_\lambda(P,Q)$ can be written as:

$$\begin{aligned} C_\lambda(P,Q) = & \|\mu_1 - \mu_2\|^2 + \text{tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \lambda^2 I)^{1/2}) \\ & -2\lambda \log|(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \lambda^2 I)^{1/2} - \lambda I| - 2\lambda n \log(2\pi\lambda) - 4\lambda n \log(2\pi) - 2\lambda n \end{aligned} \qquad (18)$$

and the relative entropy version can be written as:

$$\tilde{C}_\lambda(P,Q) = C_\lambda(P,Q) + 2\lambda \log|\Sigma_1||\Sigma_2| + 4\lambda n\{\log(2\pi) + 1\}. \qquad (19)$$

We note that we use the regularization parameter 4λ in $(*)$ for the sake of simplicity.

Proof. Although the first half of the proof can be derived directly from Lemma 2, we provide a proof of this theorem by Lagrange calculus, which will be used later for the extension to q-normal distributions. Now, we define an optimization problem that is equivalent to the entropy-regularized optimal transport as follows:

$$\text{minimize} \int \|x - y\|^2 \pi(x,y) dx dy - 4\lambda \text{Ent}(\pi) \qquad (20)$$

$$\text{subject to} \int \pi(x,y) dx = q(y) \text{ for } \forall y \in \mathbb{R}^n,$$

$$\int \pi(x,y) dy = p(x) \text{ for } \forall x \in \mathbb{R}^n. \qquad (21)$$

Here, $p(x)$ and $q(y)$ are probability density functions of P and Q, respectively. Let $\alpha(x)$, $\beta(y)$ be Lagrange multipliers that correspond to the above two constraints. The Lagrangian function of (21) is defined as:

$$\begin{aligned} L(\pi, \alpha, \beta) := & \int \|x - y\|^2 \pi(x,y) dx dy + 4\lambda \int \pi(x,y) \log \pi(x,y) dx dy \\ & - \int \alpha(x)\pi(x,y) dx dy + \int \alpha(x)p(x) dx \\ & - \int \beta(y)\pi(x,y) dx dy + \int \beta(y)q(y) dy. \end{aligned} \qquad (22)$$

Taking the functional derivative of (22) with respect to π, we obtain:

$$\delta L(\pi, \alpha, \beta) = \int \left(\|x - y\|^2 + 4\lambda \log \pi(x,y) - \alpha(x) - \beta(y) \right) \delta\pi(x,y) dx dy. \qquad (23)$$

By the fundamental lemma of the calculus of variations, we have:

$$\pi(x,y) \propto \exp\left(\alpha(x) + \beta(y) - \frac{\|x - y\|^2}{4\lambda} \right). \qquad (24)$$

Here, $\alpha(x), \beta(y)$ are determined from the constraints (21). We can assume that π is a $2n$-variate normal distribution, because for a fixed covariance matrix $\text{Cov}(X, Y)$, $-\text{Ent}(\pi)$ takes the infimum when the coupling π is a multivariate normal distribution by Lemma 2. Therefore, we can express π by using $z = (x^{\mathsf{T}}, y^{\mathsf{T}})^{\mathsf{T}}$ and a covariance matrix $\Sigma := \text{Cov}(X, Y)$ as:

$$\pi(x, y) \propto \exp\left\{-\frac{1}{2} z^{\mathsf{T}} \begin{pmatrix} \Sigma_1 & \Sigma \\ \Sigma^{\mathsf{T}} & \Sigma_2 \end{pmatrix}^{-1} z \right\}. \tag{25}$$

Putting:

$$\begin{pmatrix} \tilde{\Sigma}_1 & \tilde{\Sigma} \\ \tilde{\Sigma}^{\mathsf{T}} & \tilde{\Sigma}_2 \end{pmatrix} := \begin{pmatrix} \Sigma_1 & \Sigma \\ \Sigma^{\mathsf{T}} & \Sigma_2 \end{pmatrix}^{-1}, \tag{26}$$

we write:

$$-\frac{1}{2} z^{\mathsf{T}} \begin{pmatrix} \Sigma_1 & \Sigma \\ \Sigma^{\mathsf{T}} & \Sigma_2 \end{pmatrix}^{-1} z = -\frac{1}{2} \begin{pmatrix} x^{\mathsf{T}} & y^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \tilde{\Sigma}_1 & \tilde{\Sigma} \\ \tilde{\Sigma}^{\mathsf{T}} & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{27}$$

$$= -\frac{1}{2} x^{\mathsf{T}} \tilde{\Sigma}_1 x - \frac{1}{2} y^{\mathsf{T}} \tilde{\Sigma}_2 y - x^{\mathsf{T}} \tilde{\Sigma} y. \tag{28}$$

According to block matrix inversion formula [41], $\tilde{\Sigma} = -\Sigma_1^{-1} \Sigma A^{-1}$ holds, where $A := \Sigma_2 - \Sigma^{\mathsf{T}} \Sigma_1^{-1} \Sigma$ is positive definite. Then, comparing the term $x^{\mathsf{T}} y$ between (24) and (28), we obtain $\Sigma_1^{-1} \Sigma A^{-1} = \frac{1}{2\lambda} I$ and:

$$2\lambda \Sigma_1^{-1} \Sigma = A = \Sigma_2 - \Sigma^{\mathsf{T}} \Sigma_1^{-1} \Sigma. \tag{29}$$

Here, $\Sigma_1^{-1} \Sigma = \Sigma^{\mathsf{T}} \Sigma_1^{-1}$ holds, because A is a symmetric matrix, and thus, we obtain:

$$\lambda \Sigma_1^{-1} \Sigma + \lambda \Sigma^{\mathsf{T}} \Sigma_1^{-1} = \Sigma_2 - \Sigma^{\mathsf{T}} \Sigma_1^{-1} \Sigma. \tag{30}$$

Completing the square of the above equation, we obtain:

$$(\Sigma_1^{-1/2} (\Sigma + \lambda I) \Sigma_1^{1/2})^{\mathsf{T}} (\Sigma_1^{-1/2} (\Sigma + \lambda I) \Sigma_1^{1/2}) = \Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} + \lambda^2 I \tag{31}$$

Let Q be an orthogonal matrix; then, (31) can be solved as:

$$\Sigma_1^{-1/2} (\Sigma + \lambda I) \Sigma_1^{1/2} = Q(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} + \lambda^2 I)^{1/2}. \tag{32}$$

We rearrange the above equation as follows:

$$\Sigma_1^{1/2} (\Sigma_1^{-1} \Sigma) \Sigma_1^{1/2} + \lambda I = Q(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} + \lambda^2 I)^{1/2}. \tag{33}$$

Because the left terms and $(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} + \lambda^2 I)^{1/2}$ are all symmetric positive definite, we can conclude that Q is the identity matrix by the uniqueness of the polar decomposition. Finally, we obtain:

$$\Sigma = \Sigma_1^{1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} + \lambda^2 I)^{1/2} \Sigma_1^{-1/2} - \lambda I =: \Sigma_\lambda. \tag{34}$$

We obtain (18) by the direct calculation of C_λ using Lemma 1 with this Σ_λ. $\square$

The following corollary helps us to understand the properties of Σ_λ.

Corollary 1. *Let $\nu_{\lambda,1} \leq \nu_{\lambda,2} \leq \dot{\leq} \nu_{\lambda,n}$ be the eigenvalues of Σ_λ; then, $\nu_{\lambda,i}$ monotonically decreases with λ for any $i \in \{1, 2, ; n\}$.*

Proof. Because $\Sigma_1^{-1/2}\Sigma_\lambda\Sigma_1^{1/2} = (\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \lambda^2 I)^{1/2} - \lambda I$ has the same eigenvalues as Σ_λ, if we let $\{\nu_{0,i}\}$ be the eigenvalues of $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$, $\nu_{\lambda,i} = \sqrt{\nu_{0,i} + \lambda^2} - \lambda$, which is a monotonically decreasing function of the regularization parameter λ. □

By the proof, for large λ, we can prove $\Sigma_1^{-1/2}\Sigma_\lambda\Sigma_1^{1/2} \simeq \frac{1}{2\lambda}\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ by diagonalization and $\nu_{\lambda,i} \simeq \frac{1}{2\lambda}\nu_{0,i}$. Thus, $\Sigma_\lambda \simeq \frac{1}{2\lambda}\Sigma_1\Sigma_2$, and each element of Σ_λ converges to zero as $\lambda \to \infty$.

We show how entropy regularization behaves in two simple experiments. We calculate the entropy-regularized optimal transport cost $\mathcal{N}\left(\binom{0}{0}, \binom{1\ \ 0}{0\ \ 1}\right)$ and $\mathcal{N}\left(\binom{0}{0}, \binom{2\ \ -1}{-1\ \ 2}\right)$ in the original version and the relative entropy version in Figure 1. We separate the entropy-regularized optimal transport cost into the transport cost term and regularization term and display both of them.
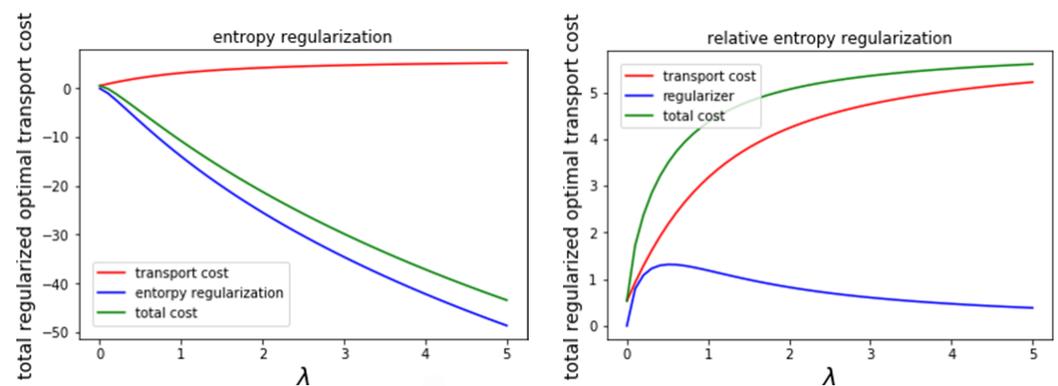

Figure 1. Graph of the entropy-regularized optimal transport cost between $\mathcal{N}\left(\binom{0}{0}, \binom{1\ \ 0}{0\ \ 1}\right)$ and $\mathcal{N}\left(\binom{0}{0}, \binom{2\ \ -1}{-1\ \ 2}\right)$ with respect to λ from zero to 10.

It is reasonable that as $\lambda \downarrow 0$, Σ_λ converges to $\Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\Sigma_1^{-1/2}$, which is equal to the original optimal coupling of nonregularized optimal transport and as $\lambda \to \infty$, Σ_λ converges to 0. This is a special case of Corollary 1. The larger λ becomes, the less correlated the optimal coupling is. We visualize this behavior by computing the optimal couplings of two one-dimensional normal distributions in Figure 2.
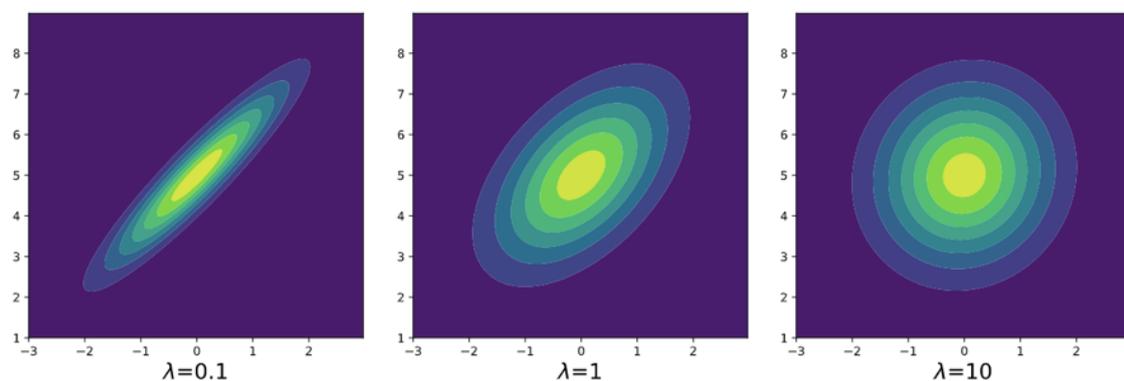

Figure 2. Contours of the density functions of the entropy-regularized optimal coupling of $\mathcal{N}(0,1)$ and $\mathcal{N}(5,2)$ in three different parameters $\lambda = 0.1, 1, 10$. All of the optimal couplings are two-variate normal distributions.

The left panel shows the original version. The transport cost is always positive, and the entropy regularization term can take both signs in general; then, the sign and total cost depend on their balance. We note that the transport cost as a function of λ is bounded, whereas the entropy regularization is not. The boundedness of the optimal cost is deduced from (1) and Corollary 1, and the unboundedness of the entropy regularization is due to the regularization parameter λ multiplied by the entropy. The right panel shows the relative

entropy version. It always takes a non-negative value. Furthermore, because the total cost is bounded by the value for the independent joint distribution (which is always a feasible coupling), both the transport cost and the relative entropy regularization regularization term are also bounded. Nevertheless, the larger the regularization parameter λ, the greater the influence of entropy regularization over the total cost.

It is known that a specific Riemannian metric can be defined in the space of multivariate normal distributions, which induces the Wasserstein distance [42]. To understand the effect of entropy regularization, we illustrate how entropy regularization deforms this geometric structure in Figure 3. Here, we generate 100 two-variate normal distributions $\{\mathcal{N}(0, \Sigma_{r,k})\}_{r,k \in \{1,2,;10\}}$, where $\{\Sigma_{r,k}\}$ is defined as:

$$\Sigma_{r,k} = \begin{pmatrix} \cos\left(2\pi \cdot \frac{k}{10}\right) & -\sin\left(2\pi \cdot \frac{k}{10}\right) \\ \sin\left(2\pi \cdot \frac{k}{10}\right) & \cos\left(2\pi \cdot \frac{k}{10}\right) \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{\frac{r}{10}} \end{pmatrix} \begin{pmatrix} \cos\left(2\pi \cdot \frac{k}{10}\right) & -\sin\left(2\pi \cdot \frac{k}{10}\right) \\ \sin\left(2\pi \cdot \frac{k}{10}\right) & \cos\left(2\pi \cdot \frac{k}{10}\right) \end{pmatrix}. \tag{35}$$

To visualize the geometric structure of these two-variate normal distributions, we compute the relative entropy-regularized optimal transport cost $\tilde{C}_\lambda$ between each pairwise two-variate normal distributions. Then, we apply multidimensional scaling [43] to embed them into a plane (see Figure 3). We can see entropy regularization deforming the geometric structure of the space of multivariate normal distributions. The deformation for distributions close to the isotopic normal distribution is more sensitive to the change in λ.
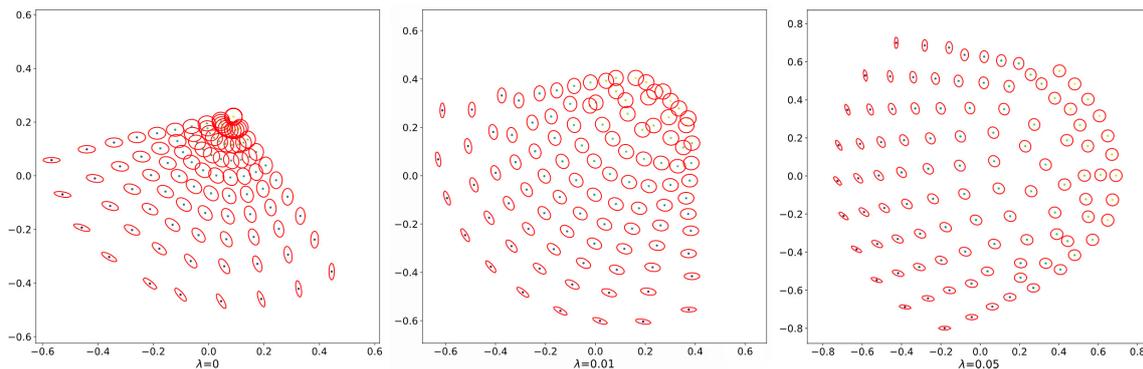

Figure 3. Multidimensional scaling of two-variate normal distributions. The pairwise dissimilarities are given by the square root of the entropy-regularized optimal transport cost $\tilde{C}_\lambda$ for three different regularization parameters $\lambda = 0, 0.01, 0.05$. Each ellipse in the figure represents a contour of the density function $\{\mathcal{N}(0, \Sigma_{r,k})\}$.

The following corollary states that if we allow orthogonal transformations of two multivariate normal distributions with fixed covariance matrices, then the minimum and maximum of C_λ are attained when Σ_1 and Σ_2 are diagonalizable by the same orthogonal matrix or, equivalently, when the ellipsoidal contours of the two density functions are aligned with the same orthogonal axes.

Corollary 2. *With the same settings as in Theorem 1, fix μ_1, μ_2, Σ_1, and all eigenvalues of Σ_2. When Σ_1 is diagonalized as $\Sigma_1 = \Gamma^{\mathrm{T}} \Lambda_1^{\downarrow} \Gamma$, where $\Lambda_1^{\downarrow}$ is the diagonal matrix of the eigenvalues of Σ_1 in descending order and Γ is an orthogonal matrix,*

(i) $C_\lambda(P, Q)$ is minimized by $\Sigma_2 = \Gamma^{\mathrm{T}} \Lambda_2^{\downarrow} \Gamma$ and

(ii) $C_\lambda(P, Q)$ is maximized by $\Sigma_2 = \Gamma^{\mathrm{T}} \Lambda_2^{\uparrow} \Gamma$,

where $\Lambda_2^{\downarrow}$ and $\Lambda_2^{\uparrow}$ are the diagonal matrices of the eigenvalues of Σ_2 in descending and ascending order, respectively. Therefore, neither the minimizer, nor the maximizer depend on the choice of λ.

Proof. Because μ_1, μ_2, Σ_1, and all eigenvalues of Σ_2 are fixed,

$$C_\lambda(P, Q) = -2\mathrm{tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \lambda^2 I)^{1/2}\right) - \frac{\lambda}{2}\log|(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \lambda^2 I)^{1/2} - \lambda I| + (\text{constant}) \quad (36)$$

$$= \sum_{i=1}^n -2(\nu_i + \lambda^2)^{1/2} - \frac{\lambda}{2}\log\{(\nu_i + \lambda^2)^{1/2} - \lambda\} + (\text{constant}) \quad (37)$$

$$= \sum_{i=1}^n g_\lambda(\log(\nu_i)) + (\text{constant}) \quad (38)$$

where $\nu_1 \leq \cdots \leq \nu_n$ are the eigenvalues of $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ and:

$$g_\lambda(x) := -2(e^x + \lambda^2)^{1/2} - \frac{\lambda}{2}\log\{(e^x + \lambda^2)^{1/2} - \lambda\}. \quad (39)$$

Note that $g_\lambda(x)$ is a concave function, because:

$$g_\lambda''(x) = -\frac{e^x(4e^x + 7\lambda^2)}{8(e^x + \lambda^2)^{3/2}} < 0. \quad (40)$$

Let $\nu_1^{\downarrow\downarrow} \leq \cdots \leq \nu_n^{\downarrow\downarrow}$ and $\nu_1^{\downarrow\uparrow} \leq \cdots \leq \nu_n^{\downarrow\uparrow}$ be the eigenvalues of $\Lambda_1^\downarrow\Lambda_2^\downarrow$ and $\Lambda_1^\downarrow\Lambda_2^\uparrow$, respectively. By Exercise 6.5.3 of [44] or Theorem 6.13 and Corollary 6.14 of [45],

$$(\log(\nu_i^{\downarrow\uparrow})) \prec (\log(\nu_i)) \prec (\log(\nu_i^{\downarrow\downarrow})), \quad (41)$$

Here, for $(a_i), (b_i) \in \mathbb{R}^n$ such that $a_1 \geq \cdots \geq a_n$ and $b_1 \geq \cdots \geq b_n$, $(a_i) \prec (b_i)$ means:

$$\sum_{i=1}^k a_i \leq \sum_{i=1}^k b_i \text{ for } k = 1, \ldots, n-1, \text{ and } \sum_{i=1}^n a_i = \sum_{i=1}^n b_i \quad (42)$$

and (a_i) is said to be majorized by (b_i). Because $g_\lambda(x)$ is concave,

$$g_\lambda(\log(\nu_i^{\downarrow\uparrow})) \prec^{\mathrm{w}} g_\lambda(\log(\nu_i)) \prec^{\mathrm{w}} g_\lambda(\log(\nu_i^{\downarrow\downarrow})), \quad (43)$$

where $\prec^{\mathrm{w}}$ represents weak supermajorization, i.e., $(a_i) \prec^{\mathrm{w}} (b_i)$ means:

$$\sum_{i=k}^n a_i \geq \sum_{i=k}^n b_i \text{ for } k = 1, \ldots, n \quad (44)$$

(see Theorem 5.A.1 of [46], for example). Therefore,

$$\sum_{i=1}^n g_\lambda(\log(\nu_i^{\downarrow\uparrow})) \geq \sum_{i=1}^n g_\lambda(\log(\nu_i)) \geq \sum_{i=1}^n g_\lambda(\log(\nu_i^{\downarrow\downarrow})). \quad (45)$$

As in Case (i) (or (ii)), the eigenvalues of $\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}$ correspond to the eigenvalues of $\Lambda_1^\downarrow\Lambda_2^\downarrow$ (or $\Lambda_1^\downarrow\Lambda_2^\uparrow$, respectively), the corollary follows. □

Note that a special case of Corollary 2 for the ordinary Wasserstein metric ($\lambda = 0$) has been studied in the context of fidelity and the Bures distance in quantum information theory. See Lemma 3 of [47]. Their proof is not directly applicable to our generalized result; thus, we used another approach to prove it.

4. Extension to Tsallis Entropy Regularization

In this section, we consider a generalization of entropy-regularized optimal transport. We now focus on the Tsallis entropy [32], which is a generalization of the Shannon entropy and appears in nonequilibrium statistical mechanics. We show that the optimal coupling of Tsallis entropy-regularized optimal transport between two q-normal distributions is also

a *q*-normal distribution. We start by recalling the definition of the *q*-exponential function and *q*-logarithmic function based on [32].

Definition 7. *Let q be a real parameter, and let u > 0. The q-logarithmic function is defined as:*

$$
\log_q(u) := \begin{cases} \frac{1}{1-q}(u^{1-q} - 1) & \text{if } q \neq 1, \\ \log(u) & \text{if } q = 1 \end{cases}
\tag{46}
$$

and the q-exponential function is defined as:

$$
\exp_q(u) := \begin{cases} [1 + (1-q)u]_+^{\frac{1}{1-q}} & \text{if } q \neq 1, \\ \exp(u) & \text{if } q = 1 \end{cases}
\tag{47}
$$

Definition 8. *Let $q < 1$ or $1 < q < 1 + \frac{2}{n}$; an n-variate q-normal distribution is defined by two parameters: $\mu \in \mathbb{R}^n$ and a positive definite matrix Σ, and its density function is:*

$$
f(x) := \frac{1}{C_q(\Sigma)} \exp_q\left(-(x-\mu)^{\mathrm{T}} \Sigma^{-1}(x-\mu)\right),
\tag{48}
$$

where $C_q(\Sigma)$ is a normalizing constant. μ and Σ are called the location vector and scale matrix, respectively.

In the following, we write the multivariate *q*-normal distribution $\mathcal{N}_q(\mu, \Sigma)$. We note that the property of the *q*-normal distribution changes in accordance with *q*. The *q*-normal distribution has an unbounded support for $1 < q < \frac{2}{n}$ and a bounded support for $q < 1$. The second moment exists for $q < 1 + \frac{2}{n+2}$, and the covariance becomes $\frac{1}{2+(n+2)(1-q)}\Sigma$. We remark that each *n*-variate $\left(1 + \frac{2}{\nu+n}\right)$-normal distribution is equivalent to an *n*-variate *t*-distribution with ν degrees of freedom,

$$
\frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}|\Sigma|^{1/2}} \left[1 + \frac{1}{\nu}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]^{-(\nu+n)/2},
\tag{49}
$$

for $1 < q < 1 + \frac{2}{n+2}$ and an *n*-variate normal distribution for $q \downarrow 1$.

Definition 9. *Let p be a probability density function. The Tsallis entropy is defined as:*

$$
S_q(p) := \int p(x) \log_q \frac{1}{p(x)} dx = \frac{1}{q-1}\left(1 - \int p(x)^q dx\right).
\tag{50}
$$

Then, the Tsallis entropy-regularized optimal transport is defined as:

$$
\text{minimize } \int \|x-y\|^2 \pi(x,y) dx dy - 2\lambda S_q(\pi)
\tag{51}
$$

$$
\text{subject to } \int \pi(x,y) dx = q(y) \text{ for } \forall\, y \in \mathbb{R}^n,
$$

$$
\int \pi(x,y) dy = p(x) \text{ for } \forall\, x \in \mathbb{R}^n.
\tag{52}
$$

The following lemma is a generalization of the maximum entropy principle for the Shannon entropy shown in Section 2 of [48].

Lemma 3. *Let P be a centered n-dimensional probability measure with a fixed covariance matrix Σ; the maximizer of the Renyi α-entropy:*

$$
\frac{1}{1-\alpha} \log \int f(x)^\alpha \mathrm{d}x
\tag{53}
$$

under the constraint is $\mathcal{N}_{2-\alpha}(\mathbf{0}, ((n+2)\alpha - n)\Sigma)$ for $\frac{n}{n+2} < \alpha < 1$.

We note that the maximizers of the Renyi α-entropy and the Tsallis entropy with $q = \alpha$ coincide; thus, the above lemma also holds for the Tsallis entropy. This is mentioned, for example, in Section 9 of [49].

To prove Theorem 2, we use the following property of multivariate t-distributions, which is summarized in Chapter 1 of [50].

Lemma 4. *Let X be a random vector following an n-variate t-distribution with degree of freedom ν. Considering a partition of the mean vector μ and scale matrix Σ, such as:*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{54}$$

X_1 follows a p-variate t-distribution with degree of freedom ν, mean vector μ_1, and scale matrix Σ_{11}, where p is the dimension of X_1.

Recalling the correspondence of the parameter of the multivariate q-normal distribution and the degree of freedom of the multivariate t-distribution $q = 1 + \frac{2}{\nu+n}$, we can obtain the following corollary.

Corollary 3. *Let X be a random vector following an n-variate q-normal distribution for $1 < q < 1 + \frac{2}{n+2}$. Consider a partition of the mean vector μ and scale matrix Σ in the same way as in (54). Then, X_1 follows a p-variate $\left(1 + \frac{2(q-1)}{2-(n-p)(q-1)}\right)$-normal distribution with mean vector μ_1 and scale matrix Σ_{11}, where p is the dimension of X_1.*

Theorem 2. *Let $P \sim \mathcal{N}_q(\mu_1, \Sigma_1), Q \sim \mathcal{N}_q(\mu_2, \Sigma_2)$ be n-variate q-normal distributions for $1 < q < 1 + \frac{2}{n+2}$ and $\tilde{q} = -\frac{2(q-1)}{2-n(q-1)}$; consider the Tsallis entropy-regularized optimal transport:*

$$C_\lambda(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^2 \pi(x, y) dx dy - 2\lambda S_{1+\tilde{q}}(\pi). \tag{55}$$

Then, there exists a unique $\tilde{\lambda} = \tilde{\lambda}(q, \Sigma_1, \Sigma_2, \lambda) \in \mathbb{R}_+$ such that the optimal coupling π of the entropy-regularized optimal transport is expressed as:

$$\pi \sim \mathcal{N}_{1-\tilde{q}}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{\tilde{\lambda}} \\ \Sigma_{\tilde{\lambda}}^{\mathrm{T}} & \Sigma_2 \end{pmatrix}\right), \tag{56}$$

where:

$$\Sigma_{\tilde{\lambda}} := \Sigma_1^{1/2}(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2} + \tilde{\lambda}^2 I)^{1/2}\Sigma_1^{-1/2} - \tilde{\lambda}I. \tag{57}$$

Proof. The proof proceeds in a similar way as in Theorem 1. Let $\alpha \in L(P)$ and $\beta \in L(Q)$ be the Lagrangian multipliers. Then, the Lagrangian function $L(\pi, \alpha, \beta)$ of (52) is defined as:

$$L(\pi, \alpha, \beta) := \int \|x - y\|^2 \pi(x, y) dx dy - 2\lambda \left\{ \frac{1}{\tilde{q}}\left(1 - \int \pi(x, y)^{1+\tilde{q}} dx dy\right)\right\}$$
$$- \int \alpha(x)\pi(x, y) dx dy + \int \alpha(x)p(x) dx$$
$$- \int \beta(y)\pi(x, y) dx dy + \int \beta(y)q(y) dy \tag{58}$$

and the extremum of the Tsallis entropy-regularized optimal transport is obtained by the functional derivative with respect to π,

$$\pi(x,y) = \left(\frac{\tilde{q}}{2(\tilde{q}+1)\lambda} \left(-\alpha(x) - \beta(y) + \|x-y\|^2 \right) \right)^{\frac{1}{\tilde{q}}}. \tag{59}$$

Here, α and β are quadratic polynomials by Lemma 3. To separate the normalizing constant, we introduce a constant $c \in \mathbb{R}_+$, and π can be written as:

$$\pi(x,y) = c^{\frac{1}{\tilde{q}}} \left(\tilde{\alpha}(x) + \tilde{\beta}(y) + \frac{\tilde{q}\|x-y\|^2}{2c(\tilde{q}+1)\lambda} \right)^{\frac{1}{\tilde{q}}}, \tag{60}$$

with quadratic functions $\tilde{\alpha}(x)$ and $\tilde{\beta}(y)$.

Let $\tilde{\lambda} = \frac{c(\tilde{q}+1)\lambda}{\tilde{q}} > 0$. Then, by the same argument as in the proof of Theorem 1 and using Corollary 3, we obtain the scale matrix of π as:

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{\tilde{\lambda}} \\ \Sigma_{\tilde{\lambda}}^{\mathrm{T}} & \Sigma_2 \end{pmatrix}, \tag{61}$$

where:

$$\Sigma_{\tilde{\lambda}} = \Sigma_1^{1/2} (\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} + \tilde{\lambda}^2 I)^{1/2} \Sigma_1^{-1/2} - \tilde{\lambda} I. \tag{62}$$

Let $z = (x^{\mathrm{T}}, y^{\mathrm{T}})^{\mathrm{T}}$ and $K_{\tilde{q}} = \int (1 + z^{\mathrm{T}} z)^{\frac{1}{\tilde{q}}} dz$; π can be written as:

$$\pi(x,y) = \frac{1}{K_{\tilde{q}}|\Sigma|} (1 + z^{\mathrm{T}} \Sigma^{-1} z)^{\frac{1}{\tilde{q}}}. \tag{63}$$

The constant c is determined by:

$$\frac{1}{K_{\tilde{q}}|\Sigma|} = c^{\frac{1}{\tilde{q}}}. \tag{64}$$

We will show that the above equation has a unique solution. Let $\{\tau\}_{i=1}^n$ be the eigenvalues of $(\Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2})^{1/2}$; $|\Sigma|$ can be expressed as $\prod_{i=1}^{2n} 2\tilde{\lambda}(\sqrt{\tau_i^2 + \tilde{\lambda}^2} - \tilde{\lambda})$. We consider:

$$f(c) = \log(c^{\frac{1}{\tilde{q}}} K_{\tilde{q}} |\Sigma|) \tag{65}$$

$$= \frac{1}{\tilde{q}} \log c + \sum_{i=1}^{2n} \log(\sqrt{\tau_i^2 + \tilde{\lambda}^2} - \tilde{\lambda}) + 2n \log(2\tilde{\lambda}) + \log K_{\tilde{q}}. \tag{66}$$

Because $\tilde{q} < 0$, $f(c)$ is a monotonic decreasing function, and $\lim_{c\downarrow 0} f(c) = \infty$, $\lim_{c\to\infty} f(c) = -\infty$, (64) has a unique positive solution, and $\tilde{\lambda}$ is determined uniquely. $\square$

5. Entropy-Regularized Kantorovich Estimator

Many estimators are defined by minimizing the divergence or distance ρ between probability measures, that is $\arg\min_\mu \rho(\mu,\nu)$ for a fixed ν. When ρ is the Kullback–Leibler divergence, the estimator corresponds to the maximum likelihood estimator. When ρ is the Wasserstein distance, the following estimator is called the minimum Kantorovich estimator, according to [36]. In this section, we consider a probability measure Q^* that minimizes $C_\lambda(P,Q)$ for a fixed P over $\mathcal{P}_2(\mathbb{R}^n)$, the set of all probability measures on $\mathbb{R}^n$ with finite second moment that are absolutely continuous with respect to the Lebesgue measure. In other words, we define the entropy-regularized Kantorovich estimator $\arg\min_{Q\in\mathcal{P}_2(\mathbb{R}^n)} C_\lambda(P,Q)$. The entropy-regularized Kantorovich estimator for discrete probability measures was studied in [33], Theorem 2. We obtain the entropy-regularized Kantorovich estimator for continuous probability measures in the following theorem:

Theorem 3. *For a fixed* $P \in \mathcal{P}_2(\mathbb{R}^n)$,

$$Q^* = \arg \min_{Q \in \mathcal{P}_2(\mathbb{R}^n)} C_\lambda(P, Q) \tag{67}$$

exists, and its density function can be written as:

$$dQ^* = dP \star \phi_\lambda, \tag{68}$$

where $\phi_\lambda(x)$ *is a density function of* $\mathcal{N}(0, \frac{\lambda}{2}I)$, *and* $\star$ *denotes the convolution operator.*

We use the dual problem of the entropy-regularized optimal transport to prove Theorem 3 (for details, see Proposition 2.1 of [15] or Section 3 of [51]).

Lemma 5. *The dual problem of entropy-regularized optimal transport can be written as:*

$$\mathcal{A}_\lambda(P, Q) = \sup_{\substack{\alpha \in L_1(P) \\ \beta \in L_1(Q)}} \int \alpha(x)p(x)dx + \int \beta(y)q(y)dy$$

$$- \lambda \int \exp\left\{ \frac{\alpha(x) + \beta(y) - \|x - y\|^2}{\lambda} \right\} dxdy. \tag{69}$$

Moreover, $\mathcal{A}_\lambda(P, Q) = C_\lambda(P, Q)$ *holds.*

Now, we prove Theorem 3.

Proof. Let Q^* be the minimizer of $\min_Q C_\lambda(P, Q)$. Applying Lemma 5, there exist $\alpha^* \in L_1(P)$ and $\beta^* \in L_1(Q^*)$ such that:

$$C_\lambda(P, Q^*) = \mathcal{A}_\lambda(P, Q^*) = \int \alpha^*(x)p(x)dx + \int \beta^*(y)q^*(y)dy$$

$$- \lambda \int \exp\left\{ \frac{\alpha^*(x) + \beta^*(y) - \|x - y\|^2}{\lambda} \right\} dxdy. \tag{70}$$

Now, $\mathcal{A}_\lambda(P, Q^*)$ is the minimum value of $\mathcal{A}_\lambda$, such that the variation $\delta \mathcal{A}_\lambda(P, Q^*)$ is always zero. Then,

$$\delta \mathcal{A}_\lambda(P, Q^*) = \int \beta^*(y)\delta q^*(y)dy = 0 \Rightarrow \beta^* \equiv 0 \tag{71}$$

holds, and the optimal coupling of P, Q can be written as:

$$\pi^*(x, y) = \exp\left\{ \frac{\alpha^*(x) + \beta^*(y)}{\lambda} - \frac{\|x - y\|^2}{\lambda} \right\} \tag{72}$$

$$= \exp\left\{ \frac{\alpha^*(x)}{\lambda} \right\} \exp\left\{ -\frac{\|x - y\|^2}{\lambda} \right\}. \tag{73}$$

Moreover, we can obtain a closed-form of $\alpha^*(x)$ as follows from the equation $\int \pi(x, y)dy = p(x)$:

$$\frac{\alpha^*(x)}{\lambda} = \log p(x) - \log \int \exp\left\{ -\frac{\|x - y\|^2}{\lambda} \right\} dy = \log p(x) - \frac{n}{2} \log(\pi\lambda). \tag{74}$$

Then, by calculating the marginal distribution of $\pi(x, y)$ with respect to x, we can obtain:

$$q^*(y) = \int \frac{1}{(\pi\lambda)^{\frac{n}{2}}} \exp\left\{ -\frac{\|x - y\|^2}{\lambda} \right\} p(x)dx = (p \star \phi_\lambda)(y). \tag{75}$$

Therefore, we conclude that a probability measure Q that minimizes $C_\lambda(P, Q)$ is expressed as (75). $\quad\square$

It should be noted that when P in Theorem 3 are multivariate normal distributions, Q^* and P are simultaneously diagonalizable by a direct consequence of the theorem. This is consistent with the result of Corollary 2(1) for minimization when all eigenvalues are fixed.

We can determine that the entropy-regularized Kantorovich estimator is a measure convolved with an isotropic multivariate normal distribution scaled by the regularization parameter λ. This is similar to the idea of prior distributions in the context of Bayesian inference. Applying Theorem 3, the entropy-regularized Kantorovich estimator of the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ is $\mathcal{N}(\mu, \Sigma + \frac{\lambda}{2}I)$.

6. Numerical Experiments

In this section, we introduce experiments that show the statistical efficiency of entropy regularization in Gaussian settings. We consider two different setups, estimating covariance matrices (Section 6.1) and the entropy-regularized Wasserstein barycenter (Section 6.2). To obtain the entropy-regularized Wasserstein barycenter, we adopt the Newton–Schulz method and a manifold optimization method, which are explained in Sections 6.3 and 6.4, respectively.

6.1. Estimation of Covariance Matrices

We provide a covariance estimation method based on entropy-regularized optimal transport. Let $P = \mathcal{N}(\mu, \Sigma)$ be an n-variate normal distribution. We define an entropy-regularized Kantorovich estimator $\hat{P}_\lambda$, that is,

$$\hat{P}_\lambda = \arg\min_Q C_\lambda(P, Q). \tag{76}$$

We generate some samples from $\mathcal{N}(\mu, \Sigma)$ and estimate the mean and covariance matrix. We compare the maximum likelihood estimator $\hat{P}_{\mathrm{MLE}} = \mathcal{N}(\hat{\mu}_{\mathrm{MLE}}, \hat{\Sigma}_{\mathrm{MLE}})$ and $\hat{P}_\lambda$ with respect to the prediction error:

$$\mathrm{KL}(P, \hat{P}_{\mathrm{MLE}}), \quad \mathrm{KL}(P, \hat{P}_\lambda). \tag{77}$$

In our experiment, the dimension n is set to $5, 15, 30$, and the sample size is set to $60, 120$. The experiment proceeds as follows.

1. Obtain a random sample of size 60 (or 120) from $\mathcal{N}(0, \Sigma)$ and its sample covariance matrix $\hat{\Sigma}$.
2. Obtain the entropy-regularized minimum Kantorovich estimator of $\hat{\Sigma}$ obtained in Step 1.
3. Compute the prediction error between Σ and the entropy-regularized minimum Kantorovich estimator of $\hat{\Sigma}$
4. Repeat the above steps 1000 times and obtain a confidence interval of the prediction error.

Table 1 shows the average prediction error of the MLE and entropy-regularized Kantorovich estimator of covariance matrices from 60 samples from an n-variate normal distribution with the 95% confidential interval. We can see that the prediction error is smaller than the maximum likelihood estimator under adequately small λ for $n = 15, 30$, but not for $n = 5$. Moreover, the decrease in the prediction error is larger for $n = 30$ than for $n = 15$, which indicates that the entropy regularization is effective in a high dimension. On the other hand, Table 2 shows in all cases that the decreases in the prediction error are more moderate than Table 1. We can see that this is due to the increase in the sample size. Then, we can conclude that the entropy regularization is effective in a high-dimensional setting with a small sample size.

Table 1. Average prediction error of the MLE and entropy-regularized Kantorovich estimator of covariance matrices from 60 samples from an n-variate normal distribution with the 95% confidential interval.

λ	$\mathbf{KL}(P, \hat{P}_\mathbf{W}), n = 5$	$\mathbf{KL}(P, \hat{P}_\mathbf{W}), n = 15$	$\mathbf{KL}(P, \hat{P}_\mathbf{W}), n = 30$
0(MLE)	0.062 ± 0.005	1.346 ± 0.022	10.69 ± 0.112
0.01	0.051 ± 0.005	1.242 ± 0.021	8.973 ± 0.087
0.1	0.104 ± 0.004	0.841 ± 0.013	4.180 ± 0.033
0.5	0.647 ± 0.003	0.931 ± 0.007	3.093 ± 0.010
1.0	1.166 ± 0.003	1.670 ± 0.006	5.075 ± 0.009

Table 2. Average prediction error of the MLE and entropy-regularized Kantorovich estimator of covariance matrices from 120 samples from an n-variate normal distribution with the 95% confidential interval.

λ	$\mathbf{KL}(P, \hat{P}_\mathbf{W}), n = 5$	$\mathbf{KL}(P, \hat{P}_\mathbf{W}), n = 15$	$\mathbf{KL}(P, \hat{P}_\mathbf{W}), n = 30$
0(MLE)	0.024 ± 0.002	0.490 ± 0.007	2.810 ± 0.021
0.01	0.020 ± 0.002	0.459 ± 0.006	2.528 ± 0.018
0.1	0.101 ± 0.002	0.397 ± 0.005	1.700 ± 0.001
0.5	0.659 ± 0.002	0.875 ± 0.004	2.833 ± 0.005
1.0	1.180 ± 0.002	1.730 ± 0.004	5.124 ± 0.005

6.2. Estimation of the Wasserstein Barycenter

A barycenter with respect to the Wasserstein distance is definable [52] and is widely used for image interpolation and 3D object interpolation tasks with entropy regularization [3,33].

Definition 10. *Let $\{Q_i\}_{i=1}^m$ be a set of probability measures in $\mathcal{P}(\mathbb{R}^n)$. The barycenter with respect to C_λ (entropy-regularized Wasserstein barycenter) is defined as:*

$$\arg \min_{P \in \mathcal{P}(\mathbb{R}^n)} \sum_{i=1}^m C_\lambda(P, Q_i). \tag{78}$$

Now, we restrict P and $\{Q_i\}_{i=1}^m$ to be multivariate normal distributions and apply our theorem to illustrate the effect of entropy regularization.

The experiment proceeds as follows. The dimensionality and the sample size were set the same as in the experiments in Section 6.1.

1.　Obtain a random sample of size 60 (or 120) from $\mathcal{N}(0, \Sigma)$ and its sample covariance matrix $\hat{\Sigma}$.
2.　Repeat Step 1 three times, and obtain $\{\hat{\Sigma}\}_{i=1}^3$.
3.　Obtain the barycenter of $\{\hat{\Sigma}_i\}_{i=1}^3$.
4.　Compute the prediction error between Σ and the barycenter obtained in step 3.
5.　Repeat the above steps 100 times and obtain a confidence interval of the prediction error.

We show the results for several values of the regularization parameter λ in Tables 3 and 4. A decrease in the prediction error can be seen in Table 3 for $n = 30$, as well as Tables 1 and 2. However, because the computation of the entropy-regularized Wasserstein barycenter uses more data than that of the minimum Kantorovich estimator, the decrease in the prediction error is mild. The entropy-regularized Kantorovich estimator is a special case of the entropy-regularized Wasserstein barycenter (78) for $m = 1$. Our experiments show that the appropriate range of λ to decrease the prediction error depends on m and becomes narrow as m increases. In addition, we note that there is a small decrease in the prediction error in Table 4 for $n = 30$.

Table 3. Average prediction error of the entropy-regularized barycenter with the 95% confidential interval (random sample of size 60).

λ	$\text{KL}(P, \hat{P}_W), n = 5$	$\text{KL}(P, \hat{P}_W), n = 15$	$\text{KL}(P, \hat{P}_W), n = 30$
0	0.455 ± 0.395	1.318 ± 0.006	4.875 ± 0.035
0.001	0.429 ± 0.396	1.318 ± 0.004	4.887 ± 0.036
0.01	0.434 ± 0.270	1.344 ± 0.006	4.551 ± 0.164
0.025	0.780 ± 0.223	1.456 ± 0.064	5.710 ± 0.536
0.005	1.047 ± 0.029	1.537 ± 0.064	7.570 ± 0.772

Table 4. Average prediction error of the entropy-regularized barycenter with the 95% confidential interval (random sample of size 120).

λ	$\text{KL}(P, \hat{P}_W), n = 5$	$\text{KL}(P, \hat{P}_W), n = 15$	$\text{KL}(P, \hat{P}_W), n = 30$
0	0.154 ± 0.600	1.303 ± 0.010	5.091 ± 0.035
0.001	0.212 ± 0.070	1.305 ± 0.010	5.072 ± 0.037
0.01	0.306 ± 0.046	1.328 ± 0.008	5.274 ± 0.252
0.025	0.671 ± 0.028	1.337 ± 0.073	5.851 ± 0.424
0.005	1.109 ± 0.063	1.603 ± 0.184	8.072 ± 0.725

6.3. Gradient Descent on $\text{Sym}_+(n)$

We use a gradient descent method to compute the entropy-regularized barycenter. Applying the gradient descent method to the loss function defined by the Wasserstein distance was proposed in [4]. This idea is extendable to entropy-regularized optimal transport. The detailed algorithm is shown below. Because $C_\lambda(P, Q)$ is a function of a positive definite matrix, we used a manifold gradient descent algorithm on the manifold of positive definite matrices.

We review the manifold gradient descent algorithm used in our numerical experiment. Let $\text{Sym}_+(n)$ be the manifold of n-dimensional positive definite matrices. We require a formula for a gradient operator and the inner product of $\text{Sym}_+(n)$ in the gradient descent algorithm. In this paper, we use the following inner product from [44], Chapter 6. For a fixed $X \in \text{int}(\text{Sym}_+(n))$, we define an inner product of $\text{Sym}_+(n)$ as:

$$g_X(Y, Z) = \text{tr}\left(YX^{-1}ZX^{-1}\right), \ Y, Z \in \text{Sym}_+(n), \tag{79}$$

Equation (79) is the best choice in terms of the convergence speed according to [53]. Let $f : \text{Sym}_+(n) \to \mathbb{R}$ be a differential matrix function. Then, the induced gradient of f under (79) is:

$$\text{grad}\, f(X) = X\left(\frac{\partial f(X)}{\partial X}\right)X. \tag{80}$$

We consider the updating step after obtaining the gradient of f. $\text{grad}f(X)$ is an element of the tangent space, and we have to project it to $\text{Sym}_+(n)$. This projection map is called a retraction. It is known that the Riemannian metric g_X leads to the following retraction:
$\exp_X x = X \text{Exp}(X^{-1}x)$, where Exp is the matrix exponential. Then, the corresponding gradient descent method becomes as shown in Algorithm 1.

6.4. Approximate the Matrix Square Root

To compute the gradient of the square root of a matrix in the objective function, we approximate it using the Newton–Schulz method [54], which can be implemented by matrix operations as shown in Algorithm 2. It is amenable to automatic differentiation, such that we can easily apply the gradient descent method to our algorithm.

Algorithm 1 Gradient descent on the manifold of positive definite matrices.

Input: $f(X)$
    initialize X
    while no convergence **do**
      η : step size
      $\text{grad} \leftarrow X\left(\frac{\partial f(X)}{\partial X}\right)X$
      $X \leftarrow \exp_X(\eta\text{grad}) = X\text{Exp}(\eta X^{-1}\text{grad})$
    end while
Output: X

Algorithm 2 Newton–Schulz method.

Input: $A \in \text{Sym}_+(n), \epsilon > 0$
    $Y \leftarrow \frac{A}{(1+\epsilon)\|A\|}, \;\; Z \leftarrow I$
    while no convergence **do**
      $T \leftarrow (3I - ZY)/2$
      $Y \leftarrow YT, \; Z \leftarrow TZ$
    end while
Output: $\sqrt{(1+\epsilon)\|A\|}\,Y$

---

7. Conclusions and Future Work

In this paper, we studied entropy-regularized optimal transport and derived several result. We summarize these as follows and add notes on future work.

- We obtain the explicit form of entropy-regularized optimal transport between two multivariate normal distributions and derived Corollaries 1 and 2, which clarified the properties of optimal coupling. Furthermore, we demonstrate experimentally how entropy regularization affects the Wasserstein distance, the optimal coupling, and the geometric structure of multivariate normal distributions. Overall, the properties of optimal coupling were revealed both theoretically and experimentally. We expect that the explicit formula can be a replacement for the existing methodology using the (non-regularized) Wasserstein distance between normal distributions (for example, [4,5]).
- Theorem 2 derives the explicit form of the optimal coupling of the Tsallis entropy-regularized optimal transport between multivariate q-normal distributions. The optimal coupling of the Tsallis entropy-regularized optimal transport between multivariate q-normal distributions is also a multivariate q-normal distribution, and the obtained result has an analogy to that of the normal distribution. We believe that this result can be extended to other elliptical distribution families.
- The entropy-regularized Kantorovich estimator of a probability measure in $\mathcal{P}_2(\mathbb{R})$ is the convolution of a multivariate normal distribution and its own density function. Our experiments show that both the entropy-regularized Kantorovich estimator and the Wasserstein barycenter of multivariate normal distributions outperform the maximum likelihood estimator in the prediction error for adequately selected λ in a high dimensionality and small sample setting. As future work, we want to show the efficiency of entropy regularization using real data.

Author Contributions: Conceptualization, Q.T.; methodology, Q.T.; software, Q.T.; writing—original draft preparation, Q.T.; writing—review and editing, K.K.; supervision, K.K. Both authors read and agreed to the published version of the manuscript.

Data Availability Statement: All the data used are artificial and generated by pseudo-random numbers.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008; Volume 338.
2. Rubner, Y.; Tomasi, C.; Guibas, L.J. The Earth Mover's Distance as a metric for Image Retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121. [CrossRef]
3. Solomon, J.; De Goes, F.; Peyré, G.; Cuturi, M.; Butscher, A.; Nguyen, A.; Du, T.; Guibas, L. Convolutional Wasserstein Distances: Efficient optimal transportation on geometric domains. *ACM Trans. Graph. TOG* **2015**, *34*, 1–11. [CrossRef]
4. Muzellec, B.; Cuturi, M. Generalizing point embeddings using the Wasserstein space of elliptical distributions. In *Advances in Neural Information Processing Systems*; ACM: New York, NY, USA, 2018; pp. 10237–10248.
5. 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*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
6. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 214–223.
7. Nitanda, A.; Suzuki, T. Gradient layer: Enhancing the convergence of adversarial training for generative models. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, Playa Blanca, Spain, 9–11 April 2018; Volume 84, pp. 1008–1016.
8. Sonoda, S.; Murata, N. Transportation analysis of denoising autoencoders: A novel method for analyzing deep neural networks. *arXiv* **2017**, arXiv:1712.04145.
9. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
10. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*; ACM: New York, NY, USA, 2013; pp. 2292–2300.
11. Sinkhorn, R.; Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pac. J. Math.* **1967**, *21*, 343–348. [CrossRef]
12. Lin, T.; Ho, N.; Jordan, M.I. On the efficiency of the Sinkhorn and Greenkhorn algorithms and their acceleration for optimal transport. *arXiv* **2019**, arXiv:1906.01437.
13. Lee, Y.T.; Sidford, A. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science, Philadelphia, PA, USA, 18–21 October 2014; pp. 424–433.
14. Dvurechensky, P.; Gasnikov, A.; Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1367–1376.
15. Aude, G.; Cuturi, M.; Peyré, G.; Bach, F. Stochastic optimization for large-scale optimal transport. *arXiv* **2016**, arXiv:1605.08527.
16. Altschuler, J.; Weed, J.; Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *arXiv* **2017**, arXiv:1705.09634.
17. Blondel, M.; Seguy, V.; Rolet, A. Smooth and sparse optimal transport. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Playa Blanca, Spain, 9–11 April 2018; pp. 880–889.
18. Cuturi, M.; Peyré, G. A smoothed dual approach for variational Wasserstein problems. *SIAM J. Imaging Sci.* **2016**, *9*, 320–343. [CrossRef]
19. Lin, T.; Ho, N.; Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 3982–3991.
20. Allen-Zhu, Z.; Li, Y.; Oliveira, R.; Wigderson, A. Much faster algorithms for matrix scaling. In Proceedings of the 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, USA, 15–17 October 2017; pp. 890–901.
21. Cohen, M.B.; Madry, A.; Tsipras, D.; Vladu, A. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In Proceedings of the 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, USA, 15–17 October 2017; pp. 902–913.
22. Blanchet, J.; Jambulapati, A.; Kent, C.; Sidford, A. Towards optimal running times for optimal transport. *arXiv* **2018**, arXiv:1810.07717.
23. Quanrud, K. Approximating optimal transport with linear programs. *arXiv* **2018**, arXiv:1810.05957.
24. Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; Poggio, T.A. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*; ACM: New York, NY, USA, 2015; pp. 2053–2061.
25. Courty, N.; Flamary, R.; Tuia, D.; Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1853–1865. [CrossRef]
26. Lei, J. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli* **2020**, *26*, 767–798. [CrossRef]
27. Mena, G.; Niles-Weed, J. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*; ACM: New York, NY, USA, 2019; pp. 4543–4553.
28. Rigollet, P.; Weed, J. Entropic optimal transport is maximum-likelihood deconvolution. *Comptes Rendus Math.* **2018**, *356*, 1228–1235. [CrossRef]

29. Balaji, Y.; Hassani, H.; Chellappa, R.; Feizi, S. Entropic GANs meet VAEs: A Statistical Approach to Compute Sample Likelihoods in GANs. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 414–423.

30. Amari, S.I.; Karakida, R.; Oizumi, M. Information geometry connecting Wasserstein distance and Kullback–Leibler divergence via the entropy-relaxed transportation problem. *Inf. Geom.* **2018**, *1*, 13–37. [CrossRef]

31. Dowson, D.; Landau, B. The Fréchet distance between multivariate normal distributions. *J. Multivar. Anal.* **1982**, *12*, 450–455. [CrossRef]

32. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]

33. Amari, S.I.; Karakida, R.; Oizumi, M.; Cuturi, M. Information geometry for regularized optimal transport and barycenters of patterns. *Neural Comput.* **2019**, *31*, 827–848. [CrossRef]

34. Janati, H.; Muzellec, B.; Peyré, G.; Cuturi, M. Entropic optimal transport between (unbalanced) Gaussian measures has a closed form. In *Advances in Neural Information Processing Systems*; ACM: New York, NY, USA, 2020.

35. Mallasto, A.; Gerolin, A.; Minh, H.Q. Entropy-Regularized 2-Wasserstein Distance between Gaussian Measures. *arXiv* **2020**, arXiv:2006.03416.

36. Peyré, G.; Cuturi, M. Computational optimal transport. *Found. Trends® Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]

37. Monge, G. *Mémoire sur la Théorie des Déblais et des Remblais*; Histoire de l'Académie Royale des Sciences de Paris: Paris, France, 1781.

38. Kantorovich, L.V. On the translocation of masses. *Proc. USSR Acad. Sci.* **1942**, *37*, 199–201. [CrossRef]

39. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [CrossRef]

40. Mardia, K.V. Characterizations of directional distributions. In *A Modern Course on Statistical Distributions in Scientific Work*; Springer: Berlin/Heidelberg, Germany, 1975; pp. 365–385.

41. Petersen, K.; Pedersen, M. *The Matrix Cookbook*; Technical University of Denmark: Lyngby, Denmark, 2008; Volume 15.

42. Takatsu, A. Wasserstein geometry of Gaussian measures. *Osaka J. Math.* **2011**, *48*, 1005–1026.

43. Kruskal, J.B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **1964**, *29*, 115–129. [CrossRef]

44. Bhatia, R. *Positive Definite Matrices*; Princeton University Press: Princeton, NJ, USA, 2009; Volume 24.

45. Hiai, F.; Petz, D. *Introduction to Matrix Analysis and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2014.

46. Marshall, A.W.; Olkin, I.; Arnold, B.C. *Inequalities: Theory of Majorization and Its Applications*; Springer: Berlin/Heidelberg, Germany, 1979; Volume 143.

47. Markham, D.; Miszczak, J.A.; Puchała, Z.; Życzkowski, K. Quantum state discrimination: A geometric approach. *Phys. Rev. A* **2008**, *77*, 42–111. [CrossRef]

48. Costa, J.; Hero, A.; Vignat, C. On solutions to multivariate maximum *α*-entropy problems. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 211–226.

49. Naudts, J. *Generalised Thermostatistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.

50. Kotz, S.; Nadarajah, S. *Multivariate t-Distributions and their Applications*; Cambridge University Press: Cambridge, UK, 2004.

51. Clason, C.; Lorenz, D.A.; Mahler, H.; Wirth, B. Entropic regularization of continuous optimal transport problems. *J. Math. Anal. Appl.* **2021**, *494*, 124432. [CrossRef]

52. Agueh, M.; Carlier, G. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **2011**, *43*, 904–924. [CrossRef]

53. Jeuris, B.; Vandebril, R.; Vandereycken, B. A survey and comparison of contemporary algorithms for computing the matrix geometric mean. *Electron. Trans. Numer. Anal.* **2012**, *39*, 379–402.

54. Higham, N.J. Newton's method for the matrix square root. *Math. Comput.* **1986**, *46*, 537–549.