Entropic Gromov-Wasserstein between Gaussian Distributions

Khang Le [*1] **Dung Le** [*2] **Huy Nguyen** [*3] **Dat Do** [4] **Tung Pham** [**3] **Nhat Ho** [**1]

Abstract

We study the entropic Gromov-Wasserstein and its unbalanced version between (unbalanced) Gaussian distributions with different dimensions. When the metric is the inner product, which we refer to as inner product Gromov-Wasserstein (IGW), we demonstrate that the optimal transportation plans of entropic IGW and its unbalanced variant are (unbalanced) Gaussian distributions. Via an application of von Neumann's trace inequality, we obtain closed-form expressions for the entropic IGW between these Gaussian distributions. Finally, we consider an entropic inner product Gromov-Wasserstein barycenter of multiple Gaussian distributions. We prove that the barycenter is a Gaussian distribution when the entropic regularization parameter is small. We further derive a closed-form expression for the covariance matrix of the barycenter.

1. Introduction

The recent advance in computation of optimal transport (Cuturi, 2013; Bonneel et al., 2015; Kolouri et al., 2016; Altschuler et al., 2017; Dvurechensky et al., 2018; Lin et al., 2019; Deshpande et al., 2019; Nguyen et al., 2021a) has led to a surge of interest in using optimal transport for applications in machine learning and statistics. These applications include generative modeling (Arjovsky et al., 2017; Tolstikhin et al., 2018; Salimans et al., 2018; Genevay et al., 2018; Liutkus et al., 2019), unsupervised learning (Ho et al., 2017; 2019), domain adaptation (Courty et al., 2016; Damodaran et al., 2018; Le et al., 2021), mini-batch computation (Fatras et al., 2020; Nguyen et al., 2021b), and other applications (Solomon et al., 2016; Nguyen et al., 2021c). In these above applications, optimal transport has been used to quantify the discrepancy of the locations and mass between two probability measures which must be in the same space.

When probability measures lie in different spaces, their locations are not comparable, their inner structures become the main concern. In this case, the Gromov-Wasserstein distance is adopted to measure the discrepancy between their inner structures. The main idea of Gromov-Wasserstein distance is to consider the transportation of similarity matrices of points which are in the same spaces. However, such formulation of Gromov-Wasserstein distance is computationally expensive as we need to solve a quadratic programming problem. To reduce the computational cost of Gromov-Wasserstein distance, (Peyré et al., 2016) propose to regularize the objective function of Gromov-Wasserstein based on the entropy of the transportation plan, which we refer to as *entropic Gromov-Wasserstein*. The entropic regularization idea had also been used earlier in optimal transport and had been shown to improve greatly the computation of optimal transport (Cuturi, 2013; Altschuler et al., 2017; Dvurechensky et al., 2018; Lin et al., 2019). The improvement in approximation of Gromov-Wasserstein distance via the entropic regularization has led to an increasing interest of using that divergence to several applications, including deep generative models (Bunne et al., 2019), computer graphics (Solomon et al., 2016; Xu et al., 2019; Chen et al., 2020), and natural language processing (Alvarez-Melis & Jaakkola, 2018; Grave et al., 2019). When measures are not probability measures and unbalanced, i.e., they may have different total masses, unbalanced version of Gromov-Wasserstein, named unbalanced Gromov-Wasserstein, via the idea of unbalanced optimal transport (Chizat et al., 2018) had been introduced in the recent work of (Séjourné et al., 2020). The entropic unbalanced Gromov-Wasserstein has been used in robust machine learning applications (Séjourné et al., 2020). Despite their practicalities, theoretical understandings of entropic (unbalanced) Gromov-Wasserstein are still nascent. The recent work of (Salmona et al., 2021) studied the closed-form expression of Gromov-Wasserstein between Gaussian distributions in different dimensions. However, to the best of our knowledge, the full theoretical analysis of entropic Gromov-Wasserstein and its unbalanced version between (unbalanced) Gaussian distributions in different dimensions has remained poorly understood.

* The first three authors contributed equally to this work, ** Nhat Ho and Tung Pham are co-last authors, equally contributed to this work. [1]University of Texas at Austin [2]École Polytechnique [3]VinAI Research [4]University of Michigan, Ann Arbor. Correspondence to: Nhat Ho, Tung Pham <minhnhat@utexas.edu, v.tungph4@vinai.io>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Our contribution. In this paper, we present a comprehensive study of the entropic Gromov-Wasserstein and its unbalanced version between (unbalanced) Gaussian distributions when inner product is considered as a cost function, which we refer to as entropic (unbalanced) inner product Gromov-Wasserstein (IGW). We also study its corresponding barycenter problem among multiple Gaussian distributions. Our work complements the works of (Gelbrich, 1990), (Janati et al., 2020) and (Salmona et al., 2021) when we give an explicit explanation of the effect of the entropic regularization on the OT and the geometric structure of an optimal transport plan between Gaussians. More particular, our contribution is four-fold and can be summarized as follows:

1. Balanced Gaussian measures: We first provide a closed-form expression of the entropic inner product Gromov-Wasserstein between Gaussian probability measures with zero means. We demonstrate that the expression depends mainly on eigenvalues of the covariance matrices of the two measures. Furthermore, an associate optimal transportation plan is shown to be also a zero-mean Gaussian measure. Our analysis hinges upon an application of von Neuman's trace inequality (Kristof, 1969; Horn & Johnson, 1991) for singular values of matrices.

2. Unbalanced Gaussian measures: Second, by relaxing marginal constraints via Kullback-Leibler (KL) divergence, we further study the entropic unbalanced IGW between two unbalanced Gaussian measures with zero means. The main challenge comes from the two KL terms that add another level of constraint in the objective function. To overcome that challenge, we show that the objective function can be broken down into several sub-problems that can be solved explicitly through some cubic and quadratic equations. That leads to an almost closed-form formulation of an optimal Gaussian transport plan of the entropic unbalanced IGW between the unbalanced Gaussian measures.

3. Barycenter problem: Finally, we investigate the (entropic) Gromov-Wasserstein barycenter problems with inner product cost, which we refer to as (entropic) IGW barycenter, among zero-mean balanced Gaussian measures. We prove that the barycenter of zero-mean Gaussian distributions is also a zero-mean Gaussian distribution with a diagonal covariance matrix. Reposing on that Gaussian barycenter, we can directly obtain a closed-form expression for the barycenter problem when the regularized parameter is sufficiently small, which is also the setting people widely use in practice.

4. Revisiting entropic optimal transport between Gaussian distributions. We note in passing that based on the new proof technique introduced for computing the entropic Gromov-Wasserstein between Gaussian distributions, we revisit the entropic Wasserstein between (unbalanced) Gaussian distributions in Appendix F and provide a simpler proof than that in (Janati et al., 2020) to derive the closed-form expression for that problem.

Paper organization. The paper is organized as follows. In Section 2, we provide backgrounds for the (unbalanced) inner product Gromov-Wasserstein and its corresponding entropic version. Next, we establish a closed-form expression of the entropic IGW between two balanced Gaussian distributions in Section 3 while proving that the entropic unbalanced IGW between unbalanced Gaussian measures also has a closed form in Section 4. Subsequently, we present our study of the (entropic) IGW barycenter problem among multiple Gaussian distributions in Section 5. In Section 6, we carry out empirical studies to investigate the behavior of algorithms solving the entropic Gromov-Wasserstein before concluding with a few discussions in Section 7. Additional proofs and auxiliary results are presented in the supplementary material.

Notation. We use the following notations throughout the paper. For a non-negative definite matrix A, if we do not specify the spectral decomposition of A, then generally, $\lambda_{a,i}$ or $\lambda_i(A)$ denotes the i-th largest eigenvalues of A. For random vectors $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_n)$, let K_{xy} be the covariance matrix between X and Y, which means that $\left(K_{xy}\right)_{i,j} = \mathrm{Cov}(X_i, Y_j)$. For square matrices A and B, we write $A \succeq B$ or $B \preceq A$ if $A - B$ is non-negative definite. For a positive integer n, we denote by Id_n the identity matrix of size $n \times n$ while $[n]$ stands for the set $\{1, 2, \ldots, n\}$. For any real number a, $[a]^+$ indicates $\max\{a, 0\}$. Lastly, we denote by $\mathcal{M}^+(\mathcal{X})$ the set of all positive measures in a space $\mathcal{X}$ and by $\mathcal{N}(\gamma, \Sigma)$ the Gaussian distribution in $\mathbb{R}^d$ with mean γ and covariance matrix Σ where d will be specified in each case.

2. Preliminaries

In this section, we first provide background for the Gromov-Wasserstein distance between two probability distributions and discuss its properties when using ℓ_2-norm as a cost function. Then, we present the formulation of an inner product Gromov-Wasserstein problem between two (unbalanced) Gaussian measures with a discussion about its geometric properties.

2.1. Gromov-Wasserstein distance

Let μ and ν be two probability measures on two Polish spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, and let $c_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be two measurable functions. Then, with $p \geq 1$, the p-Gromov-Wasserstein distance between μ and

ν, denoted by $\mathsf{GW}_p(\mu,\nu)$, is defined as follows:

$$\left(\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\left|c_{\mathcal{X}}(X,X') - c_{\mathcal{Y}}(Y,Y')\right|^p\right)^{\frac{1}{p}}, \quad (1)$$

where (X,Y) and (X',Y') are independent and identically distributed with probability distribution π that belongs to the set of joint probability distributions $\Pi(\mu,\nu)$, in which their marginal distributions are the corresponding distributions μ and ν.

Although the above problem always admits an optimal solution (Vayer, 2020), it is computationally expensive as we need to solve a quadratic optimization problem. For example, when the dimensions of $\mathcal{X}$ and $\mathcal{Y}$ are m and n, respectively, the complexity of solving the Gromov-Wasserstein minimization problem is $\mathcal{O}(m^2n+n^2m)$ (Peyré et al., 2016; Vincent-Cuaz et al., 2022).

2.2. Gromov-Wasserstein distance using ℓ_2-norm

Similar to the Wasserstein distance, there is no closed-form expression for the Gromov-Wasserstein distance between general distributions. Even in the important and specific cases when the input distributions are Gaussian, we still do not know if an optimal transport plan needs to be Gaussian or not (Salmona et al., 2021). However, there are some certain understandings of the Gromov-Wasserstein distance when the cost functions are Euclidean norms (Salmona et al., 2021; Vayer, 2020; Gelbrich, 1990),

$$\left(\inf_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\left|\|X-X'\|_m^2 - \|Y-Y'\|_n^2\right|^2\right)^{\frac{1}{2}}, \quad (2)$$

where $\|\cdot\|_m$ and $\|\cdot\|_n$ are Euclidean norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

Theorem 2.1. *(Salmona et al., 2021) Let $\mu\sim\mathcal{N}(m_0,\Sigma_0)$ and $\nu\sim\mathcal{N}(m_1,\Sigma_1)$ where $m_0\in\mathbb{R}^m$,$m_1\in\mathbb{R}^n$ are two vector means and $\Sigma_0\in\mathbb{R}^{m\times m}$ and $\Sigma_1\in\mathbb{R}^{n\times n}$ are the corresponding non-degenerated covariance matrices. Let P_0,D_0 and P_1,D_1 be respective diagonalizations of Σ_0 and Σ_1. Let us define $T_0: x\in\mathbb{R}^m\to P_0^\top(x-m_0)$ and $T_1: y\in\mathbb{R}^n\to P_1^\top(y-m_1)$. The Gromov-Wasserstein problem in equation (2) is equivalent to*

$$\sup_{X\sim T_0\#\mu, T_1\#\nu} \sum_{i,j} \mathsf{Cov}(X_i^2,Y_j^2) + 2\left\|\mathsf{Cov}(X,Y)\right\|_F^2$$

where $X=(X_1,\dots,X_m)^\top, Y=(Y_1,\dots,Y_n)^\top$ and $\|.\|_F$ is the Frobenius norm.

This theorem shows that the ℓ_2-norm Gromov-Wasserstein problem is equivalent to first shifting two input distributions to have mean zeros and then comparing covariance structures of the shifted distributions. After the shifting step, the translation invariant property is no longer needed and the problem is reduced to finding a match between the covariance structures.

2.3. Inner product Gromov-Wasserstein

In this section, we will provide background on the inner product Gromov-Wasserstein between two (unbalanced) Gaussian measures with a discussion about its geometric properties and the effect of an entropic regularization term on an optimal transport plan.

Although the Gromov-Wasserstein distance with ℓ_2-norm cost function has been used in several practical applications, finding its closed-form expression has remained an open problem due to its dependence on co-moments of order 2 and 4 of feasible transport plans (see Theorem 2.1) (Salmona et al., 2021). In particular, they can only derive the upper and lower bounds for $\mathsf{GW}_2(\mu,\nu)$ and they are even not able to show that an optimal transport plan is a Gaussian distribution. By considering the inner product cost function, our work can be seen as the first step towards better understanding and new insights on the ℓ_2-norm Gromov-Wasserstein distance between Gaussian distributions. Note that proof techniques ((Salmona et al., 2021), (Janati et al., 2020)) prior to our work to study the Gromov-Wasserstein distance between Gaussian measures were quite complicated.

Now, we reintroduce the Gromov-Wasserstein using the inner product as a *cost* function and its unbalanced variant.

Definition 2.2 (Inner product Gromov-Wasserstein). (Salmona et al., 2021) Let $X,X'\sim\mu$ be i.i.d random multivariates in $\mathbb{R}^m$ while $Y,Y'\sim\nu$ are i.i.d random multivariates in $\mathbb{R}^n$. Next, we assume that both (X,Y) and (X',Y') are jointly distributed as π. Then, the inner product Gromov-Wasserstein (IGW) problem is defined as

$$\mathsf{IGW}(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\}.$$

In this paper, we utilize an entropy-regularized approach by adding the relative entropy term to the above objective function As a result, we have the following definition, named *entropic IGW*, which is given by

$$\mathsf{IGW}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} + \varepsilon\mathrm{KL}\big(\pi\|\mu\otimes\nu\big), \quad (3)$$

where $\varepsilon>0$ is a regularized parameter and $\mathrm{KL}(\alpha||\beta) := \int\log\left(\frac{d\alpha}{d\beta}\right)d\alpha + \int(d\beta - d\alpha)$ for any positive measures α and β. Given that the entropic IGW between Gaussian distributions has a closed-form expression (see Theorem 3.1), it might be useful for some applications such as graph matching and word embedding alignment, which we leave for future development.

Entropic unbalanced IGW problem: When μ and ν are unbalanced Gaussian measures which are not probability distributions, the entropic IGW formulation in equation (3)

is no longer valid. One solution to deal with this issue is by regularizing the marginal constraints via KL divergence (Chizat et al., 2018). The entropic IGW problem between unbalanced measures, which we refer to as *entropic unbalanced IGW*, admits the following form:

$$\begin{aligned}\mathsf{UIGW}_{\varepsilon,\tau}(\mu,\nu) := \min_{\pi} \mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} \\ + \tau \mathrm{KL}^{\otimes}(\pi_x\|\mu) + \tau\mathrm{KL}^{\otimes}(\pi_y\|\nu) + \varepsilon\mathrm{KL}^{\otimes}\big(\pi\|\mu\otimes\nu\big),\end{aligned} \tag{4}$$

where $\tau > 0$, $\pi \in \mathcal{M}^+(\mathbb{R}^m \times \mathbb{R}^n)$ in the minimum, π_x and π_y are the projections of π on $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and $\mathrm{KL}^{\otimes}(\alpha||\beta) := \mathrm{KL}(\alpha\otimes\alpha||\beta\otimes\beta)$. Using the quadratic-KL divergence $\mathrm{KL}^{\otimes}$ (Séjourné et al., 2021) in place of the standard KL will result in $\mathsf{UIGW}_{\varepsilon,\tau}$ being 2-homogeneous, i.e., if μ and ν are multiplied by $\theta \geq 0$, then the value of $\mathsf{UIGW}_{\varepsilon,\tau}(\mu,\nu)$ is multiplied by θ^2.

Invariant property: Unlike the ℓ_2-norm Gromov Wasserstein in equation (2), which is invariant to both translation and rotation (Salmona et al., 2021), our IGW_ε and $\mathsf{UIGW}_{\varepsilon,\tau}$ only keeps unchanged under the latter transformation, which is shown in the following lemma with a note that the KL divergence is both rotation and translation invariant.

Lemma 2.3 (Invariant to rotation). *Let μ and ν be two probability measures on $\mathbb{R}^m$ and $\mathbb{R}^n$. Let $T_m : x \to O_m x$ and $T_n : y \to O_n y$ be two functions in which O_m and O_n are orthogonal matrices of size $m \times m$ and $n \times n$, respectively. Then, we have*

$$\mathsf{IGW}_\varepsilon(T_m\#\mu, T_n\#\nu) = \mathsf{IGW}_\varepsilon(\mu,\nu),$$

where $\#$ denotes the push-forward operator.

The proof of Lemma 2.3 can be found in Appendix A.1.

Mean-zero assumption: As being discussed above, the IGW_ε and $\mathsf{UIGW}_{\varepsilon,\tau}$ are not invariant to translation, because the IGW depends on the means of distributions, which is apparently undesirable. To get around this cumbersome, we assume that the means of two distributions are equal to zero, which is equivalent to the shifting step as shown in Section 2.2. Then, the translation invariant property is no longer a concern, and the problem is reduced to working with the covariance structures. By putting the mean-zero assumption, we still follow the main idea of the Gromov-Wasserstein distance, which is comparing the inner structures of two input distributions.

3. Entropic Gromov-Wasserstein between Balanced Gaussians

We present in this section a closed-form expression of the entropic inner product Gromov-Wasserstein between two Gaussian measures and a formula of a corresponding optimal transport plan in Theorem 3.1.

Let $\mu = \mathcal{N}(\mathbf{0}, \Sigma_\mu)$ and $\nu = \mathcal{N}(\mathbf{0}, \Sigma_\nu)$ be two Gaussian measures in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. To ease the ensuing presentation, in the entropic IGW problem, we denote the covariance matrix of the joint distribution π of μ and ν by

$$\Sigma_\pi = \begin{pmatrix} \Sigma_\mu & K_{\mu\nu} \\ K_{\mu\nu}^\top & \Sigma_\nu \end{pmatrix}. \tag{5}$$

In addition, the rotation invariant property of entropic IGW in Lemma 2.3 allows us to assume without loss of generality that Σ_μ and Σ_ν are diagonal matrices with positive diagonal entries sorted in descending order. In particular,

$$\begin{aligned}\Sigma_\mu &= \mathrm{diag}\big(\lambda_{\mu,1},\ldots,\lambda_{\mu,m}\big), \\ \Sigma_\nu &= \mathrm{diag}\big(\lambda_{\nu,1},\ldots,\lambda_{\nu,n}\big).\end{aligned} \tag{6}$$

This step simplifies significantly the unnecessarily complicated form of the general covariance matrix, since the Gromov-Wasserstein depends on the eigenvalues of the covariance matrix. The following result gives the closed-form expression for the entropic IGW between two balanced Gaussian distributions.

Theorem 3.1 (Entropic IGW has a closed form). *Suppose without loss of generality that $m \geq n$. Let $\mu = \mathcal{N}(\mathbf{0}, \Sigma_\mu)$ and $\nu = \mathcal{N}(\mathbf{0}, \Sigma_\nu)$ be two zero-mean Gaussian measures in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, where Σ_μ and Σ_ν are diagonal matrices given in equation (6). The entropic inner product Gromov-Wasserstein between μ and ν then equals to*

$$\begin{aligned}\mathsf{IGW}_\varepsilon(\mu,\nu) = \mathrm{tr}(\Sigma_\mu^2) + \mathrm{tr}(\Sigma_\nu^2) - 2\sum_{k=1}^n \Big[\lambda_{\mu,k}\lambda_{\nu,k} - \frac{\varepsilon}{4}\Big]^+ \\ + \frac{\varepsilon}{2}\sum_{k=1}^n \Big[\log(\lambda_{\mu,k}\lambda_{\nu,k}) - \log\Big(\frac{\varepsilon}{4}\Big)\Big]^+.\end{aligned}$$

Furthermore, an optimal transportation plan of the entropic IGW problem is a zero-mean Gaussian measure $\pi^* = \mathcal{N}(\mathbf{0}, \Sigma_{\pi^*})$, where

$$\Sigma_{\pi^*} = \begin{pmatrix} \Sigma_\mu & K_{\mu\nu}^* \\ (K_{\mu\nu}^*)^\top & \Sigma_\nu \end{pmatrix}, \tag{7}$$

with $K_{\mu\nu}^* = \mathrm{diag}\Big(\Big[\lambda_{\mu,k}\lambda_{\nu,k}\big[1 - \frac{\varepsilon}{4\lambda_{\mu,k}\lambda_{\nu,k}}\big]^+\Big]^{\frac{1}{2}}\Big)_{k=1}^n$ is an $m \times n$ matrix.

Remark 3.2. It can be seen that the quantity $\mathsf{IGW}_\varepsilon(\mu,\nu)$ depends only on the eigenvalues of the covariance matrices of μ and ν. In addition, as $\varepsilon \to 0$, $\mathsf{IGW}_\varepsilon(\mu,\nu)$ converges to $\sum_{k=1}^n(\lambda_{\mu,k} - \lambda_{\nu,k})^2 + \sum_{k=n+1}^m \lambda_{\mu,k}^2$, which is the squared Euclidean distance between covariance matrices Σ_μ and Σ_ν. In other words, $\mathsf{IGW}_\varepsilon(\mu,\nu)$ reflects the difference in the inner structures of the two input measures. Regarding the

optimal transport plan, when Σ_μ and Σ_ν are not diagonal, the covariance matrix $K^*_{\mu\nu}$ in equation (7) turns into

$$K^{*,\text{new}}_{\mu\nu} = P_\mu D_\mu^{\frac{1}{2}} \Lambda^{\frac{1}{2}}_{\mu\nu} D_\nu^{\frac{1}{2}} P_\nu^\top = P_\mu K^*_{\mu\nu} P_\nu^\top.$$

where $\Lambda^{\frac{1}{2}}_{\mu\nu} = \operatorname{diag}\Big(\Big\{\Big[1 - \frac{\varepsilon}{4\lambda_{\mu,k}\lambda_{\nu,k}}\Big]^+\Big\}^{\frac{1}{2}}\Big)_{k=1}^n$, P_μ, D_μ and P_ν, D_ν are the orthogonal diagonalizations of $\Sigma_\mu (= P_\mu D_\mu P_\mu^\top)$ and $\Sigma_\nu (= P_\nu D_\nu P_\nu^\top)$, respectively (see Appendix A.2 for the derivation).

Proof Sketch of Theorem 3.1. The full proof of this theorem is in Appendix A.2. First of all, we note that the objective function depends only on the covariance matrix of the joint distribution and the KL term. Given a fixed covariance matrix, the entropic term in equation (3) forces an optimal transport plan for the entropic IGW problem to be a Gaussian distribution (see Lemma 3.3 below). Subsequently, Lemma 3.4 (cf. the end of this proof) is used to attack the the determinant and the trace involving the covariance matrix $K_{\mu\nu}$ between X and Y. Then, the entropic IGW is reduced to a minimization problem of single variables in a particular interval, which has a closed-form solution. □

Lemma 3.3 (KL divergence minimum). *Let Q_v be the Gaussian distribution in $\mathbb{R}^d$ with mean v and covariance matrix Σ_v. Denote by Π_{u,Σ_u} the family of all probability distributions in $\mathbb{R}^d$ which have mean u and covariance matrix Σ_u, and are dominated by Q_v. Then,*

$$\mathcal{N}\big(u, \Sigma_u\big) = \operatorname*{arg\,min}_{P_u \in \Pi_{u,\Sigma_u}} \operatorname{KL}(P_u \| Q_v).$$

The proof of this lemma is deferred to Appendix A.3

Lemma 3.4 (Eigenvalues, determinant and trace). *Let $U_{\mu\nu}\Lambda^{\frac{1}{2}}_{\mu\nu}V^\top_{\mu\nu}$ be the SVD of $\Sigma_\mu^{-\frac{1}{2}} K_{\mu\nu} \Sigma_\nu^{-\frac{1}{2}}$, where $\Lambda^{\frac{1}{2}}_{\mu\nu} = \operatorname{diag}(\kappa^{\frac{1}{2}}_{\mu\nu,k})_{k=1}^n$ with $\kappa^{\frac{1}{2}}_{\mu\nu,k}$ is the k-th largest singular value of $\Sigma_\mu^{-\frac{1}{2}} K_{\mu\nu} \Sigma_\nu^{-\frac{1}{2}}$. Here, $\Lambda^{\frac{1}{2}}_{\mu\nu}$ is an $m \times n$ matrix, while $U_{\mu\nu}$ and $V_{\mu\nu}$ are unitary matrices of sizes $m \times m$ and $n \times n$, respectively. Then,*

(a) The values of $\kappa_{\mu\nu,1}, \ldots, \kappa_{\mu\nu,n}$ lie between 0 and 1.

(b) The determinant of matrix Σ_π could be computed as $\det(\Sigma_\pi) = \det(\Sigma_\mu)\det(\Sigma_\nu) \prod_{j=1}^n \big(1 - \kappa_{\mu\nu,j}\big)$.

(c) We have $\operatorname{tr}\big(K^\top_{\mu\nu} K_{\mu\nu}\big) \le \sum_{j=1}^n \lambda_{\mu,j}\lambda_{\nu,j}\kappa_{\mu\nu,j}$. The equality occurs when $U_{\mu\nu}$ and $V_{\mu\nu}$ are identity matrices of sizes m and n, respectively.

The proof of Lemma 3.4 can be found in Appendix A.4.

This lemma allows us to optimize the entropic IGW problem between Gaussian measures μ and ν over the singular values $(\kappa_{\mu\nu,k})_{k=1}^n$ of the matrix $K_{\mu\nu}$ rather than the transport plan π which is quite challenging to directly deal with. Part (a) of Lemma 3.4 shows a representation of the covariance matrix $K_{\mu\nu}$ through its singular values, which is useful for calculating its determinant in part (b). Our key technique to derive the closed-form expression for the entropic IGW lies in part (c) of Lemma 3.4. In that part, we utilize the von Neumann's trace inequality which says that the sum of singular values of the product of two matrices is maximized when their SVDs share the same orthogonal bases and their corresponding singular values match that of each other by their magnitudes. Specifically,

Lemma 3.5 (von Neumann's inequality). *(Kristof, 1969) Let $A_1, \ldots, A_n$ be fixed complex matrices and $Z_1, \ldots, Z_n$ run independently over all unitary matrices, all matrices are of size $m \times m$. Let $\alpha_{k1}, \ldots, \alpha_{km}$ be the singular values of A_k for all $k \in [n]$, all in the same sense monotonically ordered. Then, the maximum of the real part of $\operatorname{tr}\big([Z_1A_1]\ldots[Z_nA_n]\big)$ is given by $\sum_{\ell=1}^m \alpha_{1\ell}\ldots\alpha_{n\ell}$.*

In comparison with previous works such as (Janati et al., 2020) and (Salmona et al., 2021), we have a direct attack to the problem by figuring out a detailed formula for the covariance matrix and utilising the von Neumann's inequality to avoid solving a complicated system of equations of the first derivatives of the objective function.

Remark 3.6. In the step of applying part (c) of Lemma 3.4, the equality of the von Neumann's inequality for $\operatorname{tr}(K^\top_{\mu\nu}K_{\mu\nu})$ can occur in many cases (see Appendix A.4). Therefore, the optimal transport plan mentioned in Theorem 3.1 is not unique. For instance, $\pi' = \mathcal{N}(\mathbf{0}, \Sigma_{\pi'})$ where

$$\Sigma_{\pi'} = \begin{pmatrix} \Sigma_\mu & -K^*_{\mu\nu} \\ -K^*_{\mu\nu} & \Sigma_\nu \end{pmatrix},$$

is also an optimal transport plan.

4. Entropic Unbalanced Gromov-Wasserstein between Unbalanced Gaussians

In this section, we consider a more general setting when the two distributions μ and ν are unbalanced Gaussian measures in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Specifically,

$$\begin{aligned} \mu &= m_\mu \mathcal{N}(\mathbf{0}_m, \Sigma_\mu), \\ \nu &= m_\nu \mathcal{N}(\mathbf{0}_n, \Sigma_\nu), \end{aligned} \tag{8}$$

where $m \ge n$; $m_\mu, m_\nu > 0$ are their masses and Σ_μ, Σ_ν are given in equation (6). Here, μ and ν do not necessarily have the same mass, i.e., m_μ could be different from m_ν. For any positive measure α, we denote by $\overline{\alpha}$ the normalized measure of α. Thus, $\overline{\mu} = \mathcal{N}(\mathbf{0}_m, \Sigma_\mu)$ and $\overline{\nu} = \mathcal{N}(\mathbf{0}_n, \Sigma_\nu)$.

Under the setting of entropic unbalanced IGW, we denote

the covariance matrix of a transportation plan π by

$$\Sigma_\pi = \begin{pmatrix} \Sigma_x & K_{xy} \\ K_{xy}^\top & \Sigma_y \end{pmatrix}, \tag{9}$$

where Σ_x and Σ_y are covariance matrices of distributions π_x and π_y, respectively.

Note that the objective function of entropic unbalanced IGW in equation (4) involves two new KL divergence terms compare to that of the entropic IGW problem. Thus, before presenting the main theorem of this section, we introduce a tight lower bound for the KL divergence between two Gaussian distributions in the following lemma.

Lemma 4.1 (Lower bound for KL divergence). *Let π be given in equation (9), $\lambda_x = (\lambda_{x,i})_{i=1}^m$ be the eigenvalues of Σ_x, and $\lambda_y = (\lambda_{y,j})_{j=1}^n$ be the eigenvalues of Σ_y. Similarly, we define λ_μ for Σ_μ and λ_ν for Σ_ν. Then, we find that*

(a) $\mathrm{KL}(\overline{\pi}_x \| \overline{\mu}) \geq \frac{1}{2}\sum_{i=1}^m \Psi\big(\lambda_{x,i}\lambda_{\mu,i}^{-1}\big)$;

(b) $\mathrm{KL}(\overline{\pi}_y \| \overline{\nu}) \geq \frac{1}{2}\sum_{j=1}^n \Psi\big(\lambda_{y,j}\lambda_{\nu,j}^{-1}\big)$;

(c) $\mathrm{KL}(\overline{\pi} \| \overline{\mu} \otimes \overline{\nu}) = \mathrm{KL}(\overline{\pi}_x \| \overline{\mu}) + \mathrm{KL}(\overline{\pi}_y \| \overline{\nu}) + \frac{1}{2}\sum_{k=1}^n \log(1 - \kappa_{xy,k})$,

where $\Psi(x) = x - \log(x) - 1$, and $(\kappa_{xy,k}^{\frac{1}{2}})_{k=1}^n$ are singular values of matrix $\Sigma_x^{-\frac{1}{2}} K_{xy} \Sigma_y^{-\frac{1}{2}}$.

The proof of Lemma 4.1 is in Appendix A.5. Next, we define some functions and quantities that will be used in our analysis. Given $\tau, \varepsilon > 0$, let

$$\begin{aligned}
&g_{\varepsilon,\tau,+}(x, y; a, b) := x^2 + y^2 + \varepsilon\left[\log(xy) - \log\frac{\varepsilon}{2}\right]^+ \\
&\quad + (\tau+\varepsilon)\left[\frac{x}{a} + \frac{y}{b} - \log\left(\frac{xy}{ab}\right) - 2\right] - 2\left[xy - \frac{\varepsilon}{2}\right]^+; \\
&h_{\varepsilon,\tau}(x; a) := x^2 + (\tau+\varepsilon)\left[\frac{x}{a} - \log\left(\frac{x}{a}\right) - 1\right]; \\
&(\varphi_k, \phi_k) := \operatorname*{arg\,min}_{x,y>0}\ g_{\varepsilon,\tau,+}(x, y; \lambda_{\mu,k}\lambda_{\nu,k}),\ k \in [n]; \\
&\varphi_k := \operatorname*{arg\,min}_{x>0}\ h_{\varepsilon,\tau}(x, \lambda_{\mu,k}),\ k = n+1, \ldots, m;
\end{aligned}$$

for any $x, y, a, b > 0$. The detailed calculation of $(\varphi_k)_{k=1}^m$ and $(\phi_k)_{k=1}^n$ can be found in Appendix B.

Now, we are ready to state our main result regarding the entropic unbalanced IGW between two unbalanced Gaussian distributions.

Theorem 4.2 (Entropic unbalanced IGW has a closed form). *Let μ and ν be two unbalanced Gaussian distributions given in equation (8). Then, the entropic unbalanced IGW between μ and ν can be computed as follows:*

$$\begin{aligned}
\mathsf{UIGW}_{\varepsilon,\tau}(\mu, \nu) &= m_{\pi^*}^2 \Upsilon^* + \varepsilon \mathrm{KL}(m_{\pi^*}^2 \| m_\mu^2 m_\nu^2) \\
&\quad + \tau\Big\{\mathrm{KL}(m_{\pi^*}^2 \| m_\mu^2) + \mathrm{KL}(m_{\pi^*}^2 \| m_\nu^2)\Big\},
\end{aligned} \tag{10}$$

where

$$\begin{aligned}
m_{\pi^*} &:= (m_\mu m_\nu)^{\frac{\tau+\varepsilon}{2\tau+\varepsilon}} \exp\Big\{\frac{-\Upsilon^*}{2(2\tau+\varepsilon)}\Big\}, \\
\Upsilon^* &:= \sum_{k=1}^n g_{\varepsilon,\tau,+}(\varphi_k, \phi_k; \lambda_{\mu,k}, \lambda_{\nu,k}) \\
&\quad + \sum_{k=n+1}^m h_{\varepsilon,\tau}(\varphi_k; \lambda_{\mu,k}).
\end{aligned}$$

Furthermore, an optimal transportation plan is an unbalanced Gaussian measure $\pi^* = m_{\pi^*}\mathcal{N}(\mathbf{0}, \Sigma_{\pi^*})$, where

$$\Sigma_{\pi^*} = \begin{pmatrix} \Sigma_x^* & K_{xy}^* \\ (K_{xy}^*)^\top & \Sigma_y^* \end{pmatrix},$$

in which $\Sigma_x^* = \mathrm{diag}\big(\varphi_k\big)_{k=1}^m$, $\Sigma_y^* = \mathrm{diag}\big(\phi_k\big)_{k=1}^n$, and $K_{xy}^* = \mathrm{diag}\big(\psi_k\big)_{k=1}^n$ is an $m \times n$ matrix with

$$\psi_k := \Big\{\big[1 - \frac{\varepsilon}{2\varphi_k\phi_k}\big]^+ \varphi_k\phi_k\Big\}^{\frac{1}{2}},$$

for all $k \in [n]$.

Remark 4.3. Similar to $\mathsf{IGW}_\varepsilon(\mu, \nu)$, equation (10) shows that $\mathsf{UIGW}_\varepsilon(\mu, \nu)$ still depends on the eigenvalues of the covariance matrices of μ and ν via the quantity Υ^*, the detailed calculation of which is deferred to Appendix B. In addition, as μ and ν are assumed to be unbalanced in this case, the masses of these two measures are also included in the closed-form of $\mathsf{UIGW}_\varepsilon(\mu, \nu)$.

Proof Sketch of Theorem 4.2. The full proof of this theorem can be found in Appendix A.6. Recall that feasible transport plans π in the problem (4) are not necessarily probability measures. Therefore, we aim to optimize over the shape $\bar{\pi}$ and the masses m_π of π. To do so, we separate those two factors in the objective function (4) as follows:

$$\begin{aligned}
&m_\pi^2 \Upsilon + \varepsilon \mathrm{KL}(m_\pi^2 \| m_\mu^2 m_\nu^2) \\
&\quad + \tau\Big\{\mathrm{KL}(m_\pi^2 \| m_\mu^2) + \mathrm{KL}(m_\pi^2 \| m_\nu^2)\Big\},
\end{aligned} \tag{11}$$

where

$$\begin{aligned}
\Upsilon &:= \mathbb{E}_{\bar{\pi}\otimes\bar{\pi}}\big[\langle X, X'\rangle - \langle Y, Y'\rangle\big]^2 + 2\varepsilon \mathrm{KL}(\bar{\pi} \| \bar{\mu} \otimes \bar{\nu}) \\
&\quad + 2(\varepsilon + \tau)\Big\{\mathrm{KL}(\bar{\pi}_x \| \bar{\mu}) + \mathrm{KL}(\bar{\pi}_y \| \bar{\nu})\Big\}.
\end{aligned} \tag{12}$$

Shape optimization. In a similar fashion to the proof of Theorem 3.1, we show that the sum of the expectation and entropic terms in equation (12) depend only the covariance matrix Σ_π of π as below:

$$\begin{aligned}
&\|\lambda_x\|_2^2 + \|\lambda_y\|_2^2 - 2\sum_{k=1}^n \Big(\lambda_{x,k}\lambda_{y,k} - \frac{\varepsilon}{2}\Big)^+ \\
&\quad + \varepsilon \sum_{k=1}^n \Big[\log(\lambda_{x,k}\lambda_{y,k}) - \log\frac{\varepsilon}{2}\Big]^+,
\end{aligned}$$

where $\lambda_x, \lambda_y, \lambda_\mu$ and λ_ν are defined as in Lemma 4.1. Putting this result together with parts (a) and (b) of Lemma 4.1, the problem (12) is reduced to

$$\min_{\lambda_x,\lambda_y>0}\Big\{\sum_{k=1}^{n} g_{\varepsilon,\tau,+}(\lambda_{x,k},\lambda_{y,k};\lambda_{\mu,k},\lambda_{\nu,k}) + \sum_{k=n+1}^{m} h_{\varepsilon,\tau}(\lambda_{x,k};\lambda_{\mu,k})\Big\}.$$

Finally, Lemma C.4 (in Appendix C) allows us to optimize each summation term in the above problem independently but still preserving the decreasing order of eigenvalue sequences $(\lambda_{x,k})_{k=1}^n$. Eventually, we obtain the optimal value Υ^*, which will be calculated in detail in Appendix B.

Mass optimization. Based on the equation (11), the optimal mass m_{π^*} is the square root of the minimizer of the function

$$f(x) := x\Upsilon^* + \varepsilon \mathrm{KL}(x\|m_\mu^2 m_\nu^2) + \tau\Big\{\mathrm{KL}(x\|m_\mu^2) + \mathrm{KL}(x\|m_\nu^2)\Big\},$$

which is obtained by solving the equation $f'(x) = 0$. □

5. (Entropic) Gromov-Wasserstein Barycenter of Gaussian Distributions

In this section, we consider the problem of computing the (entropic) Gromov-Wasserstein barycenter of T Gaussian measures $\mu_1, \ldots, \mu_T$ defined in spaces of various dimensions. To tackle this problem, we need to fix the desired dimension of a barycenter, e.g, d, and choose beforehand the positive weighting coefficients $\alpha_1, \ldots, \alpha_T$ such that $\sum_{\ell=1}^T \alpha_\ell = 1$ associated with the measures $\mu_1, \ldots, \mu_T$. Here we allow the dimension of the barycenter to be flexible, since the given Gaussian measures could be in different dimensional spaces. Then, the barycenter of T positive measures under the inner-product Gromov-Wasserstein distance is defined as follows:

Definition 5.1 (Inner product Gromov-Wasserstein barycenter). Let μ_ℓ be a positive measure in $\mathbb{R}^{m_\ell}$ for any $\ell \in [T]$. Let $\alpha_1, \ldots, \alpha_T$ be positive constants such that $\sum_{\ell=1}^T \alpha_\ell = 1$. The inner product Gromov-Wasserstein (IGW) barycenter of $\{\mu_\ell, \alpha_\ell\}_{\ell=1}^T$ is defined as the probability distribution of the random vector Y in $\mathbb{R}^d$ which is a solution of the following problem

$$\operatorname*{arg\,min}_{\substack{X_\ell\sim\mu_\ell;\ \dim(Y)=d;\\ (X_\ell,Y),(X'_\ell,Y)\sim\pi_{\ell,y}}} \sum_{\ell=1}^{T} \alpha_\ell \mathbb{E}_{\pi_{\ell,y}\otimes\pi_{\ell,y}}\Big\{\big[\langle X_\ell, X'_\ell\rangle - \langle Y, Y'\rangle\big]^2\Big\}, \tag{13}$$

where in the above minimum, (X_ℓ, Y) and (X'_ℓ, Y') are i.i.d random vectors for each $\ell = 1, \ldots, T$.

Based on this definition, the following theorem provides a closed-form expression for the IGW barycenter of T Gaussian measures.

Theorem 5.2 (Inner product Gromov-Wasserstein barycenter has a closed form). *Let $\mu_\ell = \mathcal{N}(\mathbf{0}, \Sigma_\ell)$ be Gaussian distribution in $\mathbb{R}^{m_\ell}$ for all $\ell \in [T]$, where $\Sigma_\ell = \mathrm{diag}\big(\lambda_{\ell,i}\big)_{i=1}^{m_\ell}$ is an $m_\ell \times m_\ell$ positive definite matrix. Let d be a positive integer and assume that $d \le \max_{\ell\in[T]} m_\ell$, the IGW barycenter of the formulation (13) admits a Gaussian solution: $\mu^* = \mathcal{N}(\mathbf{0}, \Sigma_{\mu^*})$, where*

$$\Sigma_{\mu^*} = \mathrm{diag}\big(\lambda_{\mu^*,j}\big)_{j=1}^{d},$$

in which $\lambda_{\mu^*,j} = \sum_{\ell=1}^T \alpha_\ell \lambda_{\ell,j} \mathbf{1}_{j\le m_\ell}$ for all $j \in [d]$.

The proof of Theorem 5.2 can be found in Appendix A.7. Subsequently, we define the formulation of the barycenter of positive measures using the entropic IGW.

Definition 5.3 (Entropic inner product Gromov-Wasserstein barycenter). Let μ_ℓ be a positive measure in $\mathbb{R}^{m_\ell}$ for any $\ell \in [T]$. Let $\alpha_1, \ldots, \alpha_T$ be positive constants such that $\sum_{\ell=1}^T \alpha_\ell = 1$. The entropic inner-product Gromov-Wasserstein barycenter of $\{\mu_\ell, \alpha_\ell\}_{\ell=1}^T$ is defined as the probability distribution of the random vector Y in $\mathbb{R}^d$ which is a solution of the following problem

$$\operatorname*{arg\,min}_{\substack{X_\ell\sim\mu_\ell; Y\in\mathbb{R}^d; Y\sim\mu;\\ (X_\ell,Y),(X'_\ell,Y)\sim\pi_{\ell,y}}} \sum_{\ell=1}^{T} \alpha_\ell \Big\{\mathbb{E}_{\pi_{\ell,y}\otimes\pi_{\ell,y}}\big[\langle X_\ell, X'_\ell\rangle - \langle Y, Y'\rangle\big]^2 + \varepsilon \mathrm{KL}(\pi_{\ell,y}\|\mu_\ell\otimes\mu)\Big\}, \tag{14}$$

where in the above minimum, $\{X'_\ell\}$ and $\{X_\ell\}$ are i.i.d. random vectors and Y and Y' are i.i.d. random vectors.

From that definition, we have the following result for the entropic IGW barycenter of T Gaussian distributions.

Theorem 5.4 (Entropic inner product Gromov-Wasserstein barycenter has a closed form). *Let $\mu_\ell = \mathcal{N}(\mathbf{0}, \Sigma_\ell)$ be Gaussian distribution in $\mathbb{R}^{m_\ell}$ for all $\ell \in [T]$, where $\Sigma_\ell = \mathrm{diag}\big(\lambda_{\ell,i}\big)_{i=1}^{m_\ell}$ is an $m_\ell \times m_\ell$ positive definite matrix. Let d be a positive integer such that $d \le \max_{\ell\in[T]} m_\ell$. Define $d_\ell = \min\{d, m_\ell\}$. Let ε be a positive constant satisfying*

$$\varepsilon \le 2\lambda_{\ell,j}\big\{A_j + (A_j^2 - \varepsilon B_j)^{\frac{1}{2}}\big\}, \tag{15}$$

and $A_j^2 \ge \varepsilon B_j$ for any $\ell \in [T]$ and $j \in [d]$, where $A_j = \sum_{\ell=1}^T \alpha_\ell \lambda_{\ell,j} \mathbf{1}_{j\le d_\ell}$ and $B_j = \sum_{\ell=1}^T \alpha_\ell \mathbf{1}_{j\le d_\ell}$. Then, the barycenter of the formulation (14) admits a Gaussian solution which has the form: $\mu^* = \mathcal{N}(\mathbf{0}, \Sigma_{\mu^*})$, in which $\Sigma_{\mu^*} = \mathrm{diag}(\lambda_{\mu^*,j})_{j=1}^d$ with

$$\lambda_{\mu^*,j} = \sum_{\ell=1}^{T} \alpha_\ell \lambda_{\ell,j} \mathbf{1}_{j\le d_\ell}\left\{1 - \frac{\varepsilon}{\lambda_{\ell,j}\big[A_j + (A_j^2 - \varepsilon B_j)^{\frac{1}{2}}\big]}\right\}.$$

The proof of Theorem 5.4 is in Appendix A.8.

Remark 5.5. In the above theorem, we need a set of conditions for ε that could be satisfied when ε is small, which is often chosen in practice. When ε is not small, then the readers could follow the guideline in Lemma C.5 in Appendix C to find the covariance matrix Σ_{μ^*}. The proof of Theorem 5.4 needs the below lemma to show the existence of a Gaussian minimizer.

Lemma 5.6. *Let $Q_X = \mathcal{N}(\gamma_x, \Sigma_x)$ be a Gaussian distribution in $\mathbb{R}^m$ and P_Y be a probability distribution with mean $\gamma_y \in \mathbb{R}^n$ and covariance matrix Σ_y. Denote by $\Pi(Q_X, P_Y)$ the family of all probability distributions in $\mathbb{R}^{m+n}$ which have marginal distributions Q_X and P_Y, and covariance matrix $\Sigma_{X,Y} = \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{pmatrix}$. Here, Σ_x and Σ_y are non-degenerate covariance matrices of size $m \times m$ and $n \times n$, respectively. Then,*

$$\mathcal{N}\big([\gamma_x, \gamma_y]^\top, \Sigma_{X,Y}\big) \in \operatorname*{arg\,min}_{P_{X,Y} \in \Pi(Q_X, P_Y)} \mathsf{KL}\big(P_{X,Y} \| Q_X \otimes P_Y\big).$$

The proof of Lemma 5.6 is in Appendix A.9.

6. Empirical Studies

In this section, we will use the derived closed-forms to inspect the behavior of algorithms solving entropic Gromov-Wasserstein problem, in particular those studied in (Peyré et al., 2016). We use the implementation of these algorithms in Python Optimal Transport library (Flamary & Courty, 2017). The implementation is available at https://github.com/lntk/egw_gaussians.

Projected mirror descent for IGW_ε (Figure 1). The algorithm in (Peyré et al., 2016) reads

$$C^{(t)} = -D_{\mathbf{a}} P^{(t)} D_{\mathbf{b}}$$
$$P^{(t+1)} = \min_{P \in \Pi(\mathbf{a}, \mathbf{b})} \langle P, C^{(t)} \rangle - \varepsilon H(P),$$

in which $D_{\mathbf{a}}, D_{\mathbf{b}}$ correspond to inner product cost matrices on supports of μ_N and ν_N, $\Pi(\mathbf{a}, \mathbf{b})$ is the set of discrete couplings $\{P \in \mathbb{R}^{N \times N} : P\mathbf{1}_N = \mathbf{a}, P^\top \mathbf{1}_N = \mathbf{b}\}$, and $\langle P, C^{(t)} \rangle = \sum_{i,j \in [N]} P_{ij} C_{ij}^{(t)}$. Here, we set $m = 2, n = 3$ and sample N (between 10 and 2000) data points from $\mathcal{N}(\mathbf{0}_2, \Sigma_\mu)$ and $\mathcal{N}(\mathbf{0}_3, \Sigma_\nu)$, in which Σ_μ and Σ_ν are diagonal matrices whose diagonal values are uniformly sampled from the interval $[0, 2]$. The regularization parameter ε is chosen from $\{0.5, 1, 5\}$. We plot means and one standard-deviation areas over 20 independent runs for each N. As expected, with more samples we can approximate IGW_ε more accurately. However, since this algorithm only converges to a stationary point, there might still be a gap between the converged value and the optimal value.

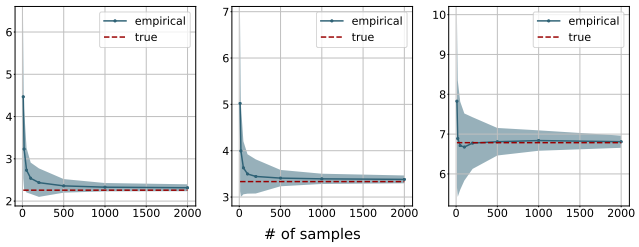

Figure 1. Empirical convergence for (Peyré et al., 2016) in computing IGW_ε. From left to right: $\varepsilon = 0.5$, $\varepsilon = 1$ and $\varepsilon = 5$. The red dashed lines correspond to the theoretical optimal values from Theorem 3.1, while the blue lines and shaded regions are the means and standard variations of the objective values computed according to (Peyré et al., 2016), respectively.

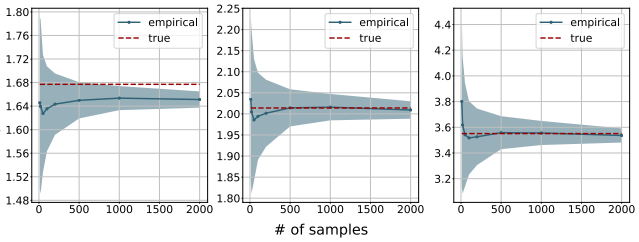

Figure 2. Empirical convergence for (Séjourné et al., 2021) in computing $\mathsf{UIGW}_{\varepsilon,\tau}$. From left to right: $\varepsilon = 0.5$, $\varepsilon = 1$ and $\varepsilon = 5$ while τ is fixed to 1. The red dashed lines correspond to the theoretical optimal values from Theorem 4.2, while the blue lines and shaded regions are the means and standard variations of the objective values computed according to (Séjourné et al., 2021), respectively.

Alternating minimization for $\mathsf{UIGW}_{\varepsilon,\tau}$ (Figure 2). Next, we use the closed-form for the unbalanced formulation (see Theorem 4.2) to demonstrate the convergence of a recently proposed algorithm (Séjourné et al., 2021, Algorithm 1), solving a bi-convex relaxation of the problem (4), which is tight for negative semi-definite kernels and the mass balance between two measures (Séjourné et al., 2021, Theorem 4). Thus, we will only consider two measures of the same mass in this section, and the setting is essentially the same as the previous section as follows: $m = 2, n = 3$ and we sample N (between 10 and 2000) data points from $\mathcal{N}(\mathbf{0}_2, \Sigma_\mu)$ and $\mathcal{N}(\mathbf{0}_3, \Sigma_\nu)$, in which Σ_μ and Σ_ν are diagonal matrices whose diagonal values are uniformly samples from the interval $[0, 2]$. The regularization parameter ε is chosen from $\{0.5, 1, 5\}$ and τ is set to 1. We plot means and one standard-deviation areas over 20 independent runs for each N. The observed behavior is also similar to what we saw previously: the approximation quality for $\mathsf{UIGW}_{\varepsilon,\tau}$ increases with more samples.

Visualizing optimal transportation plans of for $n = m = 1$ (Figure 3). Now we take a look at the trans-

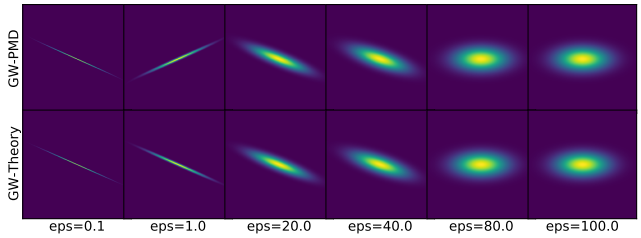

Figure 3. Empirical transportation plans between 1-dimensional Gaussians $\mathcal{N}(0,2)$ and $\mathcal{N}(0,10)$. From left to right: $\varepsilon = 0.1, \varepsilon = 1, \varepsilon = 20, \varepsilon = 40, \varepsilon = 80, \varepsilon = 100$. The first rows present plans returned by (Peyré et al., 2016) and the second row corresponds to theoretical plans in Theorem 3.1.

portation plans computed by the above algorithm for different values of ε. Specifically, we create 1000-bin histograms of $\mathcal{N}(0,\alpha)$ and $\mathcal{N}(0,\beta)$ (which will be μ_N and ν_N with $N = 1000$) in which $\alpha = 2, \beta = 10$, and set $\varepsilon \in \{0.1, 1, 20, 40, 80, 100\}$ (values in this set are chosen based on $\alpha\beta$). The algorithm is run till convergence (with tolerance 10^{-9}) and the plans are plotted in Figure 3. It is apparent that the plans returned by the projected mirror descent algorithm are Gaussians and resemble our theoretical ones (except for the case $\varepsilon = 1$, but noting that when $n = m = 1$, if $\mathcal{N}(\mathbf{0}, \left(\begin{smallmatrix}\alpha & \gamma\\ \gamma & \beta\end{smallmatrix}\right))$ is an optimal plan for IGW_ε, so is $\mathcal{N}(\mathbf{0}, \left(\begin{smallmatrix}\alpha & -\gamma\\ -\gamma & \beta\end{smallmatrix}\right))$.

Alternating minimization for the barycenter problem (Figure 4). Given T discrete measures $\{\mu_N^{(i)}\}_{i=1}^T$, a set of weights $\{\alpha_i\}_{i=1}^T$, cost matrices $C^{(i)}$ corresponding to measure $\mu_N^{(i)}$ and a probability vector p, (Peyré et al., 2016) proposes an alternating minimization scheme to find a d-dimensional barycenter μ of $\{\mu_N^{(i)}\}_{i=1}^T$ with probability mass p, i.e., finding the cost matrix C on the support of this barycenter. In our case of inner product metric, C is a Gram matrix, and we can recover the supports of μ via the Cholesky decomposition $C = LL^\top$ where $L \in \mathbb{R}^{N\times N}$ and choosing the first d columns of L. In our setting, for $i \in [T]$, we choose $\mu_N^{(i)} = \sum_{j=1}^N \frac{1}{N}\delta_{X_j}$ where $X_j \sim \mathcal{N}(0, \Sigma^{(i)})$, $C_{kl}^{(i)} = \langle X_k, X_l\rangle$ for $k, l \in [N]$ and $p = \mathbf{1}_N/N$. Specifically, we consider $N = 1000, T = 3, \Sigma^{(1)} = 5, \Sigma^{(2)} = \mathrm{diag}(10,1), \Sigma^{(3)} = \mathrm{diag}(2,1,1)$, $\alpha = (0.3, 0.6, 0.1)$ and $d = 2$. The regularization parameter ε is set to 0.1, which satisfies all the constraints in Theorem 5.4. After finding N support points for μ by the discussed method, we compute the sample covariance matrix $\hat{\Sigma}$, perform orthogonal diagonalization, and apply the corresponding transformation to the support. We then compute its sample mean and sample covariance matrix, and compare the Gaussian with these statistics with our theoretical optimal plan (see Figure 4). It is worth noting that any orthogonally-transformed version of the found barycenter is also a barycenter of $\{\mu_N^{(i)}\}_{i=1}^T$ under the Gromov-Wasserstein setting.

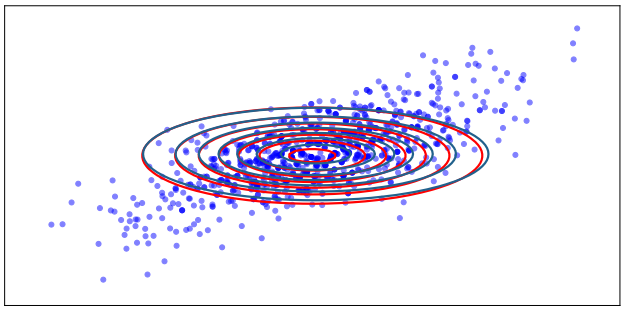

Figure 4. Visualizing the barycenter. The blue points correspond to one possible set of support for the computed barycenter by (Peyré et al., 2016) (see discussion). The blue contours are the rotated version of the Gaussian fitted to this support set (for comparison). The red contours represent the theoretical barycenter whose form is presented in Theorem 5.4.

7. Conclusion

In this paper, we provide a comprehensive study of the entropic (unbalanced) inner product Gromov-Wasserstein (IGW) between (unbalanced) Gaussian distributions. We demonstrate that the optimal transportation plan is (unbalanced) Gaussian distribution. Based on that result and a novel application of von Neumann's trace inequality, we derive the closed-form expression for the entropic (unbalanced) IGW between these distributions. Finally, we also consider the (entropic) Gromov-Wasserstein barycenter problem of multiple Gaussian measures. We prove that the barycenter problem admits a Gaussian minimizer and obtain the closed-form expression for the covariance matrix of the barycenter when the entropic regularization parameter is small.

Acknowledgements

NH gratefully acknowledges support from the NSF IFML 2019844 award and research gifts by the NSF AI Institute for Foundations of Machine Learning.

References

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in neural information processing systems*, pp. 1964–1974, 2017. (Cited on page 1.)

Alvarez-Melis, D. and Jaakkola, T. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1881–1890, 2018. (Cited

on page 1.)

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017. (Cited on page 1.)

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015. (Cited on page 1.)

Bunne, C., Alvarez-Melis, D., Krause, A., and Jegelka, S. Learning generative models across incomparable spaces. In *International Conference on Machine Learning*, pp. 851–861, 2019. (Cited on page 1.)

Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., and Liu, J. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pp. 1542–1553. PMLR, 2020. (Cited on page 1.)

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. (Cited on pages 1 and 4.)

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016. (Cited on page 1.)

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pp. 2292–2300, 2013. (Cited on page 1.)

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 447–463, 2018. (Cited on page 1.)

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019. (Cited on page 1.)

Dvurechensky, P., Gasnikov, A., and Kroshnin, A. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *ICML*, pp. 1367–1376, 2018. (Cited on page 1.)

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch Wasserstein: asymptotic and gradient properties. In *AISTATS 2020-23nd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1–20, 2020. (Cited on page 1.)

Flamary, R. and Courty, N. Pot python optimal transport library, 2017. URL https://pythonot.github.io/. (Cited on page 8.)

Gelbrich, M. On a formula for the L2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147:185–203, 1990. (Cited on pages 2 and 3.)

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018. (Cited on page 1.)

Grave, E., Joulin, A., and Berthet, Q. Unsupervised alignment of embeddings with Wasserstein procrustes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1880–1890, 2019. (Cited on page 1.)

Ho, N., Nguyen, X., Yurochkin, M., Bui, H. H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *International Conference on Machine Learning*, pp. 1501–1509, 2017. (Cited on page 1.)

Ho, N., Huynh, V., Phung, D., and Jordan, M. I. Probabilistic multilevel clustering via composite transportation distance. In *AISTATS*, 2019. (Cited on page 1.)

Horn, R. A. and Johnson, C. R. *Topics in Matrix Analysis*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511840371. (Cited on pages 2, 15, 19, 33, 37, and 38.)

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. Entropic optimal transport between unbalanced gaussian measures has a closed form. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10468–10479, 2020. (Cited on pages 2, 3, 5, 12, and 32.)

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016. (Cited on page 1.)

Kristof, W. Generalization of a theorem by John von Neumann on the trace of certain matrix products. *ETS Research Bulletin Series*, 1969:i–13, 12 1969. doi: 10.1002/j.2333-8504.1969.tb00744.x. (Cited on pages 2, 5, 15, 19, 33, 37, and 38.)

Le, T., Nguyen, T., Ho, N., Bui, H., and Phung, D. LAMDA: Label matching deep domain adaptation. In *ICML*, 2021. (Cited on page 1.)

Lin, T., Ho, N., and Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pp. 3982–3991, 2019. (Cited on page 1.)

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pp. 4104–4113. PMLR, 2019. (Cited on page 1.)

Nguyen, K., Ho, N., Pham, T., and Bui, H. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=QYjO70ACDK. (Cited on page 1.)

Nguyen, K., Nguyen, D., Nguyen, Q., Pham, T., Bui, H., Phung, D., Le, T., and Ho, N. On transportation of mini-batches: A hierarchical approach. *arXiv preprint arXiv:2102.05912*, 2021b. (Cited on page 1.)

Nguyen, T., Pham, Q.-H., Le, T., Pham, T., Ho, N., and Hua, B.-S. Point-set distances for learning representations of 3D point clouds. In *ICCV*, 2021c. (Cited on page 1.)

Peyré, G., Cuturi, M., and Solomon, J. Gromov-Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pp. 2664–2672, 2016. (Cited on pages 1, 3, 8, and 9.)

Salimans, T., Zhang, H., Radford, A., and Metaxas, D. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. (Cited on page 1.)

Salmona, A., Delon, J., and Desolneux, A. Gromov-Wasserstein distances between Gaussian distributions. *Arxiv Preprint Arxiv: 2104.07970*, 2021. (Cited on pages 1, 2, 3, 4, and 5.)

Séjourné, T., Vialard, F.-X., and Peyré, G. The unbalanced Gromov Wasserstein distance: Conic formulation and relaxation. *Arxiv Preprint Arxiv: 2009.04266*, 2020. (Cited on page 1.)

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016. (Cited on page 1.)

Séjourné, T., Vialard, F.-X., and Peyré, G. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. In *Advances in Neural Information Processing Systems*, 2021. (Cited on pages 4 and 8.)

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018. (Cited on page 1.)

Vayer, T. A contribution to optimal transport on incomparable spaces, 2020. (Cited on page 3.)

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Semi-relaxed gromov-wasserstein divergence with applications on graphs, 2022. (Cited on page 3.)

Xu, H., Luo, D., and Carin, L. Scalable Gromov-Wasserstein learning for graph partitioning and matching. In *Advances in neural information processing systems*, pp. 3052–3062, 2019. (Cited on page 1.)

Supplementary Materials for "Entropic Gromov-Wasserstein between Gaussian Distributions"

In this supplement, we first provide proofs of remaining lemmas and theorems in Appendix A. Then, we present how to derive a closed-form formulation for Υ^* in Theorem 4.2 in Appendix B. Auxiliary results are presented in Appendix C while another proof of Lemma C.4 is in Appendix D. We highlight the issue of obtaining closed-form expression for entropic inner product Gromov-Wasserstein between non-zero means Gaussian distributions in Appendix E. Finally, we revisit the entropic optimal transport between (unbalanced) Gaussian distributions in Appendix F and provide a simpler proof than that in (Janati et al., 2020) to derive the closed-form expression for that problem.

A. Proofs of remaining results

This appendix is devoted to provide the proofs of lemmas and theorems presented in the paper.

A.1. Proof of Lemma 2.3

Firstly, note that when $\pi' \in \Pi(T_m\#\mu, T_n\#\nu)$, there exists $\pi \in \Pi(\mu,\nu)$ such that $(T_m, T_n)\#\pi = \pi'$. Therefore, we have

$$\begin{aligned}\mathbb{E}_{\pi'\otimes\pi'}[\langle X, X'\rangle - \langle Y, Y'\rangle]^2 &= \mathbb{E}_{(T_m,T_n)\#\pi\otimes(T_m,T_n)\#\pi}[\langle T_m X, T_m X'\rangle - \langle T_n Y, T_n Y'\rangle]^2 \\ &= \mathbb{E}_{\pi\otimes\pi}[\langle X, X'\rangle - \langle Y, Y'\rangle]^2. \end{aligned} \tag{16}$$

Next, we will show that π is absolutely continuous w.r.t $\mu\otimes\nu$ if and only if π' is absolutely continuous w.r.t $(T_m\#\mu\otimes T_n\#\nu)$. Assume that π is absolutely continuous w.r.t $\mu\otimes\nu$. Consider two arbitrary Borel set U, V such that

$$(T_m\#\mu)(U)\times(T_n\#\nu)(V) = 0,$$

which is equivalent to $\mu(T_m^{-1}(U))\times\nu(T_n^{-1}(V)) = 0$. Since π is dominated by $\mu\otimes\nu$, we have $\pi(T_m^{-1}(U), T_n^{-1}(V)) = 0$, or equivalently,

$$\pi'(U,V) = [(T_m,T_n)\#\pi](U,V) = 0.$$

Thus, the first part of the above statement is proved, and the second part is obtained similarly. It follows that the KL terms used below are well defined.

Notice that $d\pi'(x,y) = d\pi(T_m^{-1}(x), T_n^{-1}(y))$, $d(T_m\#\mu)(x) = d\mu(T_m^{-1}(x))$ and $d(T_n\#\nu)(y) = d\nu(T_n^{-1}(y))$. Therefore, by changing of variables $x \mapsto T_m(u)$ and $y \mapsto T_n(v)$, we get

$$\begin{aligned}&\mathrm{KL}(\pi'\|(T_m\#\mu)\otimes(T_n\#\nu)) \qquad (17)\\ &= \int \log\left(\frac{d\pi'(x,y)}{d(T_m\#\mu)(x)d(T_n\#\nu)(y)}\right) d\pi'(x,y)\\ &= \int \log\left(\frac{d\pi'(T_m(u),T_n(v)))}{d(T_m\#\mu)(T_m(u))d(T_n\#\nu)(T_n(v))}\right) |\det(T_m'(u))\det(T_n'(v))| d\pi'(T_m(u),T_n(v))\\ &= \int \log\left(\frac{d\pi(u,v)}{d\mu(u)d\nu(v)}\right) d\pi(u,v)\\ &= \mathrm{KL}(\pi\|\mu\otimes\nu), \qquad (18)\end{aligned}$$

with a note that $|\det(T_m'(u)| = |\det(O_m)| = 1$ and $|\det(T_n'(v))| = |\det(O_n)| = 1$ since O_m and O_n are orthogonal matrices. Putting the results in equations (16) and (18) together, we obtain the conclusion.

A.2. Proof of Theorem 3.1

Firstly, we will show that

$$\mathbb{E}_{\pi\otimes\pi}\left\{\left[\langle X, X'\rangle - \langle Y, Y'\rangle\right]^2\right\} = \mathrm{tr}(\Sigma_\mu^2) + \mathrm{tr}(\Sigma_\nu^2) - 2\,\mathrm{tr}\left(K_{\mu\nu}^\top K_{\mu\nu}\right). \tag{19}$$

Indeed, we have

$$\mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle-\langle Y,Y'\rangle\big]^2\Big\}=\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle^2+\mathbb{E}_{\pi\otimes\pi}\langle Y,Y'\rangle^2-2\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle\langle Y,Y'\rangle.$$

As π is a coupling of two zero-mean distributions μ and ν, π also has mean zero. Combining this result with the independence between X and X' and Y and Y' leads to

$$\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle^2=\operatorname{tr}(\Sigma_\mu^2)\quad\text{ and }\quad\mathbb{E}_{\pi\otimes\pi}\langle Y,Y'\rangle^2=\operatorname{tr}(\Sigma_\nu^2).$$

Meanwhile, we have

$$\mathbb{E}_{\pi\otimes\pi}[\langle X,X'\rangle\langle Y,Y'\rangle]=\mathbb{E}_{\pi\otimes\pi}\Big[\sum_{i=1}^m\sum_{j=1}^n X_iX_i'Y_jY_j'\Big]=\sum_{i=1}^m\sum_{j=1}^n\mathbb{E}_{\pi\otimes\pi}[X_iY_j]\mathbb{E}_\pi[X_i'Y_j']=\operatorname{tr}\left(K_{\mu\nu}^\top K_{\mu\nu}\right).$$

Putting the above results together, we obtain the desired equality (19). It indicates that we can rewrite the formulation of $\mathsf{IGW}_\varepsilon(\mu,\nu)$ as follows:

$$\mathsf{IGW}_\varepsilon(\mu,\nu)=\operatorname{tr}(\Sigma_\mu^2)+\operatorname{tr}(\Sigma_\nu^2)+\min_{\pi\in\Pi(\mu,\nu)}\Big\{\varepsilon\mathrm{KL}(\pi\|\mu\otimes\nu)-2\operatorname{tr}\left(K_{\mu\nu}^\top K_{\mu\nu}\right)\Big\}.\tag{20}$$

To solve the minimization problem (20), we firstly fix the matrix $K_{\mu\nu}$, therefore, the covariance matrix of π is also fixed due to equation (5). By Lemma 3.3, the optimal transport plan of this problem needs to be a Gaussian distribution. Thus, according to Lemma C.3 in Appendix C and part (b) of Lemma 3.4, the entropic term in the objective function reads

$$\begin{aligned}\varepsilon\mathrm{KL}(\pi\|\mu\otimes\nu)&=\frac{1}{2}\varepsilon\Big\{\operatorname{tr}\left(\Sigma_\pi\Sigma_{\mu\otimes\nu}^{-1}\right)-(m+n)+\log\Big(\frac{\det(\Sigma_{\mu\otimes\nu})}{\det(\Sigma_\pi)}\Big)\Big\}\\&=-\frac{\varepsilon}{2}\sum_{i=1}^n\log(1-\kappa_{\mu\nu,i}),\end{aligned}$$

where $([\kappa_{\mu\nu,i}]^{\frac{1}{2}})_{i=1}^n$ are singular values (in descending order) of matrix $\Sigma_\mu^{-\frac{1}{2}}K_{\mu\nu}\Sigma_\nu^{-\frac{1}{2}}$. By applying part (c) of Lemma 3.4, the optimal value of the term $\operatorname{tr}\left(K_{\mu\nu}^\top K_{\mu\nu}\right)$ is achieved at $K_{\mu\nu}=\Sigma_\mu^{\frac{1}{2}}\operatorname{diag}([\kappa_{\mu\nu,k}]^{\frac{1}{2}})_{k=1}^n\Sigma_\nu^{\frac{1}{2}}$. Therefore, the optimization problem (20) reduces to

$$\mathsf{IGW}_\varepsilon(\mu,\nu)=\operatorname{tr}(\Sigma_\mu^2)+\operatorname{tr}(\Sigma_\nu^2)-\max_{\kappa_{\mu\nu,k}\in[0,1)}\Big\{2\sum_{k=1}^n\lambda_{\mu,k}\lambda_{\nu,k}\kappa_{\mu\nu,k}+\frac{\varepsilon}{2}\sum_{k=1}^n\log(1-\kappa_{\mu\nu,k})\Big\}.$$

The function $f(x)=ax+\frac{\varepsilon}{4}\log(1-x)$ for $a>0$ determined in the interval $[0,1)$ attains its maximum at $x=\big[1-\frac{\varepsilon}{4a}\big]^+$. Thus, $\kappa_{\mu\nu,k}^*=\big[1-\frac{\varepsilon}{4\lambda_{\mu,k}\lambda_{\nu,k}}\big]^+$ for all $k\in[n]$, and

$$\mathsf{IGW}_\varepsilon(\mu,\nu)=\operatorname{tr}(\Sigma_\mu^2)+\operatorname{tr}(\Sigma_\nu^2)-2\sum_{k=1}^n\Big(\lambda_{\mu,k}\lambda_{\nu,k}-\frac{\varepsilon}{4}\Big)^+-\frac{\varepsilon}{2}\sum_{k=1}^n\Big[\log(\lambda_{\mu,k}\lambda_{\nu,k})-\log\Big(\frac{\varepsilon}{4}\Big)\Big]^+.$$

Additionally, we have

$$K_{\mu\nu}^*=\operatorname{diag}\Big(\Big\{\lambda_{\mu,k}\lambda_{\nu,k}\Big[1-\frac{\varepsilon}{4\lambda_{\mu,k}\lambda_{\nu,k}}\Big]^+\Big\}^{\frac{1}{2}}\Big)_{k=1}^n.$$

As a consequence, the proof is completed.

When Σ_μ and Σ_ν are not diagonal: The value of $\mathsf{IGW}_\varepsilon(\mu,\nu)$ still remains according to Lemma 2.3 but the optimal transport plan does not. More specifically, the equality conditions mentioned in part (c) of Lemma 3.4 will be no longer valid. Instead, let P_μ, D_μ and P_ν, D_ν are the orthogonal diagonalizations of $\Sigma_\mu(=P_\mu D_\mu P_\mu^\top)$ and $\Sigma_\nu(=P_\nu D_\nu P_\nu^\top)$,

respectively. By using the same approach as in Lemma 3.4, let $U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top$ be the SVD of matrix $\Sigma_\mu^{-\frac{1}{2}}K_{\mu\nu}\Sigma_\nu^{-\frac{1}{2}}$ where $\Lambda_{\mu\nu}^{\frac{1}{2}} = \mathrm{diag}([\kappa_{\mu\nu,k}]^{\frac{1}{2}})_{k=1}^n$, we have $K_{\mu\nu} = \Sigma_\mu^{\frac{1}{2}}U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top\Sigma_\nu^{\frac{1}{2}}$. Therefore,

$$\begin{aligned}
\mathrm{tr}(K_{\mu\nu}^\top K_{\mu\nu}) &= \mathrm{tr}\left(\Sigma_\nu^{\frac{1}{2}}V_{\mu\nu}(\Lambda_{\mu\nu}^{\frac{1}{2}})^\top U_{\mu\nu}^\top\Sigma_\mu U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top\Sigma_\nu^{\frac{1}{2}}\right)\\
&= \mathrm{tr}\left(V_{\mu\nu}(\Lambda_{\mu\nu}^{\frac{1}{2}})^\top U_{\mu\nu}^\top\Sigma_\mu U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top\Sigma_\nu\right)\\
&= \mathrm{tr}\left(V_{\mu\nu}(\Lambda_{\mu\nu}^{\frac{1}{2}})^\top U_{\mu\nu}^\top P_\mu D_\mu P_\mu^\top U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top P_\nu D_\nu P_\nu^\top\right)\\
&= \mathrm{tr}\left([P_\nu^\top V_{\mu\nu}(\Lambda_{\mu\nu}^{\frac{1}{2}})^\top][U_{\mu\nu}^\top P_\mu D_\mu][P_\mu^\top U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}][V_{\mu\nu}^\top P_\nu D_\nu]\right)\\
&\leq \sum_{i=1}^n \lambda_{\mu,i}\lambda_{\nu,i}\kappa_{\mu\nu,i}.
\end{aligned}$$

The equality occurs when $U_{\mu\nu} = P_\mu$ and $V_{\mu\nu} = P_\nu$. Hence, the covariance matrix $K_{\mu\nu}^*$ in this case turns into

$$K_{\mu\nu}^{*,\mathrm{new}} = \Sigma_\mu^{\frac{1}{2}}P_\mu\Lambda_{\mu\nu}^{\frac{1}{2}}P_\nu^\top\Sigma_\nu^{\frac{1}{2}} = P_\mu D_\mu^{\frac{1}{2}}\Lambda_{\mu\nu}^{\frac{1}{2}}D_\nu^{\frac{1}{2}}P_\nu = P_\mu K_{\mu\nu}^* P_\nu^\top.$$

A.3. Proof of Lemma 3.3

Let q_v and p_u be the probability density functions of distributions Q_v and P_u, respectively, whereas q_u be the probability density function of the Gaussian distribution Q_u in $\mathbb{R}^d$ which has mean u and covariance matrix Σ_u. Then, we have

$$\begin{aligned}
\mathrm{KL}(P_u\|Q_v) &= \int_{\mathbb{R}^d} p_u\log\frac{p_u}{q_v}\mathrm{d}x = \int_{\mathbb{R}^d} p_u\log\frac{p_u}{q_u}\mathrm{d}x + \int_{\mathbb{R}^d} p_u\log\frac{q_u}{q_v}\mathrm{d}x\\
&= \mathrm{KL}(P_u\|Q_u) + \int_{\mathbb{R}^d} p_u\log\frac{q_u}{q_v}\mathrm{d}x\\
&\geq \int_{\mathbb{R}^d} p_u\log\frac{q_u}{q_v}\mathrm{d}x.
\end{aligned}$$

The equality happens when $P_u = Q_u$. Now we prove the lower bound is constant.

Note that

$$\begin{aligned}
\int_x p_u\log q_v &= \mathbb{E}_{p_u}\big[\log(q_v)\big] = \mathbb{E}_{p_u}\Big[-\frac{1}{2}\log\left[(2\pi)^d\det(\Sigma_v)\right] - \frac{1}{2}(x-v)^\top\Sigma_v^{-1}(x-v))\Big]\\
&= -\frac{1}{2}\mathbb{E}_{p_u}\Big[(x-v)^\top\Sigma_v^{-1}(x-v)\Big] + \mathsf{const}\\
&= -\frac{1}{2}\mathbb{E}_{p_u}\Big[(x-u+u-v)^\top\Sigma_v^{-1}(x-u+u-v)\Big] + \mathsf{const}\\
&= -\frac{1}{2}\mathbb{E}_{p_u}\Big[(x-u)^\top\Sigma_v^{-1}(x-u) + 2(x-u)^\top\Sigma_v^{-1}(u-v) + (u-v)^\top\Sigma_v^{-1}(u-v)\Big] + \mathsf{const}\\
&= -\frac{1}{2}\mathbb{E}_{p_u}\Big[(x-u)^\top\Sigma_v^{-1}(x-u)\Big] + \mathsf{const}\\
&= -\frac{1}{2}\mathbb{E}_{p_u}\Big[(x-u)^\top\Sigma_v^{-\frac{1}{2}}\Sigma_v^{-\frac{1}{2}}(x-u)\Big] + \mathsf{const}\\
&= -\frac{1}{2}\mathbb{E}_{p_u}\Big[\mathrm{tr}\left(\Sigma_v^{-\frac{1}{2}}(x-u)(x-u)^\top\Sigma_v^{-\frac{1}{2}}\right)\Big] + \mathsf{const}\\
&= -\frac{1}{2}\mathrm{tr}\left(\Sigma_v^{-\frac{1}{2}}\Sigma_u\Sigma_v^{-\frac{1}{2}}\right) + \mathsf{const}.
\end{aligned} \tag{21}$$

Similarly, we have $\mathbb{E}_{p_u}\big[\log(q_u)\big]$ is a constant. Hence, the proof is completed.

A.4. Proof of Lemma 3.4

(a) We start with decomposing the matrix $K_{\mu\nu}$ as $K_{\mu\nu} = \Sigma_\mu^{\frac{1}{2}} U_{\mu\nu} \Lambda_{\mu\nu}^{\frac{1}{2}} V_{\mu\nu}^\top \Sigma_\nu^{\frac{1}{2}}$. It follows from the fact Σ_π is non-negative definite that

$$\begin{aligned}
&\Sigma_\nu \succeq K_{\mu\nu}^\top \Sigma_\mu^{-1} K_{\mu\nu} \\
\Leftrightarrow\ &\Sigma_\nu \succeq \Sigma_\nu^{\frac{1}{2}} V_{\mu\nu} (\Lambda_{\mu\nu}^{\frac{1}{2}})^\top U_{\mu\nu}^\top U_{\mu\nu} \Lambda_{\mu\nu}^{\frac{1}{2}} V_{\mu\nu}^\top \Sigma_\nu^{\frac{1}{2}} \\
\Leftrightarrow\ &\Sigma_\nu \succeq \Sigma_\nu^{\frac{1}{2}} V_{\mu\nu} \Lambda_{\mu\nu} V_{\mu\nu}^\top \Sigma_\nu^{\frac{1}{2}} \\
\Leftrightarrow\ &\mathrm{Id}_n \succeq V_{\mu\nu} \Lambda_{\mu\nu} V_{\mu\nu}^\top,
\end{aligned}$$

where $\Lambda_{\mu\nu} = \mathrm{diag}\big(\kappa_{\mu\nu,k}\big)_{k=1}^n$ is an $n \times n$ matrix. Since $V_{\mu\nu}$ is an unitary matrix, all eigenvalues of matrix $\Lambda_{\mu\nu}$, which are $(\kappa_{\mu\nu,k})_{k=1}^n$, belong to the interval $[0, 1]$.

(b) By the definition of matrix Σ_π, we have

$$\begin{aligned}
\det(\Sigma_\pi) &= \det(\Sigma_\mu) \det\big(\Sigma_\nu - K_{\mu\nu}^\top \Sigma_\mu^{-1} K_{\mu\nu}\big) \\
&= \det(\Sigma_\mu) \det(\Sigma_\nu) \det\Big(\mathrm{Id}_n - V_{\mu\nu} \Lambda_{\mu\nu} V_{\mu\nu}^\top\Big) \\
&= \det(\Sigma_\mu) \det(\Sigma_\nu) \prod_{k=1}^n (1 - \kappa_{\mu\nu,k}).
\end{aligned}$$

The third equality results from the fact that all eigenvalues of matrix $\mathrm{Id}_n - V_{\mu\nu} \Lambda_{\mu\nu} V_{\mu\nu}^\top$ are $1 - \kappa_{\mu\nu,k}$ for $k \in [n]$.

(c) Using the decomposition of matrix $K_{\mu\nu}$ in part (a), $\mathrm{tr}\left(K_{\mu\nu}^\top K_{\mu\nu}\right)$ can be rewritten as

$$\mathrm{tr}\left(\Sigma_\nu^{\frac{1}{2}} V_{\mu\nu} (\Lambda_{\mu\nu}^{\frac{1}{2}})^\top U_{\mu\nu}^\top \Sigma_\mu U_{\mu\nu} \Lambda_{\mu\nu}^{\frac{1}{2}} V_{\mu\nu}^\top \Sigma_\nu^{\frac{1}{2}}\right) = \mathrm{tr}\left(\left[V_{\mu\nu} (\Lambda_{\mu\nu}^{\frac{1}{2}})^\top\right] \left[U_{\mu\nu}^\top \Sigma_\mu\right] \left[U_{\mu\nu} \Lambda_{\mu\nu}^{\frac{1}{2}}\right] \left[V_{\mu\nu}^\top \Sigma_\nu\right]\right).$$

Applying the generalization of the von Neumann's inequality (Kristof, 1969; Horn & Johnson, 1991) for singular values with a note that the equality happens when $U_{\mu\nu}$ and $V_{\mu\nu}^\top$ are identity matrices, we obtain

$$\mathrm{tr}\left(K_{\mu\nu}^\top K_{\mu\nu}\right) \leq \sum_{i=1}^n \lambda_{\mu,i} \lambda_{\nu,i} \kappa_{\mu\nu,i}.$$

Hence, we obtain the conclusion of this lemma. It is worth noting that the equality for the above inequality can also occur in other cases, for example, when $U_{\mu\nu} = \mathrm{Id}_m, V_{\mu\nu} = -\mathrm{Id}_n$ or when $U_{\mu\nu} = -\mathrm{Id}_m, V_{\mu\nu} = \mathrm{Id}_n$.

A.5. Proof of Lemma 4.1

By the formula for KL divergence between Gaussians in Lemma C.3 in Appendix C and applying the von Neumann's trace inequality (Kristof, 1969; Horn & Johnson, 1991), we have

$$\begin{aligned}
\mathrm{KL}(\overline{\pi}_x \| \overline{\mu}) &= \frac{1}{2} \left\{ \mathrm{tr}\left(\Sigma_x \Sigma_\mu^{-1}\right) - m + \log\left(\frac{\det(\Sigma_\mu)}{\det(\Sigma_x)}\right) \right\} \\
&\geq \frac{1}{2} \sum_{i=1}^m \left\{ \frac{\lambda_{x,i}}{\lambda_{\mu,i}} - \log\left(\frac{\lambda_{x,i}}{\lambda_{\mu,i}}\right) - 1 \right\}.
\end{aligned}$$

Similarly, we find that

$$\mathrm{KL}(\overline{\pi}_y \| \overline{\nu}) \geq \frac{1}{2} \sum_{j=1}^n \left\{ \frac{\lambda_{y,j}}{\lambda_{\nu,j}} - \log\left(\frac{\lambda_{y,j}}{\lambda_{\nu,j}}\right) - 1 \right\}.$$

For the last equality, $\mathrm{KL}(\overline{\pi}\|\overline{\mu}\otimes\overline{\nu})$ is equal to

$$\begin{aligned}
&\frac{1}{2}\left\{\mathrm{tr}\left(\Sigma_\pi\Sigma_{\mu\otimes\nu}^{-1}\right) - (m+n) - \log\left(\frac{\det(\Sigma_\pi)}{\det(\Sigma_{\mu\otimes\nu})}\right)\right\}\\
=\ &\frac{1}{2}\Big\{\mathrm{tr}\left(\Sigma_x\Sigma_\mu^{-1}\right) + \mathrm{tr}\left(\Sigma_y\Sigma_\nu^{-1}\right) - (m+n) - \log\left(\frac{\det(\Sigma_x)\det(\Sigma_y)}{\det(\Sigma_\mu)\det(\Sigma_\nu)}\right) - \sum_{k=1}^{n}\log(1-\kappa_{xy,k})\Big\}\\
=\ &\mathrm{KL}(\overline{\pi}_x\|\overline{\mu}) + \mathrm{KL}(\overline{\pi}_y\|\overline{\nu}) - \frac{1}{2}\sum_{k=1}^{n}\log(1-\kappa_{xy,k}).
\end{aligned}$$

The first equality results from the form of the matrix $\Sigma_{\mu\otimes\nu} = \begin{pmatrix}\Sigma_\mu & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & \Sigma_\nu\end{pmatrix}$.

A.6. Proof of Theorem 4.2

For the sake of computing, we will firstly show that all quadratic KL terms in the objective function (4) can be transformed to their normal versions as follows:

$$\mathrm{KL}^{\otimes}(\pi_x\|\mu) = 2m_\pi^2\mathrm{KL}(\bar{\pi}_x\|\bar{\mu}) + \mathrm{KL}(m_\pi^2\|m_\mu^2), \tag{22}$$
$$\mathrm{KL}^{\otimes}(\pi_y\|\nu) = 2m_\pi^2\mathrm{KL}(\bar{\pi}_y\|\bar{\nu}) + \mathrm{KL}(m_\pi^2\|m_\nu^2), \tag{23}$$
$$\mathrm{KL}^{\otimes}(\pi\|\mu\otimes\nu) = 2m_\pi^2\mathrm{KL}(\bar{\pi}\|\overline{\mu}\otimes\overline{\nu}) + \mathrm{KL}\big(m_\pi^2\|m_\mu^2 m_\nu^2\big). \tag{24}$$

By applying parts (ii) and (i) of Lemma C.1 in that order, we get the equation (22)

$$\begin{aligned}
\mathrm{KL}^{\otimes}(\pi_x\|\mu) &= \mathrm{KL}^{\otimes}(m_\pi\bar{\pi}_x\|m_\mu\bar{\mu})\\
&= m_\pi^2\mathrm{KL}^{\otimes}(\bar{\pi}_x\|\bar{\mu}) + m_\pi^2\log\left(\frac{m_\pi^2}{m_\mu^2}\right) + (m_\mu^2 - m_\pi^2)\\
&= 2m_\pi^2\mathrm{KL}(\bar{\pi}_x\|\bar{\mu}) + \mathrm{KL}(m_\pi^2\|m_\mu^2).
\end{aligned}$$

The equations (23) and (24) are obtained similarly. As a result, the objective function in equation (4) can be rewritten as

$$\begin{aligned}
&m_\pi^2\mathbb{E}_{\bar{\pi}\otimes\bar{\pi}}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} + \varepsilon\Big\{2m_\pi^2\mathrm{KL}(\bar{\pi}\|\overline{\mu}\otimes\overline{\nu}) + \mathrm{KL}\big(m_\pi^2\|m_\mu^2 m_\nu^2\big)\Big\}\\
&\qquad + \tau\Big\{2m_\pi^2\mathrm{KL}(\bar{\pi}_x\|\bar{\mu}) + \mathrm{KL}(m_\pi^2\|m_\mu^2) + 2m_\pi^2\mathrm{KL}(\bar{\pi}_y\|\bar{\nu}) + \mathrm{KL}(m_\pi^2\|m_\nu^2)\Big\}\\
=\ &m_\pi^2\Upsilon + \varepsilon\mathrm{KL}(m_\pi^2\|m_\mu^2 m_\nu^2) + \tau\Big\{\mathrm{KL}(m_\pi^2\|m_\mu^2) + \mathrm{KL}(m_\pi^2\|m_\nu^2)\Big\},
\end{aligned} \tag{25}$$

in which Υ is defined as

$$\Upsilon := \mathbb{E}_{\bar{\pi}\otimes\bar{\pi}}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} + 2\varepsilon\mathrm{KL}(\bar{\pi}\|\bar{\mu}\otimes\bar{\nu}) + 2\tau\Big\{\mathrm{KL}(\bar{\pi}_x\|\bar{\mu}) + \mathrm{KL}(\bar{\pi}_y\|\bar{\nu})\Big\}. \tag{26}$$

Now, we will divide the rest of the proof into two parts: shape optimization and mass optimization.

Shape optimization. The problem of interest now is $\min_{\bar{\pi}}\Upsilon$. We will first prove that the optimal shape $\bar{\pi}$ is a zero-mean Gaussian measure. Denote by (u,v) and Σ_π the mean vector and covariance matrix of $\bar{\pi}$ where $u\in\mathbb{R}^m$ and $v\in\mathbb{R}^n$. Let $Z = X-u, Z' = X'-u$ and $T = Y-v, T' = Y'-v$. According to Lemma C.7, we have

$$\begin{aligned}
&\mathbb{E}_{\bar{\pi}\otimes\bar{\pi}}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\}\\
=&\mathbb{E}_{\bar{\pi}\otimes\bar{\pi}}\Big\{\big[\langle Z,Z'\rangle - \langle T,T'\rangle\big]^2\Big\} + 2u^\top\Sigma_x u - 4u^\top\Sigma_{xy}v + 2v^\top\Sigma_y v + \Big[\|u\|^2 - \|v\|^2\Big]^2\\
=&\,\mathrm{tr}(\Sigma_x^2) + \mathrm{tr}(\Sigma_y^2) - 2\,\mathrm{tr}(K_{xy}^\top K_{xy}) + 2u^\top\Sigma_x u - 4u^\top\Sigma_{xy}v + 2v^\top\Sigma_y v + \Big[\|u\|^2 - \|v\|^2\Big]^2.
\end{aligned}$$

The second equality is obtained by using the same arguments for deriving equation (19). Let us fix the covariance matrix and mean vector of $\bar{\pi}$, therefore, the value of the expectation term in equation (26) is also fixed. Then, Lemma 3.3 indicates that $\bar{\pi}$ needs to be a Gaussian measure. while Lemma C.3 forces its mean to be zero.

Consequently, the problem $\min_{\bar{\pi}} \Upsilon$ is reduced to

$$\Upsilon^* := \min_{\Sigma_\pi} \Big\{ \operatorname{tr}(\Sigma_x^2) + \operatorname{tr}(\Sigma_y^2) - 2\operatorname{tr}(K_{xy}^\top K_{xy}) + 2\varepsilon \mathrm{KL}(\bar{\pi} \| \bar{\mu} \otimes \bar{\nu}) + 2\tau \mathrm{KL}(\bar{\pi}_x \| \bar{\mu}) + 2\tau \mathrm{KL}(\bar{\pi}_y \| \bar{\nu}) \Big\} \tag{27}$$

Using the same arguments as in Theorem 3.1, we obtain that the minimum value of

$$\operatorname{tr}(\Sigma_x^2) + \operatorname{tr}(\Sigma_y^2) - 2\operatorname{tr}(K_{xy}^\top K_{xy}) + 2\varepsilon \mathrm{KL}(\bar{\pi} \| \bar{\mu} \otimes \bar{\nu})$$

is equal to

$$\|\lambda_x\|_2^2 + \|\lambda_y\|_2^2 - 2\sum_{k=1}^n \Big(\lambda_{x,k}\lambda_{y,k} - \frac{\varepsilon}{2}\Big)^+ + \varepsilon \sum_{k=1}^n \Big[\log(\lambda_{x,k}\lambda_{y,k}) - \log\frac{\varepsilon}{2}\Big]^+.$$

Combining this result with parts (a) and (b) of Lemma 4.1 implies that the problem (27) is equivalent to

$$\Upsilon^* = \min_{\lambda_x, \lambda_y > 0} \Big\{ \sum_{k=1}^n g_{\varepsilon,\tau,+}(\lambda_{x,k}, \lambda_{y,k}; \lambda_{\mu,k}, \lambda_{\nu,k}) + \sum_{k=n+1}^m h_{\varepsilon,\tau}(\lambda_{x,k}; \lambda_{\mu,k}) \Big\},$$

with note that $(\lambda_{x,k})_{k=1}^m$ and $(\lambda_{y,k})_{k=1}^n$ are decreasing sequences. By Lemma C.4, each summation term of the above problem can be optimized independently but still preserving the order of $(\lambda_{x,k})_{k=1}^n$ as follows:

$$\Upsilon^* = \sum_{k=1}^n \min_{\lambda_{x,k}, \lambda_{y,k} > 0} g_{\varepsilon,\tau,+}\big(\lambda_{x,k}, \lambda_{y,k}; \lambda_{\mu,i}, \lambda_{\nu,j}\big) + \sum_{k=n+1}^m \min_{\lambda_{x,k} > 0} h_{\varepsilon,\tau}\big(\lambda_{x,k}, \lambda_{\mu,k}\big).$$

A detailed calculation of Υ^* is deferred to Appendix B.

Mass optimization. Finally, based on equation (25), the optimal mass is obtained by taking the square root of the minimizer of the following problem:

$$m_{\pi^*}^2 = \arg\min_{x>0} \; f(x) := \Upsilon^* x + \varepsilon \mathrm{KL}(x \| m_\mu^2 m_\nu^2) + \tau \mathrm{KL}(x \| m_\mu^2) + \tau \mathrm{KL}(x \| m_\nu^2).$$

By part (c) of Lemma C.6, we obtain $m_{\pi^*} = (x^*)^{\frac{1}{2}} = (m_\mu m_\nu)^{\frac{\tau+\varepsilon}{2\tau+\varepsilon}} \exp\Big\{\frac{-\Upsilon^*}{2(2\tau+\varepsilon)}\Big\}$. Hence, the proof is completed.

A.7. Proof of Theorem 5.2

Firstly, let $K_{\ell y}$ be the covariance matrix between X_ℓ and Y while v and Σ_y be the mean and covariance matrix of Y, respectively. By Lemma C.7, the objective function in equation (13) is rewritten as

$$\sum_{\ell=1}^T \alpha_\ell \mathbb{E}_{\pi_{\ell,y} \otimes \pi_{\ell,y}} \Big\{ \big[\langle X_\ell, X'_\ell \rangle - \langle Y, Y' \rangle\big]^2 \Big\} = \sum_{\ell=1}^T \alpha_\ell \Big\{ \mathbb{E}_{\pi_{\ell,y} \otimes \pi_{\ell,y}} \big[\langle X_\ell, X'_\ell \rangle - \langle Y - v, Y' - v \rangle\big]^2 + 2v^\top \Sigma_y v + \|v\|^4 \Big\}.$$

As Σ_y is a positive definite matrix, we have $2v^\top \Sigma_y v + \|v\|^4 > 0$ when $v \neq \mathbf{0}_d$. Therefore, the barycenter μ^* must have mean zero. Next, we denote by $\lambda_{\ell,i}$ and $\lambda_{y,i}$ the ith largest eigenvalues of Σ_ℓ and Σ_y, respectively. By using the same arguments for deriving equation (19), we get

$$\sum_{\ell=1}^T \alpha_\ell \mathbb{E}_{\pi_{\ell,y} \otimes \pi_{\ell,y}} \Big\{ \big[\langle X_\ell, X'_\ell \rangle - \langle Y, Y' \rangle\big]^2 \Big\} = \sum_{\ell=1}^T \alpha_\ell \Big\{ \sum_{i=1}^d \lambda_{y,i}^2 - 2\operatorname{tr}\big(K_{\ell y}^\top K_{\ell y}\big) + \sum_{j=1}^{m_\ell} \lambda_{\ell,j}^2 \Big\}.$$

According to parts (a) and (c) of Lemma 3.4, we have

$$\operatorname{tr}(K_{\ell y}^\top K_{\ell y}) \leq \sum_{j=1}^{\min\{d, m_\ell\}} \lambda_{y,j} \lambda_{\ell,j},$$

where the equality happens when $(X_\ell)_j = \sqrt{\lambda_{\ell,j}} Z_j$ and $Y_j = \sqrt{\lambda_{y,j}} Z_j$ with $(Z_j)_j$ are i.i.d. standard normal random variables. Thus, the problem now is reduced to minimize

$$\sum_{\ell=1}^{T} \alpha_\ell \Big\{ \sum_{j=1}^{d} \lambda_{y,j}^2 - 2 \sum_{j=1}^{\min\{d,m_\ell\}} \lambda_{y,j} \lambda_{\ell,j} \Big\} = \sum_{j=1}^{d} \Big\{ \lambda_{y,j}^2 - 2\lambda_{y,j} \sum_{\ell=1}^{T} \alpha_\ell \lambda_{\ell,j} \mathbf{1}_{j \le m_\ell} \Big\}.$$

This quantity has a quadratic form, so it attains its minimum at

$$\lambda_{y,j} = \sum_{\ell=1}^{T} \alpha_\ell \lambda_{\ell,j} \mathbf{1}_{j \le m_\ell}; \quad j \in [d].$$

Hence, we obtain the conclusion of the theorem.

A.8. Proof of Theorem 5.4

Firstly, we will show that the barycenter μ^* has mean zero. Let $K_{\ell y}$ be the covariance matrix between X_ℓ and Y while v and Σ_y be the mean and covariance matrix of Y, respectively. By Lemma C.7, the objective function in equation (14) can be rewritten as

$$\begin{aligned}
&\sum_{\ell=1}^{T} \alpha_\ell \Big\{ \mathbb{E}_{\pi_{\ell,y} \otimes \pi_{\ell,y}} \big[\langle X_\ell, X'_\ell \rangle - \langle Y, Y' \rangle \big]^2 + \varepsilon \mathrm{KL}(\pi_{\ell,y} \| \mu_\ell \otimes \mu) \Big\} \\
=&\sum_{\ell=1}^{T} \alpha_\ell \Big\{ \mathbb{E}_{\pi_{\ell,y} \otimes \pi_{\ell,y}} \big[\langle X_\ell, X'_\ell \rangle - \langle Y - v, Y' - v \rangle \big]^2 + 2v^\top \Sigma_y v + \|v\|^4 \Big\} + \varepsilon \mathrm{KL}(\pi_{\ell,y} \| \mu_\ell \otimes \mu).
\end{aligned}$$

Note that $\pi_{\ell,y}$ and $\mu_\ell \otimes \mu$ share the same mean and KL divergence is invariant to translation, which make the quantity $\mathrm{KL}(\pi_{\ell,y} \| \mu_\ell \otimes \mu)$ become independent of the mean vector of $\pi_{\ell,y}$. Additionally, as Σ_y is a positive definite matrix, we have $2v^\top \Sigma_y v + \|v\|^4 > 0$ when $v \neq \mathbf{0}_d$. Therefore, the barycenter $\mu*$ must have mean zero. Moreover, according to Lemma 5.6, there exists a Gaussian solution, says μ^*. Then, the objective function is reduced to

$$\sum_{\ell=1}^{T} \alpha_\ell \Big\{ \sum_{i=1}^{d} \lambda_{y,i}^2 - 2 \operatorname{tr}\left(K_{\ell y}^\top K_{\ell y}\right) + \sum_{j=1}^{m_\ell} \lambda_{\ell,j}^2 \Big\} + \varepsilon \mathrm{KL}(\pi_{\ell,y} \| \mu_\ell \otimes \mu).$$

Denote by $\Sigma_{\ell,y}$ and $\Sigma_{\ell \otimes y}$ the covariance matrices of $\pi_{\ell,y}$ and $\mu_\ell \otimes \mu$, respectively. The entropic term in the above equation is then equal to

$$\frac{\varepsilon}{2} \left\{ \operatorname{tr}(\Sigma_{\ell,y} \Sigma_{\ell \otimes y}^{-1}) + \log\left(\frac{\det(\Sigma_{\ell \otimes y})}{\det(\Sigma_{\ell,y})} \right) - (m_\ell + d) \right\} = \frac{\varepsilon}{2} \log\left(\frac{\det(\Sigma_{\ell \otimes y})}{\det(\Sigma_{\ell,y})} \right).$$

The problem is thus reduced to minimize

$$\sum_{j=1}^{d} \lambda_{y,j}^2 - 2 \sum_{\ell=1}^{T} \alpha_\ell \operatorname{tr}\left(K_{\ell y}^\top K_{\ell y}\right) + \frac{1}{2} \sum_{\ell=1}^{T} \alpha_\ell \varepsilon \log\left(\frac{\det(\Sigma_{\ell \otimes y})}{\det(\Sigma_{\ell,y})} \right). \tag{28}$$

Use the same approach of Lemma 3.4, let $U_{\ell y} \Lambda_{\ell y}^{\frac{1}{2}} V_{\ell y}^\top$ be the SVD of matrix $\Sigma_\ell^{-\frac{1}{2}} K_{\ell y} \Sigma_y^{-\frac{1}{2}}$, or equivalently,

$$K_{\ell y} = \Sigma_\ell^{\frac{1}{2}} U_{\ell y} \Lambda_{\ell y}^{\frac{1}{2}} V_{\ell y}^\top \Sigma_y^{\frac{1}{2}}.$$

Denote by $\kappa_{\ell y,j}^{\frac{1}{2}}$ the j-th largest singular value of $\Sigma_\ell^{-\frac{1}{2}} K_{\ell y} \Sigma_y^{-\frac{1}{2}}$. Then, we find that

$$\log\left(\frac{\det(\Sigma_{\ell \otimes y})}{\det(\Sigma_{\ell,y})} \right) = -\sum_{j=1}^{d_\ell} \log(1 - \kappa_{\ell y,j}).$$

By applying the von Neumann's trace inequality (Kristof, 1969; Horn & Johnson, 1991), we have

$$\begin{aligned}\operatorname{tr}(K_{\ell y}^\top K_{\ell y}) &= \operatorname{tr}\left(\Sigma_y^{\frac{1}{2}} V_{\ell y} (\Lambda_{\ell y}^{\frac{1}{2}})^\top U_{\ell y}^\top \Sigma_\ell U_{\ell y} \Lambda_{\ell y} V_{\ell y}^\top \Sigma_y^{\frac{1}{2}}\right) = \operatorname{tr}\left([V_{\ell y}(\Lambda_{\ell y}^{\frac{1}{2}})^\top] \times [U_{\ell y}^\top \Sigma_\ell] \times [U_{\ell y}\Lambda_{\ell y}^{\frac{1}{2}}] \times [V_{\ell y}^\top \Sigma_y]\right) \\ &\leq \sum_{j=1}^{d_\ell} \lambda_{\ell,j}\lambda_{y,j}\kappa_{\ell y,j},\end{aligned}$$

with a note that all $\Lambda_{\ell y}$, Σ_y and Σ_ℓ are diagonal matrices. The equality occurs when $V_{\ell y}$ and $U_{\ell y}$ are identity matrices.

Put the above results together, the problem in equation (28) is reduced to minimizing

$$\sum_{j=1}^{d} \lambda_{y,j}^2 - \sum_{\ell=1}^{T} 2\alpha_\ell \sum_{t=1}^{d_\ell} \lambda_{\ell,t}\kappa_{\ell y,t}\lambda_{y,t} - \frac{\varepsilon}{2}\sum_{\ell=1}^{T}\alpha_\ell \sum_{t=1}^{d_\ell}\log(1-\kappa_{\ell y,t}),$$

or equivalently,

$$\sum_{j=1}^{d}\Big\{\lambda_{y,j}^2 - 2\lambda_{y,j}\sum_{\ell=1}^{T}\alpha_\ell\lambda_{\ell,j}\kappa_{\ell y,j}\mathbf{1}_{j\leq d_\ell}\Big\} - \frac{\varepsilon}{2}\sum_{\ell=1}^{T}\alpha_\ell\sum_{t=1}^{d_\ell}\log(1-\kappa_{\ell y,t}).$$

By fixing the values of $\kappa_{\ell y,t} \in [0,1]$ for all $t \in [d_\ell]$, this quantity attains its minimum at

$$\lambda_{y,j} = \sum_{\ell=1}^{T}\alpha_\ell\lambda_{\ell,j}\kappa_{\ell y,j}\mathbf{1}_{j\leq d_\ell}, \quad j \in [d].$$

Therefore, the problem of entropic IGW barycenter is equivalent to finding the maximum of

$$\begin{aligned}&\sum_{j=1}^{d}\Big[\sum_{\ell=1}^{T}\alpha_\ell\lambda_{\ell,j}\kappa_{\ell y,j}\mathbf{1}_{j\leq d_\ell}\Big]^2 + \frac{\varepsilon}{2}\sum_{\ell=1}^{T}\sum_{j=1}^{d_\ell}\alpha_\ell\log(1-\kappa_{\ell y,j}) \\ &= \sum_{j=1}^{d}\Big\{\Big[\sum_{\ell=1}^{T}\alpha_\ell\lambda_{\ell,j}\kappa_{\ell y,j}\mathbf{1}_{j\leq d_\ell}\Big]^2 + \frac{\varepsilon}{2}\sum_{\ell=1}^{T}\alpha_\ell\log(1-\kappa_{\ell y,j})\mathbf{1}_{j\leq d_\ell}\Big\}.\end{aligned}$$

We need to maximize each term, which is

$$\Big[\sum_{\ell=1}^{T}\alpha_\ell\lambda_{\ell,j}\kappa_{\ell y,j}\mathbf{1}_{j\leq d_\ell}\Big]^2 + \sum_{\ell=1}^{T}\frac{\varepsilon}{2}\alpha_\ell\log(1-\kappa_{\ell y,j})\mathbf{1}_{j\leq d_\ell},$$

under the constraints that $0 \leq \kappa_{\ell y,j} \leq 1$, for all $j \in [d_\ell]$. By Lemma C.5 in Appendix C, and when ε satisfying the condition (15) for $j \leq \min\{d, m_\ell\}$, we have

$$\kappa_{\ell y,j} = 1 - \frac{\varepsilon}{2\lambda_{\ell,j}\Big\{A_j + \sqrt{A_j^2 - \varepsilon B_j}\Big\}},$$

where $A_j = \sum_{\ell=1}^{T}\alpha_\ell\lambda_{\ell,j}\mathbf{1}_{j\leq d_\ell}$ and $B_j = \sum_{\ell=1}^{T}\alpha_\ell\mathbf{1}_{j\leq d_\ell}$. Hence, we reach the conclusion of the theorem.

A.9. Proof of Lemma 5.6

Let $Q_{X,Y} = \mathcal{N}([\gamma_x,\gamma_y]^\top, \Sigma_{X,Y})$ and $Q_Y = \mathcal{N}(\gamma_y, \Sigma_y)$ be Gaussian distributions in $\mathbb{R}^{m+n}$ and $\mathbb{R}^n$, respectively. Denote by $q_X, q_Y, q_{X,Y}, p_Y$ and $p_{X,Y}$ the probability density functions of distributions $Q_X, Q_Y, Q_{X,Y}, P_Y$ and $P_{X,Y}$, respectively, we need to show that

$$\mathsf{KL}\big(P_{X,Y}\|Q_X\otimes P_Y\big) \geq \mathsf{KL}\big(Q_{X,Y}\|Q_X\otimes Q_Y\big).$$

Expanding the KL divergence, we need to prove

$$\int p_{X,Y}(x,y)\big[\log p_{X,Y}(x,y) - \log q_X(x) - \log p_Y(y)\big]dxdy \geq \int q_{X,Y}(x,y)\big[\log q_{X,Y}(x,y) - \log q_X(x) - \log q_Y(y)\big]dxdy.$$

Similar to the derivation for equation (21) in the proof Lemma 3.3 with a note that $q_{X,Y}$, q_X and q_Y are Gaussian distributions, we have

$$\begin{aligned}\int p_{X,Y}(x,y)\log q_X(x)dxdy &= \int q_{X,Y}(x,y)\log q_X(x)dxdy;\\ \int p_{X,Y}(x,y)\log q_Y(y)dxdy &= \int q_{X,Y}(x,y)\log q_Y(y)dxdy;\\ \int p_{X,Y}(x,y)\log q_{X,Y}(x,y)dxdy &= \int q_{X,Y}(x,y)\log q_{X,Y}(x,y)dxdy,\end{aligned}$$

since both sides are equal to the same function of the covariance matrices of $P_{X,Y}$ and $Q_{X,Y}$, which are both equal to $\Sigma_{X,Y}$. Hence, the inequality is equivalent to

$$\begin{aligned}&\int p_{X,Y}(x,y)\big[\log p_{X,Y}(x,y) - \log p_Y(y)\big]dxdy \geq \int p_{X,Y}(x,y)\big[\log q_{X,Y}(x,y) - \log q_Y(y)\big]dxdy\\ &\quad\Leftrightarrow \int p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)/p_Y(y)}{q_{X,Y}(x,y)/q_Y(y)}dxdy \geq 0.\end{aligned}$$

By using the formula for conditional distributions, we get

$$\begin{aligned}p_{X,Y}(x,y) &= p_Y(y)p_{X|Y}(x|y);\\ q_{X,Y}(x,y) &= q_Y(y)q_{X|Y}(x|y).\end{aligned}$$

Then the left hand side is equal to

$$\begin{aligned}\int p_Y(y)p_{X|Y}(x|y)\log\frac{p_{X|Y}(x|y)}{q_{X|Y}(x|y)}dxdy &= \int_y p_Y(y)\Big[\int_x p_{X|Y}(x|y)\log\frac{p_{X|Y}(x|y)}{q_{X|Y}(x|y)}dx\Big]dy\\ &= \int_y p_Y(y)\mathsf{KL}\big(P_{X|Y}\|Q_{X|Y}\big)dy\end{aligned}$$

which is not less than zero, since the KL term is non-negative. The inequality becomes equality when $P_{X|Y} = Q_{X|Y}$.

B. On detailed calculation of Υ^*

In this appendix, we discuss how to derive a closed-form formulation for Υ^* in Theorem 4.2. It is equivalent to finding the minimizers of functions $g_{\varepsilon,\tau,+}$ and $h_{\varepsilon,\tau}$ in Lemma B.1 and Lemma B.4, respectively.

Lemma B.1 (Minimizer of $g_{\varepsilon,\tau,+}$)**.** *Let*

$$g_{\varepsilon,\tau,+}(x,y;a,b) := x^2+y^2+(\tau+\varepsilon)\Big(\frac{x}{a}+\frac{y}{b}-\log\Big(\frac{xy}{ab}\Big)-2\Big) - 2\Big[xy-\frac{\varepsilon}{2}\Big]^+ + \varepsilon\Big[\log(xy)-\log\Big(\frac{\varepsilon}{2}\Big)\Big]^+.$$

We define

$$\begin{aligned}g_{\varepsilon,\tau,-1}(x,y;a,b) &:= x^2+y^2+(\tau+\varepsilon)\Big[\frac{x}{a}+\frac{y}{b}-\log\Big(\frac{xy}{ab}\Big)-2\Big],\\ g_{\varepsilon,\tau,1}(x,y;a,b) &:= x^2+y^2+(\tau+\varepsilon)\Big[\frac{x}{a}+\frac{y}{b}-\log\Big(\frac{xy}{ab}\Big)-2\Big] - 2(xy-\frac{\varepsilon}{2}) + \varepsilon\Big[\log(xy)-\log\Big(\frac{\varepsilon}{2}\Big)\Big],\\ (\widetilde{x},\widetilde{y}) &:= \operatorname*{arg\,min}_{x,y>0}\ g_{\varepsilon,\tau,-1}(x,y;a,b),\\ (\widehat{x},\widehat{y}) &:= \operatorname*{arg\,min}_{x,y>0}\ g_{\varepsilon,\tau,1}(x,y;a,b),\\ (x^*,y^*) &:= \operatorname*{arg\,min}_{x,y>0}\ g_{\varepsilon,\tau,+}(x,y,;a,b).\end{aligned}$$

Then if $\widetilde{x}\widetilde{y} < \frac{\varepsilon}{2}$, then $(x^*,y^*) = (\widetilde{x},\widetilde{y})$ where solutions are in Lemma B.2, and if otherwise, then $(x^*,y^*) = (\widehat{x},\widehat{y})$, where solutions are in Lemma B.3.

Proof of Lemma B.1. By definition of functions $g_{\varepsilon,\tau,1}$, $g_{\varepsilon,\tau,-1}$ and $g_{\varepsilon,\tau,+}$, we have

$$g_{\varepsilon,\tau,+}(x,y;a,b) = g_{\varepsilon,\tau,-1}(x,y;a,b)\mathbf{1}_{\{xy<\frac{\varepsilon}{2}\}} + g_{\varepsilon,\tau,1}(x,y;a,b)\mathbf{1}_{\{xy\geq\frac{\varepsilon}{2}\}}.$$

We observe that both functions $g_{\varepsilon,\tau,1}$ and $g_{\varepsilon,\tau,-1}$ are convex functions. Moreover, we have

$$\begin{aligned} g_{\varepsilon,\tau,-1}(x,y;a,b) - g_{\varepsilon,\tau,1}(x,y;a,b) &= 2\Big(xy - \frac{\varepsilon}{2}\Big) - \varepsilon\Big[\log(xy) - \log\Big(\frac{\varepsilon}{2}\Big)\Big] \\ &= \varepsilon\Big\{\frac{xy}{\varepsilon/2} - 1 - \log\Big(\frac{xy}{\varepsilon/2}\Big)\Big\} \geq 0. \end{aligned}$$

It means that $g_{\varepsilon,\tau,-1}(x,y;a,b) \geq g_{\varepsilon,\tau,1}(x,y;a,b)$.

Next, we prove that $\frac{\varepsilon}{2}$ cannot lie between $\widehat{x}\widehat{y}$ and $\widetilde{x}\widetilde{y}$. Assume the contradictory, then the segment connecting $(\widetilde{x},\widetilde{y})$ to $(\widehat{x},\widehat{y})$ cuts the hyperpole $xy = \frac{\varepsilon}{2}$ at a point with coordinates denoted by $(\overline{x},\overline{y})$. Since $g_{\varepsilon,\tau,1}$ is a convex function, we have

$$\max\Big\{g_{\varepsilon,\tau,1}(\widehat{x},\widehat{y};a,b), g_{\varepsilon,\tau,1}(\widetilde{x},\widetilde{y};a,b)\Big\} > g_{\varepsilon,\tau,1}(\overline{x},\overline{y};a,b). \tag{29}$$

Moreover, we find that

$$\begin{aligned} &g_{\varepsilon,\tau,1}(\overline{x},\overline{y};a,b) \geq g_{\varepsilon,\tau,1}(\widehat{x},\widehat{y};a,b) \\ &g_{\varepsilon,\tau,1}(\overline{x},\overline{y};a,b) = g_{\varepsilon,\tau,-1}(\overline{x},\overline{y};a,b) \geq g_{\varepsilon,\tau,-1}(\widetilde{x},\widetilde{y};a,b) \geq g_{\varepsilon,\tau,1}(\widetilde{x},\widetilde{y};a,b). \end{aligned}$$

It contradicts inequality (29). Thus, both $\widehat{x}\widehat{y}$ and $\widetilde{x}\widetilde{y}$ have to be greater or smaller than $\frac{\varepsilon}{2}$ at the same time. Given that, we have two separate cases:

Case 1: They are both greater than $\frac{\varepsilon}{2}$, then

$$g_{\varepsilon,\tau,-1}(x,y;a,b) \geq g_{\varepsilon,\tau,-1}(\widetilde{x},\widetilde{y};a,b) \geq g_{\varepsilon,\tau,1}(\widetilde{x},\widetilde{y};a,b) \geq g_{\varepsilon,\tau,1}(\widehat{x},\widehat{y};a,b).$$

It means that $\widehat{x} = x^*$ and $\widehat{y} = y^*$.

Case 2: Both of them are smaller than $\frac{\varepsilon}{2}$. For any (x,y) such that $xy \geq \frac{\varepsilon}{2}$, there exists $\big(\overline{x},\overline{y}\big)$ lies in the segment connecting (x,y) to $(\widehat{x},\widehat{y})$ such that $\overline{x}\,\overline{y} = \frac{\varepsilon}{2}$. Since the function $g_{\varepsilon,\tau,1}$ is convex, we find that

$$t g_{\varepsilon,\tau,1}(\widehat{x},\widehat{y};a,b) + (1-t) g_{\varepsilon,\tau,1}(x,y;a,b) \geq g_{\varepsilon,\tau,1}(\overline{x},\overline{y};a,b),$$

where $t = \frac{\|(\widehat{x},\widehat{y})-(\overline{x},\overline{y})\|_2}{\|(x,y)-(\widehat{x},\widehat{y})\|_2}$. However, $g_{\varepsilon,\tau,1}(\widehat{x},\widehat{y};a,b) \leq g_{\varepsilon,\tau,1}(\overline{x},\overline{y};a,b)$, which implies that

$$g_{\varepsilon,\tau,1}(x,y;a,b) \geq g_{\varepsilon,\tau,1}(\overline{x},\overline{y};a,b) = g_{\varepsilon,\tau,-1}(\overline{x},\overline{y};a,b) \geq g_{\varepsilon,\tau,-1}(\widetilde{x},\widetilde{y};a,b).$$

It means that $\widetilde{x} = x^*$ and $\widetilde{y} = y^*$. As a consequence, we obtain the conclusion of the lemma. □

Lemma B.2 (Minimizer of function $g_{\varepsilon,\tau,-1}$)**.** *With $a,b>0$, let*

$$g_{\varepsilon,\tau,-1}(x,y;a,b) = x^2 + y^2 + (\tau+\varepsilon)\left(\frac{x}{a} + \frac{y}{b} - \log(xy) - \log(ab)\right)$$

subjected to $x,y>0$. Let $(\widetilde{x},\widetilde{y}) = \arg\min_{x,y>0} g_{\varepsilon,\tau,-1}(x,y;a,b)$. Then,

$$\widetilde{x} = -\frac{\tau+\varepsilon}{4a} + \frac{1}{2}\sqrt{2(\tau+\varepsilon) + \frac{(\tau+\varepsilon)^2}{4a^2}}, \quad \widetilde{y} = -\frac{\tau+\varepsilon}{4b} + \frac{1}{2}\sqrt{2(\tau+\varepsilon) + \frac{(\tau+\varepsilon)^2}{4b^2}}.$$

Proof of Lemma B.2. 1. Taking the derivatives of $g_{\varepsilon,\tau,-1}$ with respect to x and y, we get

$$\frac{\partial g_{\varepsilon,\tau,-1}}{\partial x} = 2x + \frac{\tau+\varepsilon}{a} - \frac{\tau+\varepsilon}{x}, \quad \frac{\partial g_{\varepsilon,\tau,-1}}{\partial y} = 2y + \frac{\tau+\varepsilon}{b} - \frac{\tau+\varepsilon}{y}.$$

It follows that

$$\frac{\partial^2 g_{\varepsilon,\tau,-1}}{\partial x^2} = 2 + \frac{\tau+\varepsilon}{x^2}, \quad \frac{\partial^2 g_{\varepsilon,\tau,-1}}{\partial x \partial y} = 0, \quad \frac{\partial^2 g_{\varepsilon,\tau,-1}}{\partial y^2} = 2 + \frac{\tau+\varepsilon}{y^2}.$$

The Hessian matrix of function $g_{\varepsilon,\tau,-1}$ is positive definite, so $g_{\varepsilon,\tau,-1}$ is a convex function.

2. Solving the equations $\frac{\partial g_{\varepsilon,\tau,-1}}{\partial x} = 0$ and $\frac{\partial g_{\varepsilon,\tau,-1}}{\partial y} = 0$, we have

$$\widetilde{x} = -\frac{\tau+\varepsilon}{4a} + \frac{1}{2}\sqrt{2(\tau+\varepsilon) + \frac{(\tau+\varepsilon)^2}{4a^2}}, \quad \tilde{y} = -\frac{\tau+\varepsilon}{4b} + \frac{1}{2}\sqrt{2(\tau+\varepsilon) + \frac{(\tau+\varepsilon)^2}{4b^2}}.$$

Therefore, we reach the conclusion of the lemma. □

Lemma B.3 (Minimizer of function $g_{\varepsilon,\tau,1}$)**.** *For $a, b > 0$, let*

$$g_{\varepsilon,\tau,1}(x, y; a, b) = x^2 + y^2 + \varepsilon\Big[\log(xy) - \log\Big(\frac{\varepsilon}{2}\Big)\Big] - 2\Big(xy - \frac{\varepsilon}{2}\Big) + (\tau+\varepsilon)\left[\frac{x}{a} + \frac{y}{b} - \log\Big(\frac{xy}{ab}\Big) - 2\right],$$

subjected to $x, y > 0$. Define $(\widehat{x}, \widehat{y}) = \arg\min_{x,y>0} g_{\varepsilon,\tau,1}(x, y; a, b)$. Then, the value of $(\widehat{x}, \widehat{y})$ is given in the following equations.

$$\widehat{x} = -\frac{\tau+\varepsilon}{4a} + \frac{1}{2}\sqrt{\frac{(\tau+\varepsilon)^2}{4a^2} + 4\tau\Big(\frac{1}{2} + \frac{1}{\widehat{z}}\Big)}; \quad \widehat{y} = -\frac{\tau+\varepsilon}{4b} + \frac{1}{2}\sqrt{\frac{(\tau+\varepsilon)^2}{4b^2} + 4\tau\Big(\frac{1}{2} + \frac{1}{\widehat{z}}\Big)},$$

where $\widehat{z}$ is the unique positive root of the equation

$$\tau z^3 + \Big[8\tau - \frac{(\tau+\varepsilon)^2}{ab}\Big]z^2 + \Big[16\tau - 2(\tau+\varepsilon)^2\Big(\frac{1}{a} + \frac{1}{b}\Big)^2\Big]z - 4(\tau+\varepsilon)^2\Big(\frac{1}{a} + \frac{1}{b}\Big)^2 = 0.$$

Proof of Lemma B.3. Taking the first derivatives of $g_{\varepsilon,\tau,1}$ with respect to x and y , we have

$$\frac{\partial g_{\varepsilon,\tau,1}}{\partial x} = 2x - 2y + \frac{\tau+\varepsilon}{a} - \frac{\tau}{x}, \quad \frac{\partial g_{\varepsilon,\tau,1}}{\partial y} = 2y - 2x + \frac{\tau+\varepsilon}{b} - \frac{\tau}{y}.$$

It follows that

$$\frac{\partial^2 g_{\varepsilon,\tau,1}}{\partial x^2} = 2 + \frac{\tau}{2x^2}, \quad \frac{\partial^2 g_{\varepsilon,\tau,1}}{\partial x \partial y} = -2, \quad \frac{\partial^2 g_{\varepsilon,\tau,1}}{\partial y^2} = 2 + \frac{\tau}{y^2}.$$

Thus, the Hessian matrix of this function is positive definite, which implies that $g_{\varepsilon,\tau,1}$ is a convex function.

The equations of the stationary point give us

$$2\widehat{x} - 2\widehat{y} + \frac{\tau+\varepsilon}{a} - \frac{\tau}{\widehat{x}} = 0; \qquad 2\widehat{y} - 2\widehat{x} + \frac{\tau+\varepsilon}{b} - \frac{\tau}{\widehat{y}} = 0.$$

Taking the sum of both equations leads to

$$\frac{\tau+\varepsilon}{a} + \frac{\tau+\varepsilon}{b} = \frac{\tau}{\widehat{x}} + \frac{\tau}{\widehat{y}} \Leftrightarrow \tau\left(\frac{1}{\widehat{x}} + \frac{1}{\widehat{y}} - \frac{1}{a} - \frac{1}{b}\right) = \frac{\varepsilon}{a} + \frac{\varepsilon}{b}.$$

$$\begin{cases} 2(\widehat{x} - \widehat{y}) + \dfrac{\tau+\varepsilon}{a} = \dfrac{\tau}{\widehat{x}} \\ 2(\widehat{y} - \widehat{x}) + \dfrac{\tau+\varepsilon}{b} = \dfrac{\tau}{\widehat{y}}. \end{cases} \tag{30}$$

Taking the difference between two equations in system (30), we get

$$(\widehat{x} - \widehat{y})\Big(4 + \frac{\tau}{\widehat{x}\widehat{y}}\Big) = \frac{\tau+\varepsilon}{b} - \frac{\tau+\varepsilon}{a}.$$

Let $z = \frac{\tau}{\widehat{x}\widehat{y}}$, we have $\widehat{x} - \widehat{y} = (\tau+\varepsilon)\frac{1/b - 1/a}{4+z}$. Take the product of two equations in system (30), we have

$$\Big[2(\widehat{x}-\widehat{y}) + \frac{\tau+\varepsilon}{a}\Big]\Big[2(\widehat{y}-\widehat{x}) + \frac{\tau+\varepsilon}{b}\Big] = \frac{\tau^2}{\widehat{x}\widehat{y}}.$$

Substitute $\widehat{x} - \widehat{y} = (\tau+\varepsilon)\frac{1/b-1/a}{4+z}$ to the above equation, we obtain

$$(\tau+\varepsilon)^2\Big[2\Big(\frac{1}{a}+\frac{1}{b}\Big) + \frac{z}{a}\Big]\Big[2\Big(\frac{1}{a}+\frac{1}{b}\Big) + \frac{z}{b}\Big] = \tau z(z+4)^2,$$

which is equivalent to the following cubic equation

$$t(z) := \tau z^3 + \Big[8\tau - \frac{(\tau+\varepsilon)^2}{ab}\Big]z^2 + \Big[16\tau - 2(\tau+\varepsilon)^2\Big(\frac{1}{a}+\frac{1}{b}\Big)^2\Big]z - 4(\tau+\varepsilon)^2\Big(\frac{1}{a}+\frac{1}{b}\Big)^2 = 0.$$

Since it is a cubic equation, it has either three real roots or one real root. In the first case, we know that it has at least one positive root, since the highest coefficient of the equation is positive. Assume that it has other two positive roots. By applying Viete's theorem, we get

$$8\tau - \frac{(\tau+\varepsilon)^2}{ab} < 0; \qquad 16\tau - (\tau+\varepsilon)^2\Big(\frac{1}{a}+\frac{1}{b}\Big)^2 > 0,$$

which implies that $\big(\frac{1}{a}+\frac{1}{b}\big)^2 < \frac{2}{ab}$, it is contradiction by Cauchy-Schwarz's inequality. Thus, the equation $t(z) = 0$ has unique positive root which is denoted by $\widehat{z}$.

With $\widehat{z} = \frac{\tau}{\widehat{x}\widehat{y}}$, we have $\widehat{x}\widehat{y} = \frac{\tau}{\widehat{z}}$. Substituting it to the system of equations (30), we have

$$\begin{cases} \widehat{x}^2 + \frac{\tau+\varepsilon}{2a}\widehat{x} = \frac{\tau}{\widehat{z}} + \frac{\tau}{2} \\ \widehat{y}^2 + \frac{\tau+\varepsilon}{2b}\widehat{x} = \frac{\tau}{\widehat{z}} + \frac{\tau}{2}. \end{cases}$$

This implies that

$$\widehat{x} = -\frac{\tau+\varepsilon}{4a} + \frac{1}{2}\sqrt{\frac{(\tau+\varepsilon)^2}{4a^2} + 4\tau\Big(\frac{1}{2}+\frac{1}{\widehat{z}}\Big)}, \quad \widehat{y} = -\frac{\tau+\varepsilon}{4b} + \frac{1}{2}\sqrt{\frac{(\tau+\varepsilon)^2}{4b^2} + 4\tau\Big(\frac{1}{2}+\frac{1}{\widehat{z}}\Big)}.$$

As a consequence, we obtain the conclusion of the lemma. □

Lemma B.4 (Minimizer of function $h_{\varepsilon,\tau}$)**.** *Given a constant $a > 0$, the function*

$$h_{\varepsilon,\tau}(x;a) = x^2 + (\tau+\varepsilon)\Big[\frac{x}{a} - \log\Big(\frac{x}{a}\Big) - 1\Big]$$

attains its minimum at $x^* = -\frac{\tau+\varepsilon}{4a} + \sqrt{\frac{(\tau+\varepsilon)^2}{16a^2} + \frac{\tau+\varepsilon}{2}}$.

Proof of Lemma B.4. Taking the first and second derivatives of function $h_{\tau,\varepsilon}$, we have

$$\frac{\partial h_{\tau,\varepsilon}}{\partial x} = 2x + \frac{\tau+\varepsilon}{a} - \frac{\tau+\varepsilon}{x}, \quad \frac{\partial^2 h_{\tau,\varepsilon}}{\partial x^2} = 2 + \frac{\tau+\varepsilon}{x^2}.$$

Note that the second derivative of $h_{\tau,\varepsilon}$ is positive, therefore, the positive minimizer of this function is exactly the positive solution of equation $\frac{\partial h_{\tau,\varepsilon}}{\partial x} = 0$, which is

$$x^* = -\frac{\tau+\varepsilon}{4a} + \sqrt{\frac{(\tau+\varepsilon)^2}{16a^2} + \frac{\tau+\varepsilon}{2}}.$$

□

C. Auxiliary results

In this appendix, we provide additional lemmas that are used to derive the closed-form expressions of entropic (unbalanced) IGW between (unbalanced) Gaussian distributions.

Lemma C.1 (KL divergence between Gaussian measures)**.** *Let $\alpha = m_\alpha \mathcal{N}_k(\mathbf{0}, \Sigma_\alpha)$ and $\beta = m_\beta \mathcal{N}_k(\mathbf{0}, \Sigma_\beta)$ be scaled Gaussian measures while $\overline{\alpha} = \mathcal{N}_k(\mathbf{0}, \Sigma_\alpha)$ and $\overline{\beta} = \mathcal{N}_k(\mathbf{0}, \Sigma_\beta)$ be their normalized versions. Then, the generalized KL divergence between α and β is*

$$\mathrm{KL}(\alpha \| \beta) = m_\alpha \mathrm{KL}(\overline{\alpha} \| \overline{\beta}) + \mathrm{KL}(m_\alpha \| m_\beta),$$

where the KL divergence between $\overline{\alpha}$ and $\overline{\beta}$ is

$$\frac{1}{2}\left\{ \mathrm{tr}\left(\Sigma_\alpha \Sigma_\beta^{-1}\right) - k + \log\left(\frac{\det(\Sigma_\beta)}{\det(\Sigma_\alpha)}\right)\right\}.$$

Proof of Lemma C.1. Let $f(x, \Sigma_\alpha)$ and $f(x, \Sigma_\beta)$ be the distribution functions of $\mathcal{N}_k(\mathbf{0}, \Sigma_\alpha)$ and $\mathcal{N}_k(\mathbf{0}, \Sigma_\beta)$, respectively.

$$\begin{aligned}
\mathrm{KL}(\alpha \| \beta) &= \int_{\mathbb{R}^k} \log\Big(\frac{m_\alpha f(x, \Sigma_\alpha)}{m_\beta f(x, \Sigma_\beta}\Big) d\alpha(x) - m_\alpha + m_\beta \\
&= \int_{\mathbb{R}^k} \log\Big(\frac{f(x, \Sigma_\alpha)}{f(x, \Sigma_\beta)}\Big) m_\alpha d\overline{\alpha}(x) + \log\Big(\frac{m_\alpha}{m_\beta}\Big) m_\alpha - m_\alpha + m_\beta \\
&= m_\alpha \mathrm{KL}(\overline{\alpha} \| \overline{\beta}) + \mathrm{KL}(m_\alpha \| m_\beta).
\end{aligned}$$

The expression for $\mathrm{KL}(\overline{\alpha} \| \overline{\beta})$ results from Lemma C.3. Hence, the proof is completed. □

Lemma C.2 (Double integrals)**.** *For the sake of simple computations, we derive some relations between the quadratic-KL and the standard KL as follows:*

$$\begin{aligned}
&\textit{(i)}\ \mathrm{KL}^{\otimes}(\alpha \| \beta) = 2m_\alpha \mathrm{KL}(\alpha \| \beta) + (m_\alpha - m_\beta)^2; \\
&\textit{(ii)}\ \mathrm{KL}^{\otimes}(t\alpha \| r\beta) = t^2 \mathrm{KL}^{\otimes}(\alpha \| \beta) + t^2 \log\left(\frac{t^2}{r^2}\right) m_\alpha^2 + (r^2 - t^2) m_\beta^2, \quad t, r > 0,
\end{aligned}$$

for any positive measures α and β.

Proof of Lemma C.2. Let p_α and p_β be the Radon-Nikodym derivatives of α and β with respect to the Lebesgue measure. Then, we have
Part (i)

$$\begin{aligned}
\mathrm{KL}^{\otimes}(\alpha \| \beta) &= \int\int \log\left(\frac{p_\alpha(x) p_\alpha(x')}{p_\beta(x) p_\beta(x')}\right) p_\alpha(x) p_\alpha(x') dx dx' - m_\alpha^2 + m_\beta^2 \\
&= m_\alpha \int \log\left(\frac{p_\alpha(x)}{p_\beta(x)}\right) p_\alpha(x) dx + m_\alpha \int \log\left(\frac{p_\alpha(x')}{p_\beta(x')}\right) p_\alpha(x') dx' - m_\alpha^2 + m_\beta^2 \\
&= 2m_\alpha \int \log\left(\frac{p_\alpha(x)}{p_\beta(x)}\right) p_\alpha(x) dx - m_\alpha^2 + m_\beta^2 \\
&= 2m_\alpha [\mathrm{KL}(\alpha \| \beta) + m_\alpha - m_\beta] - m_\alpha^2 + m_\beta^2 \\
&= 2m_\alpha \mathrm{KL}(\alpha \| \beta) + (m_\alpha - m_\beta)^2.
\end{aligned}$$

Part (ii)

$$\begin{aligned}
\mathrm{KL}^{\otimes}(t\alpha\|r\beta) &= \iint \log\left(\frac{t^2 p_\alpha(x)p_\alpha(x')}{r^2 p_\beta(x)p_\beta(x')}\right) t^2 p_\alpha(x)p_\alpha(x')\mathrm{d}x\mathrm{d}x' + r^2 m_\beta^2 - t^2 m_\alpha^2 \\
&= t^2\left[\iint \log\left(\frac{p_\alpha(x)p_\alpha(x')}{p_\beta(x)p_\beta(x')}\right) p_\alpha(x)p_\alpha(x')\mathrm{d}x\mathrm{d}x' + \log\left(\frac{t^2}{r^2}\right) m_\alpha^2\right] + r^2 m_\beta^2 - t^2 m_\alpha^2 \\
&= t^2\left[\iint \log\left(\frac{p_\alpha(x)p_\alpha(x')}{p_\beta(x)p_\beta(x')}\right) p_\alpha(x)p_\alpha(x')\mathrm{d}x\mathrm{d}x' + m_\beta^2 - m_\alpha^2\right] + t^2\log\left(\frac{t^2}{r^2}\right) m_\alpha^2 + (r^2 - t^2)m_\beta^2 \\
&= t^2\mathrm{KL}^{\otimes}(\alpha\|\beta) + t^2\log\left(\frac{t^2}{r^2}\right) m_\alpha^2 + (r^2 - t^2)m_\beta^2.
\end{aligned}$$

As a consequence, we obtain the conclusion of the lemma. □

Lemma C.3 (KL divergence between Gaussians)**.** *The Kullback-Leibler divergence between Gaussian measures $\alpha = \mathcal{N}(\mu_\alpha, \Sigma_\alpha)$ and $\beta = \mathcal{N}(\mu_\beta, \Sigma_\beta)$ on $\mathbb{R}^k$ is given by*

$$\mathrm{KL}\big(\alpha\|\beta\big) = \frac{1}{2}\left[\mathrm{tr}(\Sigma_\beta^{-1}\Sigma_\alpha) + (\mu_\beta - \mu_\alpha)^\top\Sigma_\beta^{-1}(\mu_\beta - \mu_\alpha) - k + \log\left(\frac{\det(\Sigma_\beta)}{\det(\Sigma_\alpha)}\right)\right].$$

As a result, when $\Sigma_\beta, \Sigma_\alpha$ are fixed, $\mathrm{KL}(\alpha\|\beta)$ achieves its minimum value when $\mu_\alpha = \mu_\beta$.

Proof of Lemma C.3. Let $p_\alpha(x)$ and $p_\beta(x)$ be the probability density functions of α and β, respectively. By the formula for KL divergence, we have

$$\mathrm{KL}(\alpha\|\beta) = \int_{\mathbb{R}^k} \log\left(\frac{p_\alpha(x)}{p_\beta(x)}\right) p_\alpha(x)dx, \tag{31}$$

in which

$$\begin{aligned}
\log(p_\alpha(x)) &= -\frac{1}{2}[k\log(2\pi) + \log(\det(\Sigma_\alpha)) + (x-\mu_\alpha)^\top\Sigma_\alpha^{-1}(x-\mu_\alpha)], \\
\log(p_\beta(x)) &= -\frac{1}{2}[k\log(2\pi) + \log(\det(\Sigma_\beta)) + (x-\mu_\beta)^\top\Sigma_\beta^{-1}(x-\mu_\beta)],
\end{aligned}$$

where π denotes a constant number *pi* only in this lemma. Plugging this result in equation (31), we get

$$\begin{aligned}
\mathrm{KL}(\alpha\|\beta) &= \frac{1}{2}\left[\log\left(\frac{\det(\Sigma_\beta)}{\det(\Sigma_\alpha)}\right) + \int_{\mathbb{R}^k}[(x-\beta)^\top\Sigma_\beta^{-1}(x-\beta) - (x-\alpha)^\top\Sigma_\alpha^{-1}(x-\alpha)]p_\alpha(x)dx\right] \\
&= \frac{1}{2}\left[\log\left(\frac{\det(\Sigma_\beta)}{\det(\Sigma_\alpha)}\right) + (\mu_\beta-\mu_\alpha)^\top\Sigma_\beta^{-1}(\mu_\beta-\mu_\alpha) + \int_{\mathbb{R}^k}[(x-\mu_\alpha)^\top(\Sigma_\beta^{-1} - \Sigma_\alpha^{-1})(x-\mu_\alpha)]p_\alpha(x)dx\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\int_{\mathbb{R}^k}[(x-\mu_\alpha)^\top(\Sigma_\beta^{-1} - \Sigma_\alpha^{-1})(x-\mu_\alpha)]p_\alpha(x)dx &= \int_{\mathbb{R}^k}\mathrm{tr}((x-\mu_\alpha)(x-\mu_\alpha)^\top(\Sigma_\beta^{-1} - \Sigma_\alpha^{-1}))p_\alpha(x)dx \\
&= \mathrm{tr}\left(\int_{\mathbb{R}^k}[(x-\mu_\alpha)(x-\mu_\alpha)^\top(\Sigma_\beta^{-1} - \Sigma_\alpha^{-1})]p_\alpha(x)dx\right) \\
&= \mathrm{tr}\left(\Sigma_\alpha(\Sigma_\beta^{-1} - \Sigma_\alpha^{-1})\right) \\
&= \mathrm{tr}(\Sigma_\alpha\Sigma_\beta^{-1}) - k.
\end{aligned}$$

Putting the above results together, we obtain the conclusion of this lemma. □

Lemma C.4 (Order solutions)**.** *Let $\{a_i\}_{i=1}^m$ and $\{b_j\}_{j=1}^m$ be positive decreasing sequences. Let $\{\widetilde{x}_i\}_{i=1}^m$ and $\{\widetilde{y}_j\}_{j=1}^n$ be the minimizer of*

$$\begin{aligned}
\min_{x_i>0, y_j>0} \mathsf{D}\big(\{x_i\},\{y_j\};\{a_i\},\{b_j\}\big) := \sum_{i=1}^m x_i^2 + \sum_{j=1}^n y_j^2 + (\tau+\varepsilon)\Big\{\sum_{i=1}^m\big[\frac{x_i}{a_i} - \log(x_i)\big] + \sum_{j=1}^n\big[\frac{y_j}{b_j} - \log(y_j)\big]\Big\} \\
- 2\sum_{k=1}^n\big[x_k y_k - \frac{\varepsilon}{2}\big]^+ + \varepsilon\big[\log(x_k y_k) - \log\frac{\varepsilon}{2}\big]^+.
\end{aligned}$$

Then $(\widetilde{x}_i)$ and $(\widetilde{y}_j)$ are also decreasing sequences.

Proof of Lemma C.4. We have two proofs for this lemma. While the first proof will be presented here, the second one can be found in Appendix D.

Let us consider all permutations $(\widehat{x}_i)$ and $(\widehat{y}_j)$ of $(\widetilde{x}_i)$ and $(\widetilde{y}_j)$, respectively. The difference between $\mathsf{D}\big(\{\widehat{x}_i\},\{\widehat{y}_j\};\{a_i\},\{b_j\}\big)$ and $\mathsf{D}\big(\{\widetilde{x}_i\},\{\widetilde{y}_j\},\{a_i\},\{b_j\}\big)$ lies in the term

$$-2\sum_{k=1}^{n}\Big[x_k y_k - \frac{\varepsilon}{2}\Big]^+ + \varepsilon\Big[\log(x_k y_k) - \log\frac{\varepsilon}{2}\Big]^+,$$

which is actually the minimum value of

$$-2\,\mathrm{tr}\left(K_{xy}^\top K_{xy}\right) + 2\varepsilon \mathrm{KL}\big(\overline{\pi}\|\overline{\pi}_x \otimes \overline{\pi}_y\big),$$

where its solution has to match order by magnitude $\{\widetilde{x}_i\}$ to $\{\widehat{x}_i\}$ and $\{\widetilde{y}_j\}$ to $\{\widehat{y}_j\}$ of the largest eigenvalues. The part $\sum_{i=1}^m \frac{x_i}{a_i} + \sum_{j=1}^n \frac{y_j}{b_j}$ is minimized by the rearrangement inequality (cf. Lemma D.2). Therefore, we obtain the conclusion of the lemma. □

Lemma C.5. *Let $\{a_i\}_{i=1}^s$ and $\{b_i\}_{i=1}^s$ be positive sequences. The problem of finding the maximum value of*

$$\Big\{\sum_{i=1}^{s} a_i x_i\Big\}^2 + \sum_{i=1}^{s}\frac{b_i}{2}\log(1-x_i), \tag{32}$$

where $x_i \in [0,1]$, is equivalent to finding the set $S \subset [s]$ in the below problem:

$$\max_{S\subset[s]}\Big\{A_S - \frac{B_S}{2\big[A_S + (A_S^2 - B_S)^{\frac{1}{2}}\big]}\Big\}^2 + \sum_{i\in S} b_i\log(b_i) - B_S\log\big(A_S + (A_S^2 - B_S)^{\frac{1}{2}}\big),$$

where $A_S = \sum_{i\in S} a_i$, $B_S = \sum_{i\in S} b_i$ and $A_S^2 \geq B_S$ and

$$b_i \leq 2a_i\big\{A_S + (A_S^2 - B_S)^{\frac{1}{2}}\big\}; \quad i \in S.$$

Proof of Lemma C.5. Taking the derivative of function $f(x) = (ax+b)^2 + c\log(1-x)$ in $[0,1]$, we have

$$f'(x) = 2a(ax+b) - \frac{c}{1-x}.$$

The equation $f'(x) = 0$ has either two solutions or no solution. It means that either $f(x)$ is monotone or it has one minimum and one maximum. When $x \to 1^-$, $f(x)$ diverges to infinity. Thus, the $x_{\max}$ where function f attains its maximum is closer to 1 than the $x_{\min}$, where f attains its minimum. In the second case, f is monotone, then f is monotone decreasing, because $f(0) = 0$ and $f(1^-) = -\infty$. Overall,

$$x_{\max} = \begin{cases} 0 \\ \text{the larger solution of equation:} \quad 2a(ax+b) = \frac{c}{1-x}. \end{cases}$$

Hence, if $\{\widetilde{x}_i\}$ is a maximizer of the objective function, then either $\widetilde{x}_i = 0$ or $\widetilde{x}_i$ is the larger solution of the first derivative system of equations.

We consider the case that all $\widetilde{x}_i$ are not equal to zero. By taking the derivatives of the function in equation (32) with respect to x_i, we obtain

$$\begin{aligned} 2a_i\Big\{\sum_{j=1}^{s} a_j\widetilde{x}_j\Big\} - \frac{b_i}{2}\frac{1}{1-\widetilde{x}_i} &= 0, \\ 4\big\{a_i - a_i\widetilde{x}_i\big\} \times \Big\{\sum_{j=1}^{s} a_j\widetilde{x}_j\Big\} &= b_i, \\ 4\Big\{\sum_{j=1}^{s} a_j - \sum_{j=1}^{s} a_j,\widetilde{x}_j\Big\} \times \Big\{\sum_{j=1}^{s} a_j\widetilde{x}_j\Big\} &= \sum_{j=1}^{s} b_j. \end{aligned} \tag{33}$$

Denote $A = \sum_{j=1}^{s} a_j$ and $B = \sum_{j=1}^{s} b_j$. Then, by solving the quadratic equation $4X^2 - 4AX + B = 0$, we have

$$\sum_{j=1}^{s} a_j \widetilde{x}_j = \frac{1}{2}\big\{A + \sqrt{A^2 - B}\big\}.$$

Substituting it into equation (33), we obtain

$$\begin{aligned} 4a_i(1 - \widetilde{x}_i) &= \frac{2b_i}{A + \sqrt{A^2 - B}} \\ \widetilde{x}_i &= 1 - \frac{b_i}{2a_i(A + \sqrt{A^2 - B})}. \end{aligned}$$

Then, the minimum value of the objective function in this case equals to

$$\Big\{\sum_{i=1}^{s} a_i - \frac{1}{2(A + \sqrt{A^2 - B})}\sum_{i=1}^{s} b_i\Big\}^2 + \sum_{i=1}^{s} b_i \log(b_i) - \log\big(A + \sqrt{A^2 - B}\big)\sum_{i=1}^{s} b_i.$$

We have thus proved the stated result. □

Lemma C.6 (Optimizers of some functions)**.**

(a) For positive constants a and c, the function $f(x) = ax + c\log(1-x)$ for $x \in [0,1)$ attains its minimum at $x^* = \Big[1 - \frac{c}{a}\Big]^+$.

(b) For positive constants a and c, the function $f(x) = 2ax^{\frac{1}{2}} + c\log(1-x)$ for $x \in [0,1)$ attains its maximizer at $x^* = \Big[\sqrt{\frac{c^2}{4a^2} + 1} - \frac{c}{2a}\Big]^2$.

(c) For positive constants a, b, c and Υ, the function $f(x) = \Upsilon x + \tau \mathrm{KL}(x\|a) + \tau \mathrm{KL}(x\|b) + \varepsilon \mathrm{KL}(x\|c)$ attains its minimum at $x^* = a^{\frac{\tau}{2\tau+\varepsilon}} b^{\frac{\tau}{2\tau+\varepsilon}} c^{\frac{\varepsilon}{2\tau+\varepsilon}} \exp\Big\{\frac{-\Upsilon}{2\tau+\varepsilon}\Big\}$.

Proof of Lemma C.6.

For part (a), we have $f'(x) = a - \frac{c}{1-x}$ which is a decreasing function when $x \in [0,1)$. It means that the function is convex. Hence its maximum attains at the equation $f'(x) = 0$ or at the boundary. In fact, we have $f'(1 - c/a) = 0$. Therefore, the maximizer attains at $\max(0, 1 - c/a)$.

For part (b), we have $f'(x) = ax^{-\frac{1}{2}} - \frac{c}{1-x}$, which is also a decreasing function. Hence, f is concave, it means that the maximum attains at the solution of $f'(x) = 0$ or the boundary. Solving the equation $f'(x) = 0$ we obtain $x^* = \Big[\sqrt{\frac{c^2}{4a^2} + 1} - \frac{c}{2a}\Big]^2$, which belongs to $(0,1)$.

For part (c), by taking the derivative of $f(x)$, we have

$$f'(x) = \Upsilon + \tau\big\{2\log(x) - \log(ab)\big\} + \varepsilon\big\{\log(x) - \log(c)\big\}$$

Solving the equation $f'(x) = 0$, we obtain

$$x^* = \exp\Big\{\frac{\tau\log(a) + \tau\log(b) + \varepsilon\log(c) - \Upsilon}{2\tau + \varepsilon}\Big\} = a^{\frac{\tau}{2\tau+\varepsilon}} b^{\frac{\tau}{2\tau+\varepsilon}} c^{\frac{\varepsilon}{2\tau+\varepsilon}} \exp\Big\{\frac{-\Upsilon}{2\tau+\varepsilon}\Big\}.$$

□

Lemma C.7. *Let (X,Y) and (X',Y') be two i.i.d random vectors in $\mathbb{R}^m \times \mathbb{R}^n$ and follow a probability distribution π which have a mean vector $(u,v) \in \mathbb{R}^{m\times n}$ and a covariance matrix $\begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_y \end{pmatrix}$ where Σ_x and Σ_y are covariance matrices of X and Y, respectively. Denote $Z = X - u, Z' = X' - u$ and $T = Y - v, T' = Y' - v$, then*

$$\mathbb{E}_{\pi\otimes\pi}\Big[\langle X, X'\rangle - \langle Y, Y'\rangle\Big]^2 = \mathbb{E}_{\pi\otimes\pi}\Big[\langle Z, Z'\rangle - \langle T, T'\rangle\Big]^2 + 2u^\top\Sigma_x u - 4u^\top\Sigma_{xy}v + 2v^\top\Sigma_y v + \Big[\|u\|^2 - \|v\|^2\Big]^2.$$

Proof of Lemma C.7. Firstly, we have

$$\begin{aligned}\langle X, X'\rangle - \langle Y, Y'\rangle &= \langle Z+u, Z'+u\rangle - \langle T+v, T'+v\rangle \\ &= \langle Z, Z'\rangle - \langle T, T'\rangle + \langle u, Z+Z'\rangle - \langle v, T+T'\rangle + \|u\|^2 - \|v\|^2.\end{aligned}$$

Note that

$$\mathbb{E}_{\pi\otimes\pi}\langle u, Z\rangle\langle Z, Z'\rangle = \mathbb{E}_{\pi\otimes\pi}\Big[\sum_{i=1}^m Z_i Z_i' \sum_{j=1}^m u_j Z_j\Big] = \sum_{i=1,j=1}^m u_j \mathbb{E}_{\pi\otimes\pi}\big[Z_i Z_i' Z_j\big] = \sum_{i=1,j=1}^m u_j \mathbb{E}_{\pi\otimes\pi}\big[Z_i Z_j\big]\mathbb{E}_{\pi\otimes\pi}[Z_i'] = 0$$

$$\mathbb{E}_{\pi\otimes\pi}\langle Z, Z'\rangle = \mathbb{E}_{\pi\otimes\pi}\Big[\sum_{i=1}^m Z_i Z_i'\Big] = 0$$

$$\mathbb{E}_{\pi\otimes\pi}\langle Z, Z'\rangle\langle v, T\rangle = \mathbb{E}_{\pi\otimes\pi}\Big[\sum_{i=1}^m Z_i Z_i' \sum_{j=1}^n v_j T_j\Big] = \sum_{i=1}^m\sum_{j=1}^n v_j \mathbb{E}_{\pi\otimes\pi}\big[Z_i Z_i' T_j\big] = \sum_{i=1}^m\sum_{j=1}^n \mathbb{E}_{\pi\otimes\pi}[Z_i T_j]\mathbb{E}_{\pi\otimes\pi}[Z_i'] = 0.$$

Similarly, we get

$$\mathbb{E}_{\pi\otimes\pi}\langle v, T\rangle\langle T, T'\rangle = \mathbb{E}_{\pi\otimes\pi}\langle T, T'\rangle = \mathbb{E}_{\pi\otimes\pi}\langle T, T'\rangle\langle u, Z\rangle = 0.$$

As a consequence, we can deduce that

$$\mathbb{E}_{\pi\otimes\pi}\Big[\langle X, X'\rangle - \langle Y, Y'\rangle\Big]^2 = \mathbb{E}_{\pi\otimes\pi}\Big[\langle Z, Z'\rangle - \langle T, T'\rangle\Big]^2 + \mathbb{E}_{\pi\otimes\pi}\Big[\langle u, Z+Z'\rangle - \langle v, T+T'\rangle\Big]^2 + \Big[\|u\|^2 - \|v\|^2\Big]^2.$$

Next, it is sufficient to show that

$$\mathbb{E}_{\pi\otimes\pi}\Big[\langle u, Z+Z'\rangle - \langle v, T+T'\rangle\Big]^2 = 2u^\top\Sigma_x u + 2v^\top\Sigma_y v - 4u^\top\Sigma_{xy}v.$$

Indeed, we have

$$\begin{aligned}\mathbb{E}_{\pi\otimes\pi}\Big[\langle u, Z+Z'\rangle\Big]^2 &= \mathbb{E}_{\pi\otimes\pi}\Big[u^\top(Z+Z')(Z+Z')^\top u\Big] = u^\top\mathbb{V}ar(Z+Z')u = 2u^\top\Sigma_x u \\ \mathbb{E}_{\pi\otimes\pi}\Big[\langle v, T+T'\rangle\Big]^2 &= \mathbb{E}_{\pi\otimes\pi}\Big[v^\top(T+T')(T+T')^\top v\Big] = v^\top\mathbb{V}ar(T+T')v = 2v^\top\Sigma_y v \\ \mathbb{E}_{\pi\otimes\pi}\Big[\langle u, Z+Z'\rangle\langle v, T+T'\rangle\Big] &= u^\top\mathbb{E}_{\pi\otimes\pi}\Big[(Z+Z')(T+T')^\top\Big]v = u^\top\mathbb{E}_{\pi\otimes\pi}\Big[ZT^\top + Z'(T')^\top\Big]v = 2u^\top\Sigma_{xy}v.\end{aligned}$$

Hence, the proof is completed. □

D. Another proof of Lemma C.4

In this appendix, we introduce another proof of Lemma C.4 using Karamata's inequality. Firstly, we introduce an inequality which can be considered to be a generalization of Karamata's inequality in a special case.

Lemma D.1 (A quasi-Karamata inequality)**.** *Let $a_1 \geq a_2 \geq \ldots \geq a_n \geq 0$ and $b_1 \geq b_2 \geq \ldots \geq b_n \geq 0$ which satisfy*

$$\begin{gathered}a_1 \geq b_1 \\ a_1 + a_2 \geq b_1 + b_2 \\ \ldots \\ a_1 + a_2 + \ldots + a_{n-1} \geq b_1 + b_2 + \ldots + b_{n-1} \\ a_1 + a_2 + \ldots + a_n \geq b_1 + b_2 + \ldots + b_n\end{gathered}$$

Let f be a concave and non-increasing function in $[0, +\infty)$. Then, we have

$$f(a_1) + f(a_2) + \ldots + f(a_n) \leq f(b_1) + f(b_2) + \ldots + f(b_n).$$

Proof of Lemma D.1. We prove the Lemma by using induction. The Lemma is obvious when $n = 1$.

Suppose that the problem is true for $n-1$. For $i \in \{1, 2, \ldots, n\}$, let

$$c_i = a_1 + a_2 + \ldots + a_i - b_1 - b_2 - \ldots - b_i.$$

Let k ($1 \leq k \leq n$) be the smallest index such that $c_k = \min_{1 \leq i \leq n} c_i$. Let $b'_1 = b_1 + c_k$, then, we have $b'_1 \geq b_2 \geq \ldots \geq b_n$. We consider two cases:

- If $k = n$, we have

$$\begin{gathered} a_1 \geq b'_1 \\ a_1 + a_2 \geq b'_1 + b_2 \\ \ldots \\ a_1 + a_2 + \ldots + a_n = b'_1 + b_2 + \ldots + b_n \end{gathered}$$

  Applying Karamata's inequality, we have

$$f(a_1) + f(a_2) + \ldots + f(a_n) \leq f(b'_1) + f(b_2) + \ldots + f(b_n) \leq f(b_1) + f(b_2) + \ldots + f(b_n)$$

  .

- If $k < n$. We have

$$\begin{gathered} a_1 \geq b'_1 \\ a_1 + a_2 \geq b'_1 + b_2 \\ \ldots \\ a_1 + a_2 + \ldots + a_k = b'_1 + b_2 + \ldots + b_k. \end{gathered}$$

  Applying Karamata's inequality, we have

$$f(a_1) + f(a_2) + \ldots + f(a_k) \leq f(b'_1) + f(b_2) + \ldots + f(b_n) \leq f(b_1) + f(b_2) + \ldots + f(b_k)$$

  .

  Moreover, when $i > k$, we have

$$a_{k+1} + \ldots + a_i - b_{k+1} - \ldots - b_i = c_i - c_k \geq 0.$$

  Applying the inductive assumption for two arrays $a_{k+1}, \ldots, a_n$ and $b_{k+1}, \ldots, b_n$, we have

$$f(a_{k+1}) + f(a_{k+2}) + \ldots + f(a_n) \leq f(b_{k+1}) + f(b_{k+2}) + \ldots + f(b_n).$$

  which implies

$$f(a_1) + f(a_2) + \ldots + f(a_n) \leq f(b_1) + f(b_2) + \ldots + f(b_n).$$

As a consequence, we obtain the conclusion of the Lemma. □

Lemma D.2 (Extension of rearrangement inequality)**.** *Let $a_1 \geq a_2 \ldots \geq a_n \geq 0$, $b_1 \geq b_2 \geq \ldots \geq b_n \geq 0$ be two decreasing array. Consider α and β are two injections from the set $\{1, 2, \ldots, k\}$ to $\{1, 2 \ldots, n\}$ in the set $\{1, 2, \ldots, n\}$, where k is a positive integer such that $k \leq n$, we have*

$$\max a_{\alpha(1)} b_{\beta(1)} + \ldots + a_{\alpha(k)} b_{\beta(k)} = a_1 b_1 + a_2 b_2 + \ldots + a_k b_k,$$

and the maximum value is achieved when α, β are two permutation which satisfy

$$a_{\alpha(i)} = a_i, \quad b_{\beta(i)} = b_i \quad \forall i : 1 \leq i \leq k.$$

Proof of Lemma D.2. Let $I = \{i_1, i_2, \ldots, i_t\}$ be the indices which are not greater than k such that $\alpha(i_r) > k, \forall r : 1 \leq r \leq t$. Let $J = \{j_1, j_2 \ldots, j_t\}$ be the indices which are not greater than k such that $\alpha(r) \notin J, \forall r : 1 \leq r \leq k$. Consider the permutation $\tilde{\alpha} : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ which is defined by

$$\tilde{\alpha}(j) = \begin{cases} \alpha(j), & \text{if } j \notin I \\ j_r & \text{if } j = i_r \end{cases}.$$

It is obviously that

$$a_{\alpha(1)}b_{\beta(1)} + \ldots + a_{\alpha(k)}b_{\beta(k)} \leq a_{\tilde{\alpha}(1)}b_{\beta(1)} + \ldots + a_{\tilde{\alpha}(k)}b_{\beta(k)}.$$

Similarly, there exist a permutation $\tilde{\beta} : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$ such that

$$a_{\alpha(1)}b_{\beta(1)} + \ldots + a_{\alpha(k)}b_{\beta(k)} \leq a_{\tilde{\alpha}(1)}b_{\beta(1)} + \ldots + a_{\tilde{\alpha}(k)}b_{\beta(k)} \leq a_{\tilde{\alpha}(1)}b_{\tilde{\beta}(1)} + \ldots + a_{\tilde{\alpha}(k)}b_{\tilde{\beta}(k)}.$$

Applying the rearrangement inequality, we have

$$a_{\tilde{\alpha}(1)}b_{\tilde{\beta}(1)} + \ldots + a_{\tilde{\alpha}(k)}b_{\tilde{\beta}(k)} \leq a_1b_1 + a_2b_2 + \ldots + a_kb_k,$$

which implies

$$a_{\alpha(1)}b_{\beta(1)} + \ldots + a_{\alpha(k)}b_{\beta(k)} \leq a_1b_1 + a_2b_2 + \ldots + a_kb_k.$$

From this, we can conclude our proof. □

Given the above lemmas, we have another proof of Lemma C.4.

Another Proof of Lemma C.4. Let us consider a permutation $(\widehat{x}_i)$ and $(\widehat{y}_i)$ of $(\tilde{x}_i)$ and $(\tilde{y}_i)$, respectively, such that $(\widehat{x}_i)$ and $(\widehat{y}_i)$ are decreasing sequence. The difference between $\mathsf{D}\big(\{\widehat{x}_i\}, \{\widehat{y}_j\}; \{a_i\}, \{b_j\}\big)$ and $\mathsf{D}\big(\{\widetilde{x}_i\}, \{\widetilde{y}_j\}, \{a_i\}, \{b_j\}\big)$ lies in the part

$$-2\sum_{k=1}\Big[x_ky_k - \frac{\varepsilon}{2}\Big]^+ + \varepsilon\Big[\log(x_ky_k) - \log\frac{\varepsilon}{2}\Big]^+.$$

Let $f(x) = -2\left(x - \frac{\varepsilon}{2}\right)^+ + \varepsilon\left(\log(x) - \log\left(\frac{\varepsilon}{2}\right)\right)$. This function has non-positive and decreasing derivative

$$f'(x) = \begin{cases} 0, & \text{when } x \leq \frac{\varepsilon}{2} \\ \dfrac{\varepsilon}{x} - 2, & \text{when } x \geq \frac{\varepsilon}{2} \end{cases}.$$

Thus, f is concave function in $[0.\infty)$. Moreover, according to Lemma D.2, we have $(\widehat{x}_i\widehat{y}_i)$ and $(\tilde{x}_i\tilde{y}_i))$ satisfy the condition of Lemma D.1. Thus, applying this Lemma with the function f, we have the term

$$-2\sum_{k=1}\Big[x_ky_k - \frac{\varepsilon}{2}\Big]^+ + \varepsilon\Big[\log(x_ky_k) - \log\frac{\varepsilon}{2}\Big]^+.$$

is minimized when x_k and y_k are decreasing sequence. The part $\sum_{i=1}^m \frac{x_i}{a_i} + \sum_{j=1}^n \frac{y_j}{b_j}$ is minimized by the rearrangement inequality. Therefore, we obtain the conclusion of the lemma. □

E. Entropic IGW between Gaussian distributions with non-zero means

In this appendix, by considering the problem of solving entropic Gromov-Wasserstein between balanced Gaussians in a special case, we show the difficulty to solve this problem in the general case.

Let m be a positive integer. Let $\mu = \mathcal{N}(\theta_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(1, 1)$ be two Gaussian measures in $\mathbb{R}^m$ and $\mathbb{R}$, respectively, where

$$\Sigma_\mu = \mathrm{diag}\big(a_1^2, \ldots, a_m^2\big),$$
$$\theta_\mu = (b_1, \ldots, b_m)^\top.$$

To find $\mathsf{IGW}_\varepsilon(\mu,\nu)$, we must find the joint distribution π between μ and ν which has the covariance matrix

$$\Sigma_\pi = \begin{pmatrix} \Sigma_\mu & K_{\mu\nu} \\ K_{\mu\nu}^\top & 1 \end{pmatrix}$$

that minimizes the following term

$$\mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} + \varepsilon \mathrm{KL}\big(\pi\|\mu\otimes\nu\big).$$

Note that, in this case, $K_{\mu\nu}$ is an $m\times 1$ matrix, thus, we can denote that $K_{\mu\nu} = (x_1,\ldots,x_m)^\top$. Firstly, we prove the following equality

$$\mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} = \sum_{i=1}^m a_i^2 + \sum_{i=1}^m b_i^2 + 2 - 2\sum_{i=1}^m (x_i+b_i)^2. \tag{34}$$

Indeed, we have

$$\mathbb{E}_{\pi\otimes\pi}\Big\{\big[\langle X,X'\rangle - \langle Y,Y'\rangle\big]^2\Big\} = \mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle^2 + \mathbb{E}_{\pi\otimes\pi}\langle Y,Y'\rangle^2 - 2\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle\langle Y,Y'\rangle.$$

Note that the independence between X and X' and Y and Y' leads to

$$\begin{aligned}
\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle^2 &= [\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle]^2 + \mathrm{Var}_\pi\langle X,X'\rangle \\
&= [\langle\theta_\mu,\theta_\mu\rangle]^2 + \sum_{i=1}^m \mathrm{Var}_\pi[X_i]\mathrm{Var}_\pi[X_i'] \\
&= \left(\sum_{i=1}^m b_i^2\right)^2 + \sum_{i=1}^m a_i^4.
\end{aligned}$$

Similarly, we have $\mathbb{E}_{\pi\otimes\pi}\langle Y,Y'\rangle^2 = 2$. Meanwhile,

$$\mathbb{E}_{\pi\otimes\pi}\langle X,X'\rangle\langle Y,Y'\rangle = \mathbb{E}_{\pi\otimes\pi}\Big\{\sum_{i=1}^m \big[X_iX_i'YY'\big]\Big\} = \sum_{i=1}^m \mathbb{E}_{\pi\otimes\pi}\{X_iY\}\mathbb{E}_\pi\{X_i'Y'\} = \sum_{i=1}^m (x_i+b_i)^2.$$

Putting the above results together, we obtain the desired equality (34). By Lemma 3.3, the optimal transportation plan π^* is a Gaussian distribution, namely, $\pi^* = \mathcal{N}(\theta_{\pi^*},\Sigma_{\pi^*})$, where $\theta_{\pi^*} = (b_1,\ldots,b_n,1)^\top$. Thus, according to Lemma C.3, the entropic term in the objective function reads

$$\begin{aligned}
\varepsilon\mathrm{KL}(\pi^*\|\mu\otimes\nu) &= \frac{1}{2}\varepsilon\Big\{\mathrm{tr}\left(\Sigma_{\pi^*}\Sigma_{\mu\otimes\nu}^{-1}\right) - (m+n) + \log\Big(\frac{\det(\Sigma_{\mu\otimes\nu})}{\det(\Sigma_{\pi^*})}\Big)\Big\} \\
&= -\frac{\varepsilon}{2}\log\left(1-\sum_{i=1}^m \frac{x_i^2}{a_i^2}\right).
\end{aligned}$$

From the above calculation, our problem turns out to find the non-negative real numbers $x_1,\ldots,x_m$ which maximize the following function

$$\psi(x_1,\ldots,x_m) = 2\sum_{i=1}^m (x_i+b_i)^2 + \frac{\varepsilon}{2}\log\left(1-\sum_{i=1}^m \frac{x_i^2}{a_i^2}\right).$$

Taking the derivatives of the above function with respect to x_i, $i\in[m]$, we get

$$\frac{\partial\psi}{\partial x_i} = 4(x_i+b_i) - \frac{\varepsilon x_i/a_i^2}{1-\sum_{i=1}^m \frac{x_i^2}{a_i^2}}, \quad \forall i\in[m].$$

To solve the system of equations $\frac{\partial \psi}{\partial x_i} = 0$ for all $i \in [m]$, we denote $t = \frac{\varepsilon}{1 - \sum_{i=1}^m \frac{x_i^2}{a_i^2}}$. Then, $t = \frac{4(x_i + b_i)}{a_i^2 x_i}$, which implies that $x_i = \frac{4a_i^2 b_i}{t - 4a_i^2}$. By substituting this representation of x_i to the expression $t = \frac{\varepsilon}{1 - \sum_{i=1}^m \frac{x_i^2}{a_i^2}}$, we obtain

$$t\left(1 - \sum_{i=1}^m \frac{16 a_i^2 b_i^2}{(t - 4a_i^2)^2}\right) = \varepsilon.$$

After some transformations, we achieve the equation of degree $2m$, which is difficult to solve in close-form.

F. Revisiting the entropic optimal transport between Gaussian measures

In this appendix, via the technique that we use to study the entropic Gromov Wasserstein between two Gaussians, we provide a simpler proof for deriving the close-form expressions of entropic optimal transport (OT) between (unbalanced) Gaussian measures than the proof that was presented in (Janati et al., 2020).

F.1. Entropic optimal transport between balanced Gaussian measures

Before moving to the main theorem of this section, we review some backgrounds on the entropic optimal transport problem between balanced Gaussian measures.

Entropic Optimal Transport. Let α and β be two probability measures in $\mathbb{R}^d$. Then, the entropic OT between them is defined as follows:

$$\mathsf{OT}_\varepsilon(\alpha, \beta) := \min_{\pi \in \Pi(\alpha, \beta)} \mathbb{E}_\pi \|X - Y\|^2 + \varepsilon \mathrm{KL}(\pi \| \alpha \otimes \beta), \tag{35}$$

where $X \sim \alpha$ and $Y \sim \beta$ are independent random vectors in $\mathbb{R}^d$ such that $(X, Y) \sim \pi$, and $\varepsilon > 0$ denotes the entropy-regularized parameter. Now, we are ready to show the closed-form expression of entropic OT between Gaussians in the following theorem.

Theorem F.1. *Let $\mu = \mathcal{N}(\mathbf{u}, \Sigma_\mu)$ and $\nu = \mathcal{N}(\mathbf{v}, \Sigma_\nu)$ be two Gaussian measures in $\mathbb{R}^d$, where Σ_μ and Σ_ν are positive definite matrices. Then, the entropic optimal transport between μ and ν is given by*

$$\mathsf{OT}_\varepsilon(\mu, \nu) = \|\mathbf{u} - \mathbf{v}\|_2^2 + \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) - \mathrm{tr}(D_\varepsilon) + \frac{d\varepsilon}{2}(1 - \log \varepsilon) + \log\det\left(D_\varepsilon + \frac{\varepsilon}{2}\,\mathrm{Id}\right),$$

where $D_\varepsilon = \left(4\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\nu^{\frac{1}{2}} + \frac{\varepsilon^2}{4}\,\mathrm{Id}\right)^{\frac{1}{2}}$. Moreover, the optimal transportation plan admits the form $\pi^* = \mathcal{N}\big((\mathbf{u}^*, \mathbf{v}^*)^\top, \Sigma_{\pi^*}\big)$ where

$$\Sigma_{\pi^*} := \begin{pmatrix} \Sigma_\mu & K_{\mu\nu}^* \\ (K_{\mu\nu}^*)^\top & \Sigma_\nu \end{pmatrix},$$

with $K_{\mu\nu}^* = \frac{1}{2}\Sigma_\mu^{\frac{1}{2}} \left(\frac{\varepsilon^2}{4}\ \mathrm{Id} + 4\,\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma_\mu^{-\frac{1}{2}} - \frac{\varepsilon}{4}\,\mathrm{Id}$.

Remark F.2. Note that if we replace ε by $2\sigma^2$ in the above result, we obtain Theorem 1 in (Janati et al., 2020).

Proof of Theorem F.1. Note that if $\bar{\mu}$ and $\bar{\nu}$ are respective centered transformations of μ and ν, then

$$\mathsf{OT}_\varepsilon(\mu, \nu) = \mathsf{OT}_\varepsilon(\bar{\mu}, \bar{\nu}) + \|\mathbf{u} - \mathbf{v}\|_2^2.$$

Therefore, it is sufficient to solve the case when $\mathbf{u} = \mathbf{v} = \mathbf{0}_d$. Under this assumption, we have

$$\begin{aligned} \mathbb{E}_\pi \|X - Y\|^2 &= \mathbb{E}_\pi \|X\|^2 + \mathbb{E}_\pi \|Y\|^2 - 2\mathbb{E}_\pi \langle X, Y \rangle \\ &= \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) - 2\,\mathrm{tr}(K_{\mu\nu}). \end{aligned}$$

Thus, the entropic optimal transport between μ and ν can be rewritten as

$$\mathsf{OT}_\varepsilon(\mu,\nu) = \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) + \min_{\pi\in\Pi(\mu,\nu)} \Big\{\varepsilon \mathrm{KL}(\pi\|\mu\otimes\nu) - 2\,\mathrm{tr}(K_{\mu\nu})\Big\}. \tag{36}$$

By fixing the covariance matrix $\Sigma_\pi = \begin{pmatrix} \Sigma_\mu & K_{\mu\nu} \\ K_{\mu\nu}^\top & \Sigma_\nu \end{pmatrix}$, Lemma 3.3 indicates that the optimal solution must be a Gaussian measure with the form $\pi = \mathcal{N}(\mathbf{0}_{2d}, \Sigma_\pi)$. Let $U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top$ be the SVD of matrix $\Sigma_\mu^{-\frac{1}{2}}K_{\mu\nu}\Sigma_\nu^{-\frac{1}{2}}$ with $\Lambda_{\mu\nu}^{\frac{1}{2}} = \mathrm{diag}(\kappa_{\mu\nu,i}^{\frac{1}{2}})_{i=1}^d$. It follows from Lemma C.1 and Lemma 3.4 part (b) that

$$\varepsilon\mathrm{KL}(\pi\|\mu\otimes\nu) = \frac{\varepsilon}{2}\left[\mathrm{tr}(\Sigma_\pi\Sigma_{\mu\otimes\nu}^{-1}) - 2d + \log\left(\frac{\det(\Sigma_{\mu\otimes\nu})}{\det(\Sigma_\pi)}\right)\right] = -\frac{\varepsilon}{2}\sum_{i=1}^d \log(1-\kappa_{\mu\nu,i}).$$

Let $U\Lambda^{\frac{1}{2}}V^\top$ be the SVD of $\Sigma_\nu^{\frac{1}{2}}\Sigma_\mu^{\frac{1}{2}}$, where $\Lambda^{\frac{1}{2}} = \mathrm{diag}(\lambda_{\mu\nu,i}^{\frac{1}{2}})_{i=1}^d$. By using von Neumann's inequality (Kristof, 1969; Horn & Johnson, 1991), we have

$$\begin{aligned}\mathrm{tr}(K_{\mu\nu}) &= \mathrm{tr}(\Sigma_\mu^{\frac{1}{2}}U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top\Sigma_\nu^{\frac{1}{2}}) = \mathrm{tr}(U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}V_{\mu\nu}^\top\Sigma_\nu^{\frac{1}{2}}\Sigma_\mu^{\frac{1}{2}}) = \mathrm{tr}([V^\top U_{\mu\nu}\Lambda_{\mu\nu}^{\frac{1}{2}}][V_{\mu\nu}^\top U\Lambda^{\frac{1}{2}}]) \\ &\le \sum_{i=1}^d \lambda_{\mu\nu,i}^{\frac{1}{2}}\kappa_{\mu\nu,i}^{\frac{1}{2}}.\end{aligned}$$

The equality occurs when $U_{\mu\nu} = V$ and $V_{\mu\nu} = U$. Collecting the above two results, the problem in equation (36) is reduced to

$$\mathsf{OT}_\varepsilon(\mu,\nu) = \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) - \max_{\kappa_{\mu\nu,i}\in(0,1)} \sum_{i=1}^d \left[2\lambda_{\mu\nu,i}^{\frac{1}{2}}\kappa_{\mu\nu,i}^{\frac{1}{2}} + \frac{\varepsilon}{2}\log(1-\kappa_{\mu\nu})\right].$$

Note that the function $f(x) = 2ax^{\frac{1}{2}} + \frac{\varepsilon}{2}\log(1-x)$ attains its maximum at $x = \left(\frac{-\varepsilon+\sqrt{\varepsilon^2+16a^2}}{4a}\right)^2$. Then, the optimal solution of the above problem is $\kappa_{\mu\nu,i}^* = \frac{\left(-\varepsilon+\sqrt{\varepsilon^2+16\lambda_{\mu\nu,i}}\right)^2}{16\lambda_{\mu\nu,i}}$, which leads to

$$\begin{aligned}\mathsf{OT}_\varepsilon(\mu,\nu) &= \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) - \frac{1}{2}\sum_{i=1}^d \left[-\varepsilon + \sqrt{\varepsilon^2+16\lambda_{\mu\nu,i}} + \varepsilon\log\left(\frac{\varepsilon(-\varepsilon+\sqrt{\varepsilon^2+16\lambda_{\mu\nu,i}})}{8\lambda_{\mu\nu,i}}\right)\right] \\ &= \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) + \frac{d\varepsilon}{2}(1-\log\varepsilon) - \frac{1}{2}\sum_{i=1}^d \sqrt{\varepsilon^2+16\lambda_{\mu\nu,i}} - \frac{\varepsilon}{2}\sum_{i=1}^d \log\left(\frac{-\varepsilon+\sqrt{\varepsilon^2+16\lambda_{\mu\nu,i}}}{8\lambda_{\mu\nu,i}}\right) \\ &= \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) + \frac{d\varepsilon}{2}(1-\log\varepsilon) - \sum_{i=1}^d \sqrt{\frac{\varepsilon^2}{4}+4\lambda_{\mu\nu,i}} + \frac{\varepsilon}{2}\log\left(\prod_{i=1}^d\left[\frac{\varepsilon}{2}+\sqrt{\frac{\varepsilon^2}{4}+4\lambda_{\mu\nu,i}}\right]\right) \\ &= \mathrm{tr}(\Sigma_\mu) + \mathrm{tr}(\Sigma_\nu) + \frac{d\varepsilon}{2}(1-\log\varepsilon) - \mathrm{tr}\left(\left[4\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}} + \frac{\varepsilon^2}{4}\,\mathrm{Id}\right]^{\frac{1}{2}}\right) \\ &\quad + \log\det\left(\left[4\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}} + \frac{\varepsilon^2}{4}\,\mathrm{Id}\right]^{\frac{1}{2}} + \frac{\varepsilon}{2}\,\mathrm{Id}\right).\end{aligned}$$

The last equality results from the fact that $(\lambda_{\mu\nu,i})_{i=1}^d$ are eigenvalues of $\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}$. Additionally,

$$\begin{aligned}K_{\mu\nu}^* &= \Sigma_\mu^{\frac{1}{2}}V\mathrm{diag}\left(\frac{-\varepsilon+\sqrt{\varepsilon^2+16\lambda_{\mu\nu,i}}}{4\lambda_{\mu\nu,i}^{\frac{1}{2}}}\right)_{i=1}^d U^\top\Sigma_\nu^{\frac{1}{2}} \\ &= \frac{1}{2}\,\Sigma_\mu^{\frac{1}{2}}V\mathrm{diag}\left(\left[\frac{\varepsilon^2}{4\lambda_{\mu\nu,i}}+4\right]^{\frac{1}{2}}\right)_{i=1}^d U^\top\Sigma_\nu^{\frac{1}{2}} - \frac{\varepsilon}{4}\,\Sigma_\mu^{\frac{1}{2}}V\mathrm{diag}\left(\lambda_{\mu\nu,i}^{-\frac{1}{2}}\right)_{i=1}^d U^\top\Sigma_\nu^{\frac{1}{2}}.\end{aligned}$$

Recall that $\Sigma_\nu^{\frac{1}{2}}\Sigma_\mu^{\frac{1}{2}} = U\Lambda^{\frac{1}{2}}V^\top$, or equivalently $U^\top\Sigma_\nu^{\frac{1}{2}} = \Lambda^{\frac{1}{2}}V^\top\Sigma_\mu^{-\frac{1}{2}}$. Then, we have

$$\begin{aligned}K^*_{\mu\nu} &= \frac{1}{2}\,\Sigma_\mu^{\frac{1}{2}}V\mathrm{diag}\left(\left[\frac{\varepsilon^2}{4\lambda_{\mu\nu,i}}+4\right]^{\frac{1}{2}}\right)_{i=1}^d \mathrm{diag}\left(\lambda_{\mu\nu,i}^{\frac{1}{2}}\right)_{i=1}^d V^\top\Sigma_\mu^{-\frac{1}{2}} - \frac{\varepsilon}{4}\,\Sigma_\mu^{\frac{1}{2}}(U\Lambda^{\frac{1}{2}}V^\top)^{-1}\Sigma_\nu^{\frac{1}{2}}\\ &= \frac{1}{2}\,\Sigma_\mu^{\frac{1}{2}}V\mathrm{diag}\left(\left[\frac{\varepsilon^2}{4}+4\lambda_{\mu\nu,i}\right]^{\frac{1}{2}}\right)_{i=1}^d V^\top\Sigma_\mu^{-\frac{1}{2}} - \frac{\varepsilon}{4}\,\mathrm{Id}\,.\end{aligned}$$

As a result of SVD property, we get $V\mathrm{diag}(\lambda_{\mu\nu,i})_{i=1}^d V^\top = \Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}$. Thus, $V\mathrm{diag}\left(\frac{\varepsilon^2}{4}+4\lambda_{\mu\nu,i}\right)_{i=1}^d V^\top = \frac{\varepsilon^2}{4}\,\mathrm{Id}+4\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}$, which follows that

$$V\mathrm{diag}\left(\left[\frac{\varepsilon^2}{4}+4\lambda_{\mu\nu,i}\right]^{\frac{1}{2}}\right)_{i=1}^d V^\top = \left(\frac{\varepsilon^2}{4}\,\mathrm{Id}+4\,\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}}.$$

Consequently, we obtain

$$K^*_{\mu\nu} = \frac{1}{2}\Sigma_\mu^{\frac{1}{2}}\left(\frac{\varepsilon^2}{4}\,\mathrm{Id}+4\,\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_\mu^{-\frac{1}{2}} - \frac{\varepsilon}{4}\mathrm{Id} = \Sigma_\mu^{\frac{1}{2}}\left(\frac{\varepsilon^2}{16}\,\mathrm{Id}+\Sigma_\mu^{\frac{1}{2}}\Sigma_\nu\Sigma_\mu^{\frac{1}{2}}\right)^{\frac{1}{2}}\Sigma_\mu^{-\frac{1}{2}} - \frac{\varepsilon}{4}\,\mathrm{Id}\,.$$

Hence, the proof is completed. □

F.2. Entropic optimal transport between unbalanced Gaussian measures

Prior to presenting the closed-form expression of entropic optimal transport between unbalanced Gaussian measures, let us review the formulation of entropic unbalanced optimal transport (UOT).

Entropic Unbalanced Optimal Transport. When either α or β is not a probability measure, the formulation in equation (35) is no longer valid. One solution to deal with this issue is using entropic UOT, which is given by

$$\mathsf{UOT}_{\varepsilon,\tau}(\alpha,\beta) := \min_{\pi\in\mathcal{M}^+(\mathbb{R}^d\times\mathbb{R}^d)} \mathbb{E}_\pi\|X-Y\|^2 + \tau\mathrm{KL}(\pi_x\|\alpha) + \tau\mathrm{KL}(\pi_y\|\beta) + \varepsilon\mathrm{KL}(\pi\|\alpha\otimes\beta),$$

where X, Y are independent random vectors in $\mathbb{R}^d$ such that $(X,Y)\sim\pi$ while $\varepsilon,\tau>0$ are regularized parameters, and π_x,π_y are the marginal distributions of the coupling π corresponding to α and β, respectively.

Consider two unbalanced Gaussian measures in $\mathbb{R}^d$:

$$\mu = m_\mu\mathcal{N}(\mathbf{u},\Sigma_\mu) \quad\text{and}\quad \nu = m_\nu\mathcal{N}(\mathbf{v},\Sigma_\nu), \tag{37}$$

where $m_\mu, m_\nu>0$ are their masses and Σ_μ,Σ_ν are positive definite matrices. To state our main result, we need to define some quantities which are necessary for our analysis. Firstly, let us denote

$$\Sigma_{\mu,\eta} := \eta\Sigma_\mu^{-1}+\mathrm{Id}; \qquad \Sigma_{\nu,\eta} := \eta\Sigma_\nu^{-1}+\mathrm{Id},$$

where $\eta := \frac{\tau+\varepsilon}{2}$. Next, we define

$$\begin{aligned}A &:= \frac{1}{\tau}\left\{\left[\frac{\varepsilon^2}{4}\mathrm{Id}+\tau(2\eta)B^\top B\right]^{\frac{1}{2}} - \frac{\varepsilon}{2}\mathrm{Id}\right\},\ \text{where } B := \Sigma_{\nu,\eta}^{-\frac{1}{2}}\Sigma_{\mu,\eta}^{-\frac{1}{2}},\\ \Upsilon_{\mathbf{u}^*,\mathbf{v}^*} &:= (\mathbf{u}-\mathbf{v})^\top\left\{\mathrm{Id}-(\Sigma_\mu+\Sigma_\nu)\left[\eta\,\mathrm{Id}+\Sigma_\mu^{-1}+\Sigma_\nu^{-1}\right]^{-1}\right\}(\mathbf{u}-\mathbf{v}),\\ \Upsilon_{\Sigma^*} &:= 2\eta\Big\{1-\log(\eta)+\log\det(A)-\log\det(B)-\frac{1}{2}\log\det(\Sigma_\mu\Sigma_\nu)\Big\} - \frac{\varepsilon}{2}\log\det\big(\mathrm{Id}-A^2(B^\top B)^{-1}\big),\\ \Upsilon^* &:= \Upsilon_{\mathbf{u}^*,\mathbf{v}^*}+\Upsilon_{\Sigma^*}+2\eta d.\end{aligned}$$

Theorem F.3. *Let μ and ν be two unbalanced Gaussian measures given in equation (37), the entropic UOT between μ and ν can be computed as*

$$\mathsf{UOT}_{\varepsilon,\tau}(\mu,\nu) = m_{\pi^*}\Upsilon^* + \varepsilon \mathrm{KL}(m_{\pi^*}\|m_\mu m_\nu) + \tau[\mathrm{KL}(m_{\pi^*}\|m_\mu) + \mathrm{KL}(m_{\pi^*}\|m_\nu)].$$

where $m_{\pi^*} = (m_\mu m_\nu)^{\frac{\tau+\varepsilon}{2\tau+\varepsilon}} \exp\left\{\frac{-\Upsilon^*}{2\tau+\varepsilon}\right\}$. *Moreover, the optimal transport plan is also an unbalanced Gaussian measure which admits the form* $\pi^* = m_{\pi^*}\mathcal{N}\big((\mathbf{u}^*,\mathbf{v}^*)^\top, \Sigma^*_{xy}\big)$, *where*

$$\begin{aligned}
\mathbf{u}^* &= \mathbf{u} - \Sigma_\mu\Big[\eta\,\mathrm{Id} + \Sigma_\mu + \Sigma_\nu\Big]^{-1}(\mathbf{u}-\mathbf{v});\\
\mathbf{v}^* &= \mathbf{v} + \Sigma_\nu\Big[\eta\,\mathrm{Id} + \Sigma_\mu^{-1} + \Sigma_\nu^{-1}\Big]^{-1}(\mathbf{u}-\mathbf{v});\\
\Sigma^*_{xy} &= \begin{pmatrix} \Sigma^*_x & K^*_{xy} \\ (K^*_{xy})^\top & \Sigma^*_y \end{pmatrix},
\end{aligned}$$

with

$$\begin{aligned}
&\Sigma^*_x = \eta\Sigma_{\mu,\eta}^{-\frac{1}{2}} A (B^\top B)^{-\frac{1}{2}} \Sigma_{\mu,\eta}^{-\frac{1}{2}}; \qquad \Sigma^*_y = \eta\Sigma_{\nu,\eta}^{-\frac{1}{2}} A (B^\top B)^{-\frac{1}{2}} \Sigma_{\nu,\eta}^{-\frac{1}{2}};\\
&K^*_{xy} = \eta\Sigma_{\mu,\eta}^{-\frac{1}{2}} A(\mathrm{Id}-A)^{-1}(B^\top B)^{-\frac{1}{2}} B^{-1}\Sigma_{\nu,\eta}^{-\frac{1}{2}}.
\end{aligned}$$

Proof of Theorem F.3. We assume that π is a positive measure such that $\pi = m_\pi\overline{\pi}$ with $\overline{\pi}$ is a probability measure with mean $(\mathbf{u}_x,\mathbf{v}_y)$ and covariance matrix

$$\Sigma_\pi = \begin{pmatrix} \Sigma_x & K_{xy} \\ K_{xy}^\top & \Sigma_y \end{pmatrix}.$$

Here the two marginals of $\overline{\pi}$ are denoted by $\overline{\pi}_x$ and $\overline{\pi}_y$ where $\overline{\pi}_x$ has mean $\mathbf{u}_x$ and covariance matrix Σ_x while $\overline{\pi}_y$ has mean $\mathbf{v}_y$ and covariance matrix Σ_y. Let π_x and π_y be two marginals of π, then $\pi_x = m_\pi\overline{\pi}_x$ and $\pi_y = m_\pi\overline{\pi}_y$.

By Lemma 3.3, π needs to be Gaussian. We also have

$$\mathbb{E}_\pi\|X-Y\|^2 = m_\pi\Big\{\|\mathbf{u}_x-\mathbf{v}_y\|^2 + \mathrm{tr}(\Sigma_x) + \mathrm{tr}(\Sigma_y) - 2\,\mathrm{tr}(K_{xy})\Big\}.$$

According to Lemma C.1,

$$\begin{aligned}
\mathrm{KL}(\pi_x\|\mu) &= m_\pi\mathrm{KL}(\overline{\pi}_x\|\overline{\mu}) + \mathrm{KL}(m_\pi\|m_\mu),\\
\mathrm{KL}(\pi_y\|\nu) &= m_\pi\mathrm{KL}(\overline{\pi}_y\|\overline{\nu}) + \mathrm{KL}(m_\pi\|m_\nu),\\
\mathrm{KL}(\pi\|\mu\otimes\nu) &= m_\pi\mathrm{KL}(\overline{\pi}\|\overline{\mu}\otimes\overline{\nu}) + \mathrm{KL}(m_\pi\|m_\mu m_\nu).
\end{aligned}$$

Combining the above results, the entropic optimal transport between μ and ν reads as

$$\mathsf{UOT}_{\varepsilon,\tau}(\mu,\nu) = \min_{\pi\in\mathcal{M}^+(\mathbb{R}^d\times\mathbb{R}^d)}\Big\{m_\pi\Upsilon + \varepsilon\mathrm{KL}(m_\pi\|m_\mu m_\nu) + \tau[\mathrm{KL}(m_\pi\|m_\mu) + \mathrm{KL}(m_\pi\|m_\nu)]\Big\},$$

where $\Upsilon := \|\mathbf{u}_x-\mathbf{v}_y\|^2 + \mathrm{tr}(\Sigma_x) + \mathrm{tr}(\Sigma_y) - 2\,\mathrm{tr}(K_{xy}) + \varepsilon\mathrm{KL}(\overline{\pi}\|\overline{\mu}\otimes\overline{\nu}) + \tau\big[\mathrm{KL}(\overline{\pi}_x\|\overline{\mu}) + \mathrm{KL}(\overline{\pi}_y\|\overline{\nu})\big]$. We divide our proof into two main steps. The first step is to minimize Υ, and the second step is to find the optimal scale m_{π^*}.

Step 1: Minimization of Υ For the three KL terms, we have

$$\begin{aligned}
\mathsf{KL}(\overline{\pi}_x\|\overline{\mu}) &= \frac{1}{2}\Big\{\mathrm{tr}(\Sigma_x\Sigma_\mu^{-1}) - d - \log\det(\Sigma_x) + \log\det(\Sigma_\mu)\Big\} + \frac{1}{2}(\mathbf{u}_x-\mathbf{u})^\top\Sigma_\mu^{-1}(\mathbf{u}_x-\mathbf{u}),\\
\mathsf{KL}(\overline{\pi}_y\|\overline{\nu}) &= \frac{1}{2}\Big\{\mathrm{tr}(\Sigma_y\Sigma_\nu^{-1}) - d - \log\det(\Sigma_y) + \log\det(\Sigma_\nu)\Big\} + \frac{1}{2}(\mathbf{v}_y-\mathbf{v})^\top\Sigma_\nu^{-1}(\mathbf{v}_y-\mathbf{v}),\\
\mathsf{KL}(\overline{\pi}\|\overline{\mu}\otimes\overline{\nu}) &= \frac{1}{2}\Big\{\mathrm{tr}(\Sigma_x\Sigma_\mu^{-1}) - d - \log\det(\Sigma_x) + \log\det(\Sigma_\mu)\Big\} + \frac{1}{2}\Big\{\mathrm{tr}(\Sigma_y\Sigma_\nu^{-1}) - d - \log\det(\Sigma_y) + \log\det(\Sigma_\nu)\Big\}\\
&\quad + \frac{1}{2}(\mathbf{u}_x-\mathbf{u},\mathbf{v}_y-\mathbf{v})^\top\Sigma_{\mu\otimes\nu}^{-1}(\mathbf{u}_x-\mathbf{u},\mathbf{v}_y-\mathbf{v}) - \frac{1}{2}\sum_{i=1}^d\log(1-\kappa_{xy,i}),
\end{aligned}$$

where

$$\Sigma_{\mu\otimes\nu} = \begin{pmatrix} \Sigma_\mu & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \Sigma_\nu \end{pmatrix},$$

and $\kappa_{xy,i}^{\frac{1}{2}}$ is the i-th largest singular value of $\Sigma_x^{-\frac{1}{2}} K_{xy} \Sigma_y^{-\frac{1}{2}}$ for all $i \in [d]$. Here we use the equation $\operatorname{tr}(\Sigma_\pi \Sigma_{\mu\otimes\nu}^{-1}) = \operatorname{tr}(\Sigma_x \Sigma_\mu^{-1}) + \operatorname{tr}(\Sigma_y \Sigma_\nu^{-1})$ and results from part (b) of Lemma 3.4. Put the results together, we get

$$\begin{aligned}\Upsilon &= \operatorname{tr}(\Sigma_x) + \operatorname{tr}(\Sigma_y) - 2\operatorname{tr}(K_{xy}) + \eta\Big\{\operatorname{tr}(\Sigma_x\Sigma_\mu^{-1}) - \log\det(\Sigma_x) + \operatorname{tr}(\Sigma_y\Sigma_\nu^{-1}) - \log\det(\Sigma_y)\Big\} \\ &- \frac{\varepsilon}{2}\sum_{i=1}^{d}\log(1-\kappa_{xy,i}) + \|\mathbf{u}_x - \mathbf{v}_y\|^2 + \frac{1}{2}\Big\{\tau(\mathbf{u}_x-\mathbf{u})^\top\Sigma_\mu^{-1}(\mathbf{u}_x-\mathbf{u}) + \tau(\mathbf{v}_y-\mathbf{v})^\top\Sigma_\nu^{-1}(\mathbf{v}_y-\mathbf{v}) \\ &+ \varepsilon(\mathbf{u}_x-\mathbf{u},\mathbf{v}_y-\mathbf{v})^\top\Sigma_{\mu\otimes\nu}^{-1}(\mathbf{u}_x-\mathbf{u},\mathbf{v}_y-\mathbf{v})\Big\} + 2\eta d.\end{aligned}$$

The means part: We first work with the terms involving $\mathbf{u}_x$ and $\mathbf{v}_y$. Let $\mathbf{u}_x - \mathbf{u} = \widetilde{\mathbf{u}}_x$ and $\mathbf{v}_y - \mathbf{v} = \widetilde{\mathbf{v}}_y$, then

$$\begin{aligned}(\widetilde{\mathbf{u}}_x,\widetilde{\mathbf{v}}_y)^\top\Sigma_{\mu\otimes\nu}^{-1}(\widetilde{\mathbf{u}}_x,\widetilde{\mathbf{v}}_y) &= \widetilde{\mathbf{u}}_x^\top\Sigma_\mu^{-1}\widetilde{\mathbf{u}}_x + \widetilde{\mathbf{v}}_y\Sigma_\nu^{-1}\widetilde{\mathbf{v}}_y \\ \|\mathbf{u}_x-\mathbf{v}_y\|^2 &= \|\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{v}}_y + \mathbf{u} - \mathbf{v}\|^2 \\ &= \|\widetilde{\mathbf{u}}_x\|^2 + \|\widetilde{\mathbf{v}}_y\|^2 - 2\widetilde{\mathbf{u}}_x^\top\widetilde{\mathbf{v}}_y + 2\widetilde{\mathbf{u}}_x^\top(\mathbf{u}-\mathbf{v}) - 2\widetilde{\mathbf{v}}_y^\top(\mathbf{u}-\mathbf{v}) + \|\mathbf{u}-\mathbf{v}\|^2.\end{aligned}$$

Hence, sum of all terms which include $\mathbf{u}_x$ and $\mathbf{v}_y$ is equal to

$$\begin{aligned}\Upsilon_{\mathbf{u},\mathbf{v}} :=& \widetilde{\mathbf{u}}_x^\top\Big(\eta\Sigma_\mu^{-1} + \mathrm{Id}\Big)\widetilde{\mathbf{u}}_x + \widetilde{\mathbf{v}}_y^\top\Big(\eta\Sigma_\nu^{-1} + \mathrm{Id}\Big)\widetilde{\mathbf{v}}_y - 2\widetilde{\mathbf{u}}_x^\top\widetilde{\mathbf{v}}_y + 2\widetilde{\mathbf{u}}_x^\top(\mathbf{u}-\mathbf{v}) - 2\widetilde{\mathbf{v}}_y^\top(\mathbf{u}-\mathbf{v}) + \|\mathbf{u}-\mathbf{v}\|^2 \\ =& \widetilde{\mathbf{u}}_x^\top\Sigma_{\mu,\eta}\widetilde{\mathbf{u}}_x + \widetilde{\mathbf{v}}_y^\top\Sigma_{\nu,\eta}\widetilde{\mathbf{v}}_y - 2\widetilde{\mathbf{u}}_x^\top\widetilde{\mathbf{v}}_y + 2\widetilde{\mathbf{u}}_x^\top(\mathbf{u}-\mathbf{v}) - 2\widetilde{\mathbf{v}}_y^\top(\mathbf{u}-\mathbf{v}) + \|\mathbf{u}-\mathbf{v}\|^2. \qquad (38)\end{aligned}$$

Taking the derivative with respect to $\widetilde{\mathbf{u}}_x$ and $\widetilde{\mathbf{v}}_y$ of $\Upsilon_{\mathbf{u},\mathbf{v}}$, and set the equation to zero we obtain

$$\begin{aligned}\Sigma_{\mu,\eta}\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{v}}_y + (\mathbf{u}-\mathbf{v}) &= \mathbf{0}, \\ \Sigma_{\nu,\eta}\widetilde{\mathbf{v}}_y - \widetilde{\mathbf{u}}_x - (\mathbf{u}-\mathbf{v}) &= \mathbf{0}.\end{aligned}$$

Adding two equations we get

$$\begin{aligned}(\eta\Sigma_\mu^{-1} + \mathrm{Id})\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{v}}_y + (\eta\Sigma_\nu^{-1} + \mathrm{Id})\widetilde{\mathbf{v}}_y - \widetilde{\mathbf{u}}_x &= \mathbf{0}, \\ \Sigma_\mu^{-1}\widetilde{\mathbf{u}}_x &= -\Sigma_\nu^{-1}\widetilde{\mathbf{v}}_y.\end{aligned}$$

Replace them into the above equation, we obtain

$$\begin{aligned}\Sigma_{\mu,\eta}\widetilde{\mathbf{u}}_x + \Sigma_\nu\Sigma_\mu^{-1}\widetilde{\mathbf{u}}_x &= -(\mathbf{u}-\mathbf{v}) \\ \widetilde{\mathbf{u}}_x &= -\Big[\Sigma_{\mu,\eta} + \Sigma_\nu\Sigma_\mu^{-1}\Big]^{-1}(\mathbf{u}-\mathbf{v}) \\ &= -\Sigma_\mu\Big[\eta\,\mathrm{Id} + \Sigma_\mu + \Sigma_\nu\Big]^{-1}(\mathbf{u}-\mathbf{v}).\end{aligned}$$

Similarly, we obtain $\widetilde{\mathbf{v}}_y = \Sigma_\nu\Big[\eta\,\mathrm{Id} + \Sigma_\mu^{-1} + \Sigma_\nu^{-1}\Big]^{-1}(\mathbf{u}-\mathbf{v})$. The value of $\Upsilon_{\mathbf{u},\mathbf{v}}$ at the minimizer is computed as follow

$$\begin{aligned}\widetilde{\mathbf{u}}_x^\top\Sigma_{\mu,\eta}\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{u}}_x^\top\widetilde{\mathbf{v}}_y &= -\widetilde{\mathbf{u}}_x^\top(\mathbf{u}-\mathbf{v}) \\ \widetilde{\mathbf{v}}_y^\top\Sigma_{\nu,\eta}\widetilde{\mathbf{v}}_y - \widetilde{\mathbf{u}}_x^\top\widetilde{\mathbf{v}}_y &= \widetilde{\mathbf{v}}_y^\top(\mathbf{u}-\mathbf{v}).\end{aligned}$$

Adding them together

$$\widetilde{\mathbf{u}}_x^\top\Sigma_{\mu,\eta}\widetilde{\mathbf{u}}_x + \widetilde{\mathbf{v}}_y^\top\Sigma_{\nu,\eta}\widetilde{\mathbf{v}}_y - 2\widetilde{\mathbf{u}}_x^\top\widetilde{\mathbf{v}}_y = -(\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{v}}_y)^\top(\mathbf{u}-\mathbf{v}).$$

Put it back to equation (38)

$$\Upsilon_{\mathbf{u}^*,\mathbf{v}^*} = (\widetilde{\mathbf{u}}_x - \widetilde{\mathbf{v}}_y)^\top(\mathbf{u}-\mathbf{v}) + \|\mathbf{u}-\mathbf{v}\|^2 = (\mathbf{u}-\mathbf{v})^\top\left\{\mathrm{Id} - (\Sigma_\mu+\Sigma_\nu)\Big[\eta\,\mathrm{Id} + \Sigma_\mu^{-1} + \Sigma_\nu^{-1}\Big]^{-1}\right\}(\mathbf{u}-\mathbf{v}).$$

The covariance matrix part: The second part is to group all factors of Σ_x, Σ_y and $\kappa_{xy,i}$, which is

$$\Upsilon_\Sigma := \operatorname{tr}(\Sigma_x) + \operatorname{tr}(\Sigma_y) - 2\operatorname{tr}(K_{xy}) + \eta\Big\{\operatorname{tr}(\Sigma_x\Sigma_\mu^{-1}) - \log\Big(\frac{\det(\Sigma_x)}{\det(\Sigma_\mu)}\Big) + \operatorname{tr}(\Sigma_y\Sigma_\nu^{-1}) - \log\Big(\frac{\det(\Sigma_y)}{\det(\Sigma_\nu)}\Big)\Big\}$$
$$- \frac{\varepsilon}{2}\log\prod_{i=1}^{d}(1-\kappa_{xy,i}). \tag{39}$$

Grouping terms which have the same factors, the above quantity is rewritten as

$$\Upsilon_\Sigma = \operatorname{tr}(\Sigma_x\Sigma_{\mu,\eta}) + \operatorname{tr}(\Sigma_y\Sigma_{\nu,\eta}) - 2\operatorname{tr}(K_{xy}) - \eta\Big[\log\Big(\frac{\det(\Sigma_x)}{\det(\Sigma_\mu)}\Big) + \log\Big(\frac{\det(\Sigma_y)}{\det(\Sigma_\nu)}\Big)\Big] - \frac{\varepsilon}{2}\sum_{i=1}^{d}\log(1-\kappa_{xy,i}). \tag{40}$$

Let $U_{xy}\Lambda_{xy}V_{xy}^\top$ be SVD of $\Sigma_x^{-\frac{1}{2}}K_{xy}\Sigma_y^{-\frac{1}{2}}$ and denote

$$\Sigma_x^{\frac{1}{2}}U_{xy} = X; \qquad V_{xy}^\top\Sigma_y^{\frac{1}{2}} = Y. \tag{41}$$

Then we get

$$XX^\top = \Sigma_x; \qquad Y^\top Y = \Sigma_y; \qquad \Lambda_{xy} = \mathsf{diag}\big(\kappa_{xy,i}^{\frac{1}{2}}\big).$$

Note that if we fix Σ_x, Σ_y, Λ_{xy}, then we only need to maximize the term $\operatorname{tr}(K_{xy})$, which is equal to $\operatorname{tr}(\Lambda_{xy}YX)$. We could choose U_{xy} and V_{xy} such that $YX = \mathsf{diag}(\lambda_{xy,i})$ where $\lambda_{xy,i}$ is the ith singular value of $\Sigma_y^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$. By von Neumann's trace inequality (Kristof, 1969; Horn & Johnson, 1991), it is where the term $\operatorname{tr}(K_{xy})$ attains its maximum. Hence, YX is a diagonal matrix when the objective function Υ_Σ attains its minimum.

We also note that the form $\Sigma_x^{\frac{1}{2}}U_{xy}$ and $V_{xy}^\top\Sigma_y^{\frac{1}{2}}$ are the QR decompositions of X and Y, respectively. Hence there is no condition on X and Y, except that they are invertible. Then Υ_Σ now is equal to

$$\Upsilon_\Sigma = \operatorname{tr}(XX^\top\Sigma_{\mu,\eta}) + \operatorname{tr}(Y^\top Y\Sigma_{\nu,\eta}) - 2\operatorname{tr}(X\Lambda_{xy}Y) - 2\eta\Big[\log\big(\det(X)\det(Y)\big)\Big] - \frac{\varepsilon}{2}\sum_{i=1}^{d}\log(1-\kappa_{xy,i}) + \mathsf{const}.$$

Taking the derivative with respect to X and Y (since in a small local neighbourhood, they are still invertible), we get the following system of equations

$$2\Sigma_{\mu,\eta}X - 2Y^\top\Lambda_{xy}^\top - 2\eta(X^{-1})^\top = \mathbf{0} \tag{42}$$
$$2Y\Sigma_{\nu,\eta} - 2\Lambda_{xy}^\top X^\top - 2\eta(Y^{-1})^\top = \mathbf{0}. \tag{43}$$

We multiply the first equation with $X^\top$ on the left side and the second equation with $Y^\top$ on the right side,

$$X^\top\Sigma_{\mu,\eta}X = (YX)^\top\Lambda_{xy} + \eta\,\mathrm{Id} \tag{44}$$
$$Y\Sigma_{\nu,\eta}Y^\top = \Lambda_{xy}^\top(YX)^\top + \eta\,\mathrm{Id}\,. \tag{45}$$

The right hand side of (44) and (45) are diagonal matrices, since Λ_{xy} and YX are diagonal matrices. It means that $X^\top\Sigma_{\mu,\eta}X$ is also a diagonal matrix. Denote

$$(YX)^\top\Lambda_{xy} + \eta\,\mathrm{Id} = \Lambda_{xy,\eta} := \mathsf{diag}(\lambda_{xy,\eta,i}). \tag{46}$$

Then

$$X^\top\Sigma_{\mu,\eta}X = X^\top U_{\mu,\eta}\Lambda_{\mu,\eta}U_{\mu,\eta}^\top X = \Lambda_{xy,\eta}$$
$$\Big[\Lambda_{\mu,\eta}^{\frac{1}{2}}\Lambda_{xy,\eta}^{-\frac{1}{2}}X^\top U_{\mu,\eta}\Big]\Lambda_{\mu,\eta}\Big[U_{\mu,\eta}^\top X\Lambda_{xy,\eta}^{-\frac{1}{2}}\Lambda_{\mu,\eta}^{\frac{1}{2}}\Big] = \Lambda_{\mu,\eta},$$

where $U_{\mu,\eta}\Lambda_{\mu,\eta}U_{\mu,\eta}^\top$ and $U_{\nu,\eta}\Lambda_{\nu,\eta}U_{\nu,\eta}^\top$ are the spectral decomposition of $\Sigma_{\mu,\eta}$ and $\Sigma_{\nu,\eta}$, respectively. The equation has the form $A\Lambda A^\top = \Lambda$, with Λ is a diagonal matrix. Let $A\Lambda^{\frac{1}{2}} = \Lambda^{\frac{1}{2}}U$, then $\Lambda^{\frac{1}{2}}UU^\top\Lambda^{\frac{1}{2}} = A\Lambda A^\top = \Lambda$. It means that

$UU^\top = \mathrm{Id}$, then U is unitary matrix. Thus we obtain $A = \Lambda^{\frac{1}{2}} U \Lambda^{-\frac{1}{2}}$. Replace A by $\left[\Lambda_{\mu,\eta}^{\frac{1}{2}} \Lambda_{xy,\eta}^{-\frac{1}{2}} X^\top U_{\mu,\eta}\right]$, there exists unitary matrix U such that

$$\begin{aligned}
\Lambda_{\mu,\eta}^{\frac{1}{2}} \Lambda_{xy,\eta}^{-\frac{1}{2}} X^\top U_{\mu,\eta} &= \Lambda_{\mu,\eta}^{\frac{1}{2}} U \Lambda_{\mu,\eta}^{-\frac{1}{2}} \\
\Lambda_{xy,\eta}^{-\frac{1}{2}} X^\top U_{\mu,\eta} &= U \Lambda_{\mu,\eta}^{-\frac{1}{2}} \\
X^\top &= \Lambda_{xy,\eta}^{\frac{1}{2}} U \Lambda_{\mu,\eta}^{-\frac{1}{2}} U_{\mu,\eta}^\top \\
X &= U_{\mu,\eta} \Lambda_{\mu,\eta}^{-\frac{1}{2}} U^\top \Lambda_{xy,\eta}^{\frac{1}{2}}.
\end{aligned} \tag{47}$$

Similarly, there exists an unitary matrix V such that

$$Y = \Lambda_{xy,\eta}^{\frac{1}{2}} V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_{\nu,\eta}^\top. \tag{48}$$

Combine the results with a note that $U_{\mu,\eta} = U_\mu$ and $U_{\nu,\eta} = U_\nu$, we obtain

$$\begin{aligned}
YX &= \Lambda_{xy,\eta}^{\frac{1}{2}} V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_{\nu,\eta}^\top U_{\mu,\eta} \Lambda_{\mu,\eta}^{-\frac{1}{2}} U^\top \Lambda_{xy,\eta}^{\frac{1}{2}} \\
(YX)\Lambda_{xy,\eta}^{-1} &= V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_\nu^\top U_\mu \Lambda_{\mu,\eta}^{-\frac{1}{2}} U^\top := \mathsf{diag}(\alpha_i)_{i=1}^d.
\end{aligned} \tag{49}$$

Here, α_i is the ith singular values of $U_\nu \Lambda_{\nu,\tau,\varepsilon}^{-\frac{1}{2}} U_\nu^\top U_\mu \Lambda_{\mu,\eta}^{-\frac{1}{2}} U_\mu$, where U and V could be chosen such that $V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_\nu^\top U_\mu \Lambda_{\mu,\tau,\varepsilon}^{-\frac{1}{2}} U$ is a diagonal matrix $\mathsf{diag}(\alpha_i)_{i=1}^d$. By the von Neumann's trace inequality (Kristof, 1969; Horn & Johnson, 1991),

$$\mathrm{tr}(K_{xy}) = \mathrm{tr}(\Lambda_{xy} YX) \leq \sum_{i=1}^d \alpha_i \kappa_{xy,i}^{\frac{1}{2}} \lambda_{xy,\eta,i}.$$

The equality happens when $(\kappa_{xy,i})$, (α_i) and $(\lambda_{xy,\eta,i})$ are decreasing sequences with $V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_\nu^\top U_\mu \Lambda_{\mu,\eta}^{-\frac{1}{2}} U$ is diagonal.

Substitute it into the equation (46) with $\Lambda_{xy,\eta} := \mathsf{diag}(\lambda_{xy,\eta,i})_{i=1}^d$

$$\begin{aligned}
\lambda_{xy,\eta,i} &= \eta + \alpha_i \kappa_{xy,i}^{\frac{1}{2}} \lambda_{xy,\eta,i} \\
\lambda_{xy,\eta,i} &= \eta \left[1 - \alpha_i \kappa_{xy,i}^{\frac{1}{2}}\right]^{-1}.
\end{aligned} \tag{50}$$

Combine it with equations (44) and (46), we get

$$\det(X^\top \Sigma_{\mu,\eta} X) = \det(X)^2 \det(\Sigma_{\mu,\eta}) = \eta^d \prod_{i=1}^d \left(1 - \alpha_i \kappa_{xy,i}^{\frac{1}{2}}\right)^{-1}.$$

Therefore,

$$\log\det(X) = -\frac{1}{2} \sum_{i=1}^d \log\left(1 - \alpha_i \kappa_{xy,i}^{\frac{1}{2}}\right) + \mathsf{const}. \tag{51}$$

Similarly, we also have

$$\log\det(Y) = -\frac{1}{2} \sum_{i=1}^d \log\left(1 - \alpha_i \kappa_{xy,i}^{\frac{1}{2}}\right) + \mathsf{const}. \tag{52}$$

Put them (44), (45), (51), (52) together, with Σ^* to be the optimal solution of X, Y, etc, we obtain

$$\Upsilon_{\Sigma^*} = 2\eta \sum_{i=1}^d \log\left(1 - \alpha_i \kappa_{xy,i}^{\frac{1}{2}}\right) - \frac{\varepsilon}{2} \sum_{i=1}^d \log(1 - \kappa_{xy,i}) + \mathsf{const}.$$

Let us consider the function for $\alpha, t \in (0,1)$

$$f(t) = 2\eta \log\left(1 - \alpha t^{\frac{1}{2}}\right) - \frac{\varepsilon}{2}\log(1-t).$$

Taking derivative with respect to t we get

$$f'(t) = \frac{-2\eta\alpha}{2(1-\alpha t^{\frac{1}{2}})t^{\frac{1}{2}}} + \frac{\varepsilon}{2}\frac{1}{1-t} = \frac{\varepsilon(1-\alpha t^{\frac{1}{2}})t^{\frac{1}{2}} - 2\eta\alpha(1-t)}{2(1-\alpha t^{\frac{1}{2}})t^{\frac{1}{2}}(1-t)} = \frac{(2\eta-\varepsilon)\alpha t + \varepsilon t^{\frac{1}{2}} - 2\eta\alpha}{2(1-\alpha t^{\frac{1}{2}})t^{\frac{1}{2}}(1-t)}$$

The nominator is a quadratic function of $t^{\frac{1}{2}}$, which is increasing function taking negative value with $t = 0$ and positive value when $t = 1$, since $0 < \alpha < 1$. Thus, it has only one solution

$$\sqrt{t^*} = \frac{\sqrt{\varepsilon^2 + 8\tau\eta\alpha^2} - \varepsilon}{2\tau\alpha} = \frac{\sqrt{\varepsilon^2/4 + 2\tau\eta\alpha^2} - \varepsilon/2}{\tau\alpha}.$$

It also deduces that if α increases, the t^* also increases. Substituting the result into the equation (50), we obtain

$$\lambda_{xy,\eta,i} = \frac{\eta}{1 - \alpha_i \kappa_{xy,i}^{\frac{1}{2}}} = \frac{\eta}{1 - \frac{\sqrt{\varepsilon^2/4 + \tau(2\eta)\alpha_i^2} - \varepsilon/2}{\tau}}. \tag{53}$$

Due to equation (49), we get $\alpha_i \lambda_{xy,\eta,i} = (YX)_{i,i}$. Combine with equation (46), we get

$$\lambda_{xy,i} = \frac{\lambda_{xy,\eta,i} - \eta}{\alpha_i \lambda_{xy,\eta,i}}. \tag{54}$$

Covariance matrix: Recall from equations (47) and (48), we have

$$X = U_{\mu,\eta}\Lambda_{\mu,\eta}^{-\frac{1}{2}} U^\top \Lambda_{xy,\eta}^{\frac{1}{2}}; \qquad Y = \Lambda_{xy,\eta}^{\frac{1}{2}} V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_{\nu,\eta}^\top.$$

Since equation (41), $V\Lambda_{\nu,\eta}^{-\frac{1}{2}} U_\nu^\top U_\mu \Lambda_{\mu,\eta}^{-\frac{1}{2}} U^\top$ is a diagonal matrix. It means that

$$(VU_\nu^\top)U_\nu \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_\nu^\top U_\mu \Lambda_{\mu,\eta}^{-\frac{1}{2}} U_\mu^\top (U_\mu U) = (VU_\nu^\top)(\Sigma_{\nu,\eta}^{-\frac{1}{2}}\Sigma_{\mu,\eta}^{-\frac{1}{2}})(U_\mu U^\top) = \mathsf{diag}(\alpha_i)_{i=1}^d,$$

which means that the α_i are the singular values of $\Sigma_{\nu,\eta}^{-\frac{1}{2}}\Sigma_{\mu,\eta}^{-\frac{1}{2}}$. Then

$$\Sigma_{\nu,\eta}^{-\frac{1}{2}}\Sigma_{\mu,\eta}^{-\frac{1}{2}} = \left[U_\nu V^\top\right]\mathsf{diag}(\alpha_i)\left[U U_\mu^\top\right] = U_{\mu\nu,\eta}\mathsf{diag}(\alpha_i)V_{\mu\nu,\eta}^\top$$

which is a SVD of $\Sigma_{\nu,\eta}^{-\frac{1}{2}}\Sigma_{\mu,\eta}^{-\frac{1}{2}}$, where $U_\mu U^\top = V_{\mu\nu,\eta}$ and $VU_\nu^\top = U_{\mu\nu,\eta}^\top$. Put them into the equation (41)

$$\begin{aligned}
K_{xy} = X\Lambda_{xy}Y &= U_\mu \Lambda_{\mu,\eta}^{-\frac{1}{2}} U^\top \Lambda_{xy,\eta}^{\frac{1}{2}} \Lambda_{xy} \Lambda_{xy,\eta}^{\frac{1}{2}} V \Lambda_{\nu,\eta}^{-\frac{1}{2}} U_\nu^\top = \Sigma_{\mu,\eta}^{-\frac{1}{2}}\left[U_\mu U^\top\right]\left[\Lambda_{xy,\eta}^{\frac{1}{2}}\Lambda_{xy}\Lambda_{xy,\eta}^{\frac{1}{2}}\right]\left[VU_\nu^\top\right]\Sigma_{\nu,\eta}^{-\frac{1}{2}} \\
&= \Sigma_{\mu,\eta}^{-\frac{1}{2}}\left[V_{\mu\nu,\eta}\mathsf{diag}\Big(\frac{\lambda_{xy,\eta,i}-\eta}{\alpha_i}\Big)_{i=1}^d U_{\mu\nu,\eta}^\top\right]\Sigma_{\nu,\eta}^{-\frac{1}{2}} \\
&= \eta\Sigma_{\mu,\eta}^{-\frac{1}{2}} A(\mathrm{Id} - A)^{-1}(B^\top B)^{-\frac{1}{2}} B^{-1}\Sigma_{\nu,\eta}^{-\frac{1}{2}}.
\end{aligned} \tag{55}$$

We could obtain it from

$$\begin{aligned}
B = U_{\mu\nu,\eta}\mathsf{diag}(\alpha_i)_{i=1}^d V_{\mu\nu,\eta}^\top &\Rightarrow B^\top B = V_{\mu\nu,\eta}\mathsf{diag}(\alpha_i^2)_{i=1}^d V_{\mu\nu,\eta}^\top \\
A &= \frac{1}{\tau}V_{\mu\nu,\eta}\mathsf{diag}\Big(\sqrt{\varepsilon^2/4 + \tau(2\eta)\alpha_i^2} - \varepsilon/2\Big)_{i=1}^d V_{\mu\nu,\eta}^\top \\
&= \frac{1}{\tau}\left\{\left[\frac{\varepsilon^2}{4}\,\mathrm{Id} + \tau(2\eta)B^\top B\right]^{\frac{1}{2}} - \frac{\varepsilon}{2}\,\mathrm{Id}\right\} \\
\eta(\mathrm{Id} - A)^{-1} &= V_{\mu\nu,\eta}\mathsf{diag}(\lambda_{xy,\eta,i})_{i=1}^d V_{\mu\nu,\eta}^\top \\
\eta\left\{(\mathrm{Id} - A)^{-1} - \mathrm{Id}\right\}(B^\top B)^{\frac{1}{2}} &= V_{\mu\nu,\eta}\mathsf{diag}\left(\frac{\lambda_{xy,\eta,i}-\eta}{\alpha_i}\right)_{i=1}^d V_{\mu\nu,\tau}^\top.
\end{aligned}$$

Multiply both side with $B^{-1} = V_{\mu\nu,\eta}\mathsf{diag}(\alpha_i)U_{\mu\nu,\eta}^\top$, we obtain the formula in (55).

For the Σ_x, multiply with $X^\top$ to the right side of equation (42), we have

$$\begin{aligned}\Sigma_{\mu,\eta}XX^\top - Y^\top\Lambda_{xy}^\top X^\top - \eta(X^{-1})^\top X^\top &= \mathbf{0}\\ \Sigma_{\mu,\eta}\Sigma_x &= (X\Lambda_{xy}Y)^\top + \eta\,\mathrm{Id}\\ \Sigma_x &= \Sigma_{\mu,\eta}^{-1}(K_{xy}^\top + \eta\,\mathrm{Id}).\end{aligned} \tag{56}$$

Another way to derive Σ_x is to use equation (47) of X and equation (53) of $\lambda_{xy,\eta}$

$$\begin{aligned}\Sigma_x = XX^\top &= U_\mu\Lambda_{\mu,\eta}^{-\frac{1}{2}}U^\top\Lambda_{xy,\eta}U\Lambda_{\mu,\eta}^{-\frac{1}{2}}U_\mu^\top = \Sigma_{\mu,\eta}^{-\frac{1}{2}}U_\mu U^\top\Lambda_{xy,\eta}UU_\mu^\top\Sigma_{\mu,\eta}^{-\frac{1}{2}}\\ &= \Sigma_{\mu,\eta}^{-\frac{1}{2}}V_{\mu\nu,\eta}\Lambda_{xy,\eta}V_{\mu\nu,\eta}^\top\Sigma_{\mu,\eta}^{-\frac{1}{2}}\\ &= \eta\Sigma_{\mu,\eta}^{-\frac{1}{2}}A(B^\top B)^{-\frac{1}{2}}\Sigma_{\mu,\eta}^{-\frac{1}{2}},\end{aligned} \tag{57}$$

where the last equation is obtained from the equation (54)

$$\begin{aligned}V_{\mu\nu,\eta}\mathsf{diag}(\lambda_{xy,i})_{i=1}^d V_{\mu\nu,\eta}^\top &= \eta\big\{(\mathrm{Id}-A)^{-1} - \mathrm{Id}\,\big\}(\mathrm{Id}-A)(B^\top B)^{-\frac{1}{2}} = \eta\big\{\,\mathrm{Id} - (\mathrm{Id}-A)\big\}(B^\top B)^{-\frac{1}{2}}\\ &= \eta A(B^\top B)^{-\frac{1}{2}}.\end{aligned}$$

Similarly for equation (43)

$$\begin{aligned}Y^\top Y\Sigma_{\nu,\eta} - Y^\top\Lambda_{xy}^\top X^\top - \eta\,\mathrm{Id} &= \mathbf{0}\\ \Sigma_y &= (K_{xy}^\top + \eta\,\mathrm{Id})\Sigma_{\nu,\eta}^{-1}.\end{aligned} \tag{58}$$

or we could derive another way to obtain $\Sigma_y = \eta\Sigma_{\nu,\eta}^{-\frac{1}{2}}A(B^\top B)^{-\frac{1}{2}}\Sigma_{\nu,\eta}^{-\frac{1}{2}}$.

Next we calculate the Υ_{Σ^*}. Since equations (56) and (58), we have

$$\mathrm{tr}(\Sigma_x\Sigma_{\mu,\eta}) + \mathrm{tr}(\Sigma_y\Sigma_{\nu,\eta}) - 2\,\mathrm{tr}(K_{xy}) = 2\eta.$$

We also have

$$\begin{aligned}\prod_{i=1}^d(1-\kappa_{xy,i}) &= \det\big(\,\mathrm{Id} - \mathsf{diag}(\kappa_{xy,i})_{i=1}^d\big)\\ V_{\mu\nu,\eta}\mathsf{diag}(\kappa_{xy,i})_{i=1}^d V_{\mu\nu,\eta}^\top &= A^2(B^\top B)^{-1}.\end{aligned}$$

It means that $\sum_{i=1}^d\log(1-\kappa_{xy,i})_{i=1}^d = \log\det\big(\,\mathrm{Id} - A^2(B^\top B)^{-1}\big)$. Put all the results into equation (40), we obtain

$$\begin{aligned}\Upsilon_\Sigma &= \mathrm{tr}(\Sigma_x\Sigma_{\mu,\eta}) + \mathrm{tr}(\Sigma_y\Sigma_{\nu,\eta}) - 2\,\mathrm{tr}(K_{xy}) - \eta\Big[\log\Big(\frac{\det(\Sigma_x)}{\det(\Sigma_\mu)}\Big) + \log\Big(\frac{\det(\Sigma_y)}{\det(\Sigma_\nu)}\Big)\Big] - \frac{\varepsilon}{2}\sum_{i=1}^d\log(1-\kappa_{xy,i})\\ \Upsilon_{\Sigma^*} &= 2\eta - 2\eta\Big\{\log(\eta) + \log\det(A) - \log\det(B) - \frac{1}{2}\log\det(\Sigma_\mu) - \frac{1}{2}\log\det(\Sigma_\nu)\Big\}\\ &\quad - \frac{\varepsilon}{2}\log\det\big(\,\mathrm{Id} - A^2(B^\top B)^{-1}\big)\\ &= 2\eta\Big\{1 - \log(\eta) + \log\det(A) - \log\det(B) - \frac{1}{2}\log\det(\Sigma_\mu\Sigma_\nu)\Big\} - \frac{\varepsilon}{2}\log\det\big(\,\mathrm{Id} - A^2(B^\top B)^{-1}\big).\end{aligned}$$

Step 2: Calculation of m_{π^*} We need to find the minimizer of the following problem:

$$\min_{m_\pi>0}\; m_\pi\Upsilon^* + \tau\mathrm{KL}(m_\pi\|m_\mu) + \tau\mathrm{KL}(m_\pi\|m_\nu) + \varepsilon\mathrm{KL}(m_\pi\|m_\mu m_\nu).$$

By part (c) of Lemma C.6, we obtain

$$m_{\pi^*} = m_\mu^{\frac{\tau}{2\tau+\varepsilon}}m_\nu^{\frac{\tau}{2\tau+\varepsilon}}(m_\mu m_\nu)^{\frac{\varepsilon}{2\tau+\varepsilon}}\exp\Big\{\frac{-\Upsilon^*}{2\tau+\varepsilon}\Big\}.$$

Hence, we have thus proved our claims. □