On Barycenter Computation: Analyzing Semi-Unbalanced Optimal Transport-based Method on Bures-Wasserstein Manifold

Ngoc-Hai Nguyen[⋆‡] **Dung Le**[⋆†] **Hoang-Phi Nguyen**[⋄] **Tung Pham**[∘] **Nhat Ho**[†]
Tufts University[‡] The University of Texas at Austin[†] Monash University[⋄] Qualcomm AI Research[∘]

Abstract

We explore a robust version of the barycenter problem among n centered Gaussian probability measures, termed Semi-Unbalanced Optimal Transport (SUOT)-based Barycenter, wherein the barycenter remains fixed while the others are relaxed using Kullback-Leibler divergence. We develop optimization algorithms on Bures-Wasserstein manifold, named the Exact Geodesic Gradient Descent and Hybrid Gradient Descent algorithms. While the Exact Geodesic Gradient Descent method is based on computing the exact closed form of the first-order derivative of the objective function of the barycenter along a geodesic on the Bures manifold, the Hybrid Gradient Descent method utilizes optimizer components when solving the SUOT problem to replace contaminated measures before applying the Riemannian Gradient Descent. We establish the theoretical convergence guarantees for both methods and demonstrate that the Exact Geodesic Gradient Descent algorithm attains a dimension-free convergence rate. This is a novel theoretical result for Riemannian Gradient Descent applicable to an expanded class of averaging functions.

1 INTRODUCTION

Aggregating multiple data sources has garnered significant interest in data science and artificial intelligence due to its fundamental role and wide range of applications. When we have to work with data distribution, one of the useful aggregations is the barycenter of those distributions. In the context of optimal transport (Peyré et al., 2019), where the distance between distributions is defined as the optimal cost to transport masses, this problem is known as the Wasserstein Barycenter problem, which has several applications, including image processing (Ferradans et al., 2014) (Simon and Aberdam, 2020), image restoration (Mignon et al., 2023), time-series modeling (Cheng et al., 2021), domain adaptation (Montesuma and Mboula, 2021), graph representation (Simou et al., 2020), signal processing (Simou and Frossard, 2019), matching (Naas et al., 2024) and medical multi-modal large language model (Nguyen et al., 2024).

However, a common challenge arises as data often exhibit more complexity, especially in real-world scenarios where noises and outliers are prevalent, that could distort the final results of any statistical procedure. In this paper, we work on the robustness of the barycenter when the data measures contain noise, akin to extracting the true mean of clean distributions. It is known that a relaxed version of OT, which is Unbalanced Optimal Transport (UOT) (Liero et al., 2018), is able to reduce the effect of contamination, thus producing robust estimation of the OT cost between corrupted distributions. The penalty function often used in the UOT is the KL divergence, which has been well-studied (Nguyen et al., 2021; Pham et al., 2020) and is preferred because of its pleasant mathematical properties and computational advantage over other divergences. Motivated by those mathematical formulas of UOT costs, we aim to find the barycenter in the contaminated distributions setting.

The central questions are: (i) how to mathematically formulate this problem in a meaningful way, and (ii) whether standard optimization techniques, such as gradient descent, are effective for our formulation across specific or general classes of distributions. In particular, when the underlying data distributions are Gaussians, the difficulty arises because the minimization is carried out over the Bures manifold (Bhatia, 2009)—the manifold of symmetric positive definite matrices—which possesses positive curvature. Our work makes these

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. ∘: Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. ⋆: Equal contribution

two primary contributions:

1. We propose a framework for barycenter computation based on the debiased effect of Semi-Unbalanced Optimal Transport (SUOT) to measure distances between the barycenter and other distributions. By establishing that the barycenter of Gaussian distributions under SUOT remains Gaussian, the problem reduces to optimizing its mean and covariance. Building on this, we propose two efficient methods for Gaussian barycenter computation: (i) a hybrid scheme inspired by Chewi et al. (2020), and (ii) a Riemannian Gradient Descent algorithm leveraging the exact Wasserstein gradient, enabled by a closed-form expression for the SUOT distance between weighted Gaussians $\alpha_i = m_{\alpha_i}, \mathcal{N}(\mathbf{a}_i, \Sigma\alpha_i)$ in $\mathbb{R}^d$.
2. We address a challenging and meaningful theoretical question: does the dimension-independent convergence rate of Gradient Descent on the Bures manifold, established in Altschuler et al. (2021), extend to an expanded class of averaging functions? We prove that it does by providing convergence guarantees with a dimension-free rate comparable to that of Altschuler et al. (2021) on SUOT-based Barycenter. Our proof technique is non-trivial and generalizes to a far more complex and non-symmetric setting, where analyzing smoothness and geodesic convexity becomes significantly more challenging.

2 BACKGROUND

Notations. Let $\mathcal{P}_{2,ac}\left(\mathbb{R}^d\right)$ denote the space of absolutely continuous probability measures on $\mathbb{R}^d$ with finite second moment. For any $\mu, \nu \in \mathcal{P}_{2,ac}\left(\mathbb{R}^d\right)$, the generalized Kullback-Leibler divergence is defined as $\mathrm{KL}(\mu\|\nu) = \int_{\mathbb{R}^d} \mu \log\left(\frac{\mu}{\nu}\right) \mathrm{d}x$. We denote the set of symmetric matrices by $\mathbb{S}^d$, symmetric semi-positive definite matrices by $\mathbb{S}^d_+$, and symmetric positive definite matrices by $\mathbb{S}^d_{++}$. The singular values of $\Sigma \in \mathbb{S}^d_+$ in descending order are $\{\lambda_i(\Sigma)\}_{i=1}^d$. The Gaussian measure on $\mathbb{R}^d$ with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}^d_{++}$ is denoted $\mathcal{N}(m, \Sigma)$. The identity matrix in $\mathbb{R}^{d\times d}$ is Id. For a vector $x \in \mathbb{R}^d$, $\mathbf{diag}(x)$ is the diagonal matrix with x on the diagonal. We use $\|.\|_2$ as l_2-norm of vector in $\mathbb{R}^d$ and $\|.\|_F$ as Frobenius norm of matrix in $\mathbb{R}^{d\times d}$.

2.1 Wasserstein distance

Given probability measures $\mu, \nu \in \mathcal{P}_{2,ac}\left(\mathbb{R}^d\right)$ with finite second moments, the p-Wasserstein distance between μ and ν is defined as (Kantorovich, 1942; Peyré et al., 2019; Villani et al., 2009):

$$W_p^p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi \|X - Y\|_p^p,$$

where X, Y are independent random vectors in $\mathbb{R}^d$ such that $(X, Y) \sim \pi$, $\Pi(\mu, \nu)$ denotes the set of couplings of μ and ν, i.e., the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are respectively μ and ν. If μ and ν have densities with respect to the Lebesgue measure on $\mathbb{R}^d$, the infimum is attained, and the optimal coupling is supported on the graph of a map $T : \mathbb{R}^d \to \mathbb{R}^d$, such that for π-a.e. $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, $y = T(x)$. This map T is the optimal transport map from μ to ν. Recently, Korotin et al. (2022); Mokrov et al. (2023) developed neural network algorithms to compute OT maps and plans for strong and weak costs. For Gaussians, the Wasserstein distance has a closed-form expression Altschuler et al. (2021)

$$\begin{aligned} &W_2^2(\mathcal{N}(m, \Sigma), \mathcal{N}(m', \Sigma')) \\ &= \|m - m'\|_2^2 + \mathsf{tr}\Big(\Sigma + \Sigma' - 2\left[\Sigma^{1/2}\Sigma'\Sigma^{1/2}\right]^{1/2}\Big). \end{aligned}$$

We focus on centered Gaussians, represented by their covariance matrices on Bures-Wasserstein (SPD) manifold Bhatia (2009); Takatsu (2011). Gaussians cover general location-scatter families and play a central role in Optimal Transport Altschuler et al. (2021); Bunne et al. (2023); Chewi et al. (2020); Han et al. (2021), while SPD matrices have broad applications Gao et al. (2020); Herath et al. (2017); Kobler et al. (2022). All Gaussians considered are non-degenerate. For a distance function d, we write $d(\Sigma, \Sigma')$ for the distance between two centered Gaussians with covariances Σ and Σ' (e.g., $W_2(\Sigma, \Sigma')$), and $T_{\Sigma\to\Sigma'}$ for the corresponding optimal transport map.

Barycenter Problem. Let P be a probability measure over $\mathcal{P}_{2,ac}\left(\mathbb{R}^d\right)$. Then, the Wasserstein Barycenter of P is a solution of

$$\underset{\mu\in\mathcal{P}_{2,ac}(\mathbb{R}^d)}{\text{minimize}} \quad \int W_2^2(\mu, \cdot)\mathrm{d}P.$$

Cuturi and Doucet (2014) formed the basis for barycenter computation, proposing an algorithm based on the dual problem and applying it to clustering and perturbed images. Additional methods can be found in Chi et al. (2023); Korotin et al. (2021); Noble et al. (2024). A related notion of average is the entropically-regularized Wasserstein Barycenter of P, which is defined to be a solution of

$$\underset{\mu\in\mathcal{P}_{2,ac}(\mathbb{R}^d)}{\text{minimize}} \quad \int W_2^2(\mu, \cdot)\mathrm{d}P + \mathcal{H}(\mu),$$

where $\mathcal{H}$ is (negative) differential entropy i.e. $\mathcal{H}(\mu) := \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}\right) \mathrm{d}\mu(x)$ Chizat (2023). Benamou et al.

(2015) introduced Iterative Bregman Projections for solving entropic OT and applied it to the weighted barycenter problem. In addition, Bonneel et al. (2015) extended the barycenter problem to Radon barycenters (using the Radon transform) and Sliced barycenters (based on Sliced Wasserstein distance), with applications in color manipulation.

Unbalanced Optimal Transport. Unbalanced Optimal Transport (UOT) Liero et al. (2018) relaxes marginal constraints via Kullback–Leibler regularization. It solves

$$\inf_{\pi \in \mathcal{M}^+(\mathbb{R}^d \times \mathbb{R}^d)} \mathbb{E}_\pi \|X - Y\|_2^2 + \tau \mathrm{KL}\left(\pi_x \| \mu\right) + \tau \mathrm{KL}\left(\pi_y \| \nu\right),$$

where X, Y are independent random vectors in $\mathbb{R}^d$ such that $(X, Y) \sim \pi$, $\mathcal{M}^+\left(\mathbb{R}^d \times \mathbb{R}^d\right)$ is the set of all positive measures in $\mathbb{R}^d \times \mathbb{R}^d$ and π_x, π_y are the marginal distributions of the coupling π corresponding to μ, ν, respectively. $\tau > 0$ is regularized parameter. Iterative scaling methods Chizat et al. (2016) and majorization–minimization approaches Chapel et al. (2021) have been proposed for its solution. UOT is a robust extension of OT, accommodating measures with unequal mass and showing resilience to local variations, outliers, and missing components Blondel et al. (2018); Séjourné et al. (2023); Vincent-Cuaz et al. (2021).

2.2 Riemannian Gradient Descent

Optimization on SPD manifold, common in machine learning Gao et al. (2020), is challenging due to its positive curvature. Unlike Euclidean methods, Riemannian gradient algorithms respect the manifold's geometry, using tangent spaces and Riemannian operators for optimization Absil et al. (2008); Gao et al. (2020). In our case, the Riemannian gradient space at point Σ of SPD manifolds denoted by $T_\Sigma \mathbb{S}_{++}^d$ is identified with the space $\mathbb{S}^d$. The inner product for two matrices A, B, is defined as $\langle A, B \rangle_\Sigma := \mathrm{tr}\left(A^\top \Sigma B\right)$ which would induce the tangent space norm $\|.\|_\Sigma$. The Riemannian exponential map $\mathsf{Exp}_\Sigma(\cdot) : T_\Sigma \mathbb{S}_{++}^d \to \mathbb{S}_{++}^d$ maps a tangent vector to a constant-speed geodesic, while the logarithmic map $\mathsf{Log}_\Sigma(\cdot) : \mathbb{S}_{++}^d \to T_\Sigma \mathbb{S}_{++}^d$ is its inverse. Specifically, they are given by

$$\mathsf{Exp}_\Sigma(X) = (\mathsf{Id} + X)\,\Sigma\,(\mathsf{Id} + X)$$
$$\mathsf{Log}_\Sigma(\Sigma') = T_{\Sigma \to \Sigma'} - \mathsf{Id}.$$

On optimizing a differentiable function $f : \mathcal{M} \to \mathbb{R}$, we denote traditional Euclidean gradient as ∇f and Riemannian gradient as $\mathrm{grad}\, f$ is a tangent vector that satisfies for any $\mu \in T_x \mathcal{M}, \langle \mathrm{grad}\, f(x), \mu \rangle_x = \mathrm{D}_\mu f(x)$, where $\mathrm{D}_\mu f(x)$ is the directional derivative of $f(x)$ along μ. Instead of using traditional gradient descent update $\mu^{(t+1)} = \mu^{(t)} - \eta_t \nabla f(\mu^{(t)})$ which cannot guarantee that the update remains on the manifold, Riemannian Gradient Descent Chewi et al. (2020) reads the update

$$\mu^{(t+1)} = \mathsf{Exp}_{\mu^{(t)}}\Big(- \eta_t \,\mathrm{grad}\, f\big(\mu^{(t)}\big)\Big),$$

for some step size η_t. In this way, it ensures that the updates are along the geodesic and stay on the manifolds Gao et al. (2020). In the context of the Bures-Wasserstein manifold, we also call the gradient as Wasserstein gradient.

3 SEMI-UNBALANCED OPTIMAL TRANSPORT-BASED BARYCENTER

We define the distance $W_{2,\mathrm{SUOT}}^2(\alpha, \beta, \tau)$ between two measures α and β as

$$W_{2,\mathrm{SUOT}}^2(\alpha, \beta, \tau) := \inf_{\pi \in \mathcal{M}^+(\mathbb{R}^d \times \mathbb{R}^d)} \mathbb{E}_\pi \|X - Y\|_2^2 + \tau \mathrm{KL}\left(\pi_x \| \alpha\right), \quad (1)$$

where X, Y are independent random vectors in $\mathbb{R}^d$ such that $(X, Y) \sim \pi, Y \sim \beta$ while $\tau > 0$ is regularization parameter, and π_x is the marginal distribution of the coupling π corresponding to α. Now consider a set of probability measures $\{\alpha_i\}_{i=1}^n$ where the measures are contaminated. To address the above barycenter problem, we proposed the empirical SUOT-based Barycenter which is a minimization problem due to probability measure β

$$\arg\min_\beta \sum_{i=1}^n w_i W_{2,\mathrm{SUOT}}^2(\alpha_i, \beta, \tau), \quad (2)$$

where $\{w_i\}_{i=1}^n$ are given weights satisfying $w_i \geq 0; \sum_{i=1}^n w_i = 1$. Throughout most of this work, we assume uniform weights, setting $w_1 = w_2 = \cdots = w_n = \frac{1}{n}$. The intuition behind relaxing only one marginal constraint lies in the context of computing distances from measures to their barycenter. In this scenario, the data are contaminated, making the true distributions inaccessible. To address this, KL divergence is used to penalize the contaminated side, relaxing the marginal constraint. However, our goal is to find the *barycenter of the true (clean) distributions*, which inherently remains free of noise and outliers. Thus, no penalty needs to be applied to its side, making Semi-UOT the appropriate choice over UOT as commonly used in previous work. This relaxation facilitates the detection of outliers by adapting the measures slightly, ensuring that the barycenter remains representative of the overall distribution without being overly influenced by individual outliers. The SUOT-based barycenter approach provides a robust framework for handling noise and outliers in data distributions, making it highly reliable for

real-world applications. This is particularly valuable for tasks like robust learning, data augmentation, interpolation, and other downstream AI applications Séjourné et al. (2023). A toy example in Figure 2a demonstrates the effectiveness of the Semi-UOT approach, highlighting its robustness and practical relevance.

4 ANALYSIS ON BURES-WASSERSTEIN MANIFOLD

Comparison with Related work. To highlight the contribution of our work, the first part of this section is to clarify the difference between this paper and other highly related works.

1. Broader more than Chewi et al. (2020) and Altschuler et al. (2021), we generalize the barycenter objective to the SUOT framework, which recovers the classical OT problem of them as $\tau \to \infty$ while preserving dimension-free convergence. Our main contribution is to establish explicit convergence guarantees for Riemannian gradient methods over an expanded class of barycenter problems. The technical challenge arises because the SUOT closed form is substantially more complex and nonsymmetric compared to OT, making smoothness and convexity analysis as well as gradient derivation significantly harder. Unlike prior work, where the gradient reduces to a simple derivative on the Bures manifold (with the form $\frac{1}{2}W_2^2(A + tC, B)$), our setting requires handling minimization over SPD matrices with nontrivial operations such as matrix square roots, necessitating advanced tools like Lyapunov's equation.

2. In Álvarez-Esteban et al. (2016), they proposed a fixed-point approach to solve the standard barycenter for Gaussians. Álvarez et al.'s approach is based on equating the first derivative of standard OT to zero at optimal points to derive equations for the optimum. However, the closed-form expressions for SUOT in Theorem 4.2 and its first-order derivative in Theorem 4.4 are significantly more complex than those in standard OT, rendering such an approach infeasible in our case.

3. Janati et al. (2020) and Mallasto et al. (2022) derive closed-form expressions for unbalanced Entropic Optimal Transport (EOT) between Gaussian measures. Specifically, Janati et al. (2020) obtained a closed-form expression for unbalanced OT solutions but relied on the dual form of the objective function. In contrast, our approach directly addresses the primal objective function, resulting in a shorter and more streamlined proof. Similarly, Mallasto et al. (2022) used the dual form of the objective function to derive solutions by solving a system of derivative equations.

4. Our work and Gazdieva et al. (2024) can be viewed as the first parallel efforts to introduce and study SUOT for barycenters. Despite starting from a similar problem formulation, the two works diverge significantly in terms of theoretical focus and empirical goals. While their work then pursues a solver with numerical approximation at many steps and focus on applications, our work concentrates on the geometric structure, deriving an exact closed-form algorithm and aiming to fully develop the underlying theory. Additionally, Gazdieva et al. (2024) is formulated on a general measure framework, while our work specializes to the case of Gaussian measures. Our restriction to Gaussian measure indeed does not diminish the generality or applicability of the problem due to the universality of Gaussian distribution in the real application. In the theoretical aspect: their work provides theoretical support via a dual formulation by constructing an appropriate dual function and recasting the original problem into a maximin framework. However, due to the generality of their setting, neural network parameterization, and stochastic and heuristic components of their algorithmic design, they do not provide a rigorous proof for the convergence. Meanwhile, our work develops a Riemannian gradient descent method in Bure-Wasserstein manifold, and we provide a linear convergence rate to optimal solution. Moreover, the experiments in (Gazdieva et al., 2024) are designed to demonstrate the effectiveness of SUOT barycenters for downstream tasks such as generative modelling. Our experiments, on the other hand, focus on the optimization behavior and training dynamics, serving primarily to support and validate our theoretical results.

4.1 Semi-Unbalanced Optimal Transport has a Closed-Form Expression for Gaussians

Agueh and Carlier (2011) showed that the normal Wasserstein Barycenter of Gaussians distributions, for ℓ_2 cost is also a Gaussian. In our work, we also demonstrate a similar result for SUOT-based Barycenter.

Theorem 4.1. *Let $(\alpha_i)_{i=1}^n$ be zero-mean Gaussian distributions in $\mathbb{R}^d$ which have covariance matrices $(\Sigma_i)_{i=1}^n$. Let $\Sigma_\beta \in \mathbb{S}_{++}^d$ be a SPD matrix. Consider the SUOT-based Barycenter problem in Equation* (2)

$$\arg\min_{\beta \in \mathcal{P}(\Sigma_\beta)} L = \frac{1}{n} \sum_{i=1}^{n} W_{2_{\mathrm{SUOT}}}^2(\alpha_i, \beta, \tau),$$

where $\mathcal{P}(\Sigma_\beta)$ is the set of zero-mean probability distributions in $\mathbb{R}^d$ which have covariance matrix Σ_β. Then, β is itself a Gaussian.

The result of Theorem 4.1 states that within the family of zero-mean distributions sharing the same covariance matrix, Gaussian is the optimal solution. In subsequent development in this work, this does not impose an additional covariance constraint; rather, it shows that for whatever covariance the SUOT-barycenter optimization selects, the optimal distribution within that class is necessarily Gaussian, which justifies narrowing the problem to the Gaussian family. From a technical perspective, this result asserts that the Gaussian form of the barycenter is preserved when the Kullback–Leibler (KL) divergence is incorporated into the formula. This holds because the minimum of the KL divergence, given specified means and covariance matrices, results in a Gaussian distribution. We give the proof for the Theorem 4.1 in Appendix A.

Then, Theorem 4.2 below shows the closed form for the solution of problem (1) in the case of Gaussians.

Theorem 4.2. *Consider two Gaussian measures with masses in $\mathbb{R}^d$: $\alpha = m_\alpha \mathcal{N}(\mathbf{a}, \Sigma_\alpha)$ and $\beta = m_\beta \mathcal{N}(\mathbf{b}, \Sigma_\beta)$. Assume that $\Sigma_\alpha, \Sigma_\beta$ are SPD matrices. Consider minimization problem (1). Assume that optimal solution π^* is a positive measure such that $\pi^* = m_\pi \overline{\pi}$ with $\overline{\pi}$ is a probability measure with mean $(\mathbf{a}_x, \mathbf{b})$ and covariance matrix*

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

We denote

$$\begin{aligned}
\Sigma_{\alpha,\tau} &= \mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}, \qquad \Sigma_{\alpha,\tau,\beta} = \Sigma_\beta^{-\frac{1}{2}} \Sigma_{\alpha,\tau} \Sigma_\beta^{-\frac{1}{2}}, \\
\Sigma_\gamma &= \frac{\tau}{2}\mathsf{Id} + \frac{1}{2}\Sigma_{\alpha,\tau,\beta}^{-1}\left(\mathsf{Id} + (\mathsf{Id} + \tau\Sigma_{\alpha,\tau,\beta})^{\frac{1}{2}}\right), \\
\mathbf{S}_1 &= \frac{\tau}{2}\Sigma_{\alpha,\tau,\beta}^{-1} + \frac{1}{2}\Sigma_{\alpha,\tau,\beta}^{-2}\left(\mathsf{Id} + (\mathsf{Id} + 2\tau\Sigma_{\alpha,\tau,\beta})^{\frac{1}{2}}\right) \\
\mathbf{S}_2 &= \mathsf{tr}(\Sigma_\gamma) + \mathsf{tr}(\Sigma_\beta) + \mathsf{tr}\left(\left[\Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\gamma\right]^{\frac{1}{2}}\right), \\
\mathbf{S}_3 &= -\frac{\tau}{2}\log\left(\det\left[\Sigma_\gamma \Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\beta^{-1}\Sigma_\alpha^{-1}\right]\right), \\
\mathbf{S}_4 &= \left(\left\|\Sigma_{\alpha,\tau}^{-1}\right\|_F^2 + 1\right)\mathsf{Id} - 2\Sigma_{\alpha,\tau}^{-1} + \frac{\tau}{2}\Sigma_{\alpha,\tau}^{-1}\Sigma_\alpha^{-1}\Sigma_{\alpha,\tau}^{-1}, \\
\mathbf{S}_5 &= (\mathbf{a}-\mathbf{b})^\top \mathbf{S}_4 (\mathbf{a}-\mathbf{b}) - \frac{\tau d}{2}, \\
\Upsilon &= \mathbf{S}_2 + \mathbf{S}_3 + \mathbf{S}_5.
\end{aligned}$$

Then, we find that

$$\mathbf{a}_x = \Sigma_{\alpha,\tau}^{-1}(\mathbf{b}-\mathbf{a}) + \mathbf{a}, \quad \Sigma_x = \Sigma_\beta^{-\frac{1}{2}}\mathbf{S}_1\Sigma_\beta^{-\frac{1}{2}},$$
$$K_{x,\beta} = \Lambda, \quad m_\pi = m_\alpha \exp\left\{\frac{-\Upsilon}{\tau}\right\},$$

where Λ is a diagonal matrix containing singular values of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$ in descending order. Moreover, for two centered Gaussian distributions α, β $(m_\alpha = m_\beta = 1; \mathbf{a} = \mathbf{b} = \mathbf{0})$, we have

$$W_{2_{\mathrm{SUOT}}}^2(\Sigma_\alpha, \Sigma_\beta) = W_2^2(\Sigma_x, \Sigma_\beta) + \tau \mathrm{KL}(\Sigma_x \| \Sigma_\alpha),$$

Our proof is given in Appendix B.1. On the other hand, we present a proposition linking our optimizers Σ_x with Σ_β and Σ_α through hyperparameter τ in Appendix E. In Theorem 4.2, m_α and m_β are the scales of Gaussian measures, when the scale is equal 1, we obtain a Gaussian distribution. It also becomes apparent in the proof that SUOT distance between two entities can be succinctly expressed as the sum of the 2-Wasserstein distance and a KL divergence term. In the same way, we could derive the closed-form for sparse solution achieved by Semi-Unbalanced Entropic Optimal Transport, which is given by adding a regularization term $\mathrm{KL}(\pi \| \alpha \otimes \beta)$. In particular, we have Theorem 4.3 as follows.

Theorem 4.3. *Consider two centered Gaussian measures in $\mathbb{R}^d$:*

$$\alpha = \mathcal{N}(\mathbf{0}, \Sigma_\alpha) \text{ and } \beta = \mathcal{N}(\mathbf{0}, \Sigma_\beta).$$

Assume that $\Sigma_\alpha, \Sigma_\beta$ are SPD matrices. Consider minimization problem

$$\begin{aligned}
&W_{2_{\mathrm{SUOT}},\delta}^2(\alpha, \beta; \tau) \\
&:= \min_\pi \mathbb{E}_\pi \|X - Y\|^2 + \tau \mathrm{KL}(\pi_x \| \alpha) + \delta \mathrm{KL}(\pi \| \alpha \otimes \beta), \\
&\qquad s.t\ \pi \in \mathcal{M}^+\left(\mathbb{R}^d \times \mathbb{R}^d\right),
\end{aligned}$$

where X, Y are independent random vectors in $\mathbb{R}^d$ such that $(X, Y) \sim \pi, Y \sim \beta$ while $\tau > 0$ is regularized parameter, and π_x is the marginal distribution of the coupling π corresponding to α. We note that $\Sigma_{\alpha\otimes\beta} = \begin{pmatrix} \Sigma_\alpha & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \Sigma_\beta \end{pmatrix}$. Assume that optimal solution π^* is a probability measure with mean $\mathbf{0}_{d\times d}$ and covariance matrix

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

Moreover, δ is small enough that all eigenvalues of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$ are not smaller than $\frac{\delta}{4}$. Denote

$$\Sigma_{\alpha,\beta,\tau,\delta} = \Sigma_\beta^{-\frac{1}{2}}\left(\mathsf{Id} + \frac{\tau+\delta}{2}\Sigma_\alpha^{-1}\right)\Sigma_\beta^{-\frac{1}{2}}.$$

Then we have

$$\begin{aligned}
K_{x\beta} &= \Sigma_x^{\frac{1}{2}}\Sigma_\beta^{\frac{1}{2}} - \frac{\delta}{4}\mathsf{Id}, \\
\Sigma_x &= \Sigma_\beta^{-\frac{1}{2}}\left[\frac{\tau}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-1} + \frac{1}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-2}\mathbf{S}\right]\Sigma_\beta^{-\frac{1}{2}},
\end{aligned}$$

where $\mathbf{S} = \left[\mathsf{Id} + \left(\mathsf{Id} + (2\tau + 3\delta)\Sigma_{\alpha,\beta,\tau,\delta}\right)^{\frac{1}{2}}\right]$.

The proof is given in Appendix B.1. We observe that both hyperparameters, τ and δ, contribute to the structure of Σ_x. In terms of the structure of the paper, theorem 4.3 serves as a theoretical extension that completes the concept of SUOT and provides a foundation for its variants, such as those incorporating negative entropy.

Under the restriction of theorem 4.1, theorem 4.2 becomes essential, as it provides a closed-form expression for SUOT when α, β are Gaussian. Together, they form the theoretical foundation for analysing the structure of the barycenter of the α_i's and for developing the latter methods which are the main contributions. Next, we will present our algorithms: Exact Geodesic Gradient Descent and Hybrid Gradient Descent, which focus on updating the covariance matrices of the barycenter for centered Gaussians. Without loss of generality, we assume that all weights are equal.

4.2 Exact Geodesic Gradient Descent for SUOT-based Barycenter

We first show in theorem 4.4 that $W^2_{2_{\text{SUOT}}}(\alpha, \beta, \tau)$ has a closed form for its Wasserstein gradient on Bures manifold, then it is more precise to apply Riemannian Gradient Descent on SPD manifold for our barycenter problem.

Theorem 4.4. *Consider $W^2_{2_{\text{SUOT}}}(\Sigma_\alpha, \Sigma_\beta, \tau)$ where Σ_α, τ are fixed and Σ_β is seen as the variable. Then the Wasserstein gradient of $W^2_{2_{\text{SUOT}}}(\Sigma_\alpha, \Sigma_\beta, \tau)$ with respect to Σ_β on Bures manifold is formulated by*

$$2\mathsf{Id} - \Big(2\Sigma^{-1}_{\alpha,\tau} + \frac{1}{2}(U + \tau M)\Big) + \frac{3}{2}\tau\Sigma^{-1}_\beta + \frac{\tau^2}{2}(P + Q),$$

where

$$\widetilde{\Sigma}_{\beta,\alpha,\tau} = \left\{\left[\Sigma^{-\frac{1}{2}}_{\alpha,\tau}\Sigma_\beta\Sigma^{-\frac{1}{2}}_{\alpha,\tau}\right]^2 + \tau\left[\Sigma^{-\frac{1}{2}}_{\alpha,\tau}\Sigma_\beta\Sigma^{-\frac{1}{2}}_{\alpha,\tau}\right]\right\}^{\frac{1}{2}},$$

$$M = \Sigma^{-\frac{1}{2}}_{\alpha,\tau}\widetilde{\Sigma}^{-1}_{\beta,\alpha,\tau}\Sigma^{-\frac{1}{2}}_{\alpha,\tau}, \quad U = \Sigma^{-1}_{\alpha,\tau}\Sigma_\beta M + M\Sigma_\beta\Sigma^{-1}_{\alpha,\tau},$$

$$V = \left[\mathsf{Id} + \tau\Sigma^{\frac{1}{2}}_{\alpha,\tau}\Sigma^{-1}_\beta\Sigma^{\frac{1}{2}}_{\alpha,\tau}\right]^{\frac{1}{2}},$$

$$P = \Sigma^{-1}_\beta\Sigma^{\frac{1}{2}}_{\alpha,\tau}\big[\mathsf{Id} + V\big]^{-1}\Sigma^{\frac{1}{2}}_{\alpha,\tau}\Sigma^{-1}_\beta,$$

$$Q = \Sigma^{\frac{1}{2}}_{\alpha,\tau}\big[\mathsf{Id} + V\big]^{-1}\Sigma^{\frac{1}{2}}_{\alpha,\tau}\Sigma^{-2}_\beta.$$

This directly leads to the subsequent formula for a condition of barycenter of $\{\Sigma_{\alpha_i}\}^n_{i=1}$ by summing the first derivatives of the components. The proof is given in Appendix C. Consequently, thanks to the closed form of the Wasserstein gradient, we derive the details of the Exact Geodesic Bures-Wasserstein Gradient Descent used to solve the SUOT-based Barycenter problem, as presented in Algorithm **1** with the initial point $\Sigma_{\beta^{(0)}}$ and learning rate η.

Algorithm 1 Exact Geodesic Bures-Wasserstein Gradient Descent

Require: $\mathcal{P} = \{\mathcal{N}(\mathbf{0}, \Sigma_{\alpha_i})\}^n_{i=1}, \mathcal{N}(\mathbf{0}, \Sigma^{(0)}_\beta), \eta, T, \epsilon$
  for $k = 1, \dots, T$ **do**
    $\mathbf{G}^{(k)}_1 = 2\mathsf{Id} + \frac{3}{2}\tau\Sigma^{(k-1)}_\beta$
    $M^{(k)}_i = \Sigma^{-\frac{1}{2}}_{\alpha_i,\tau}[\widetilde{\Sigma}^{(k-1)}_{\beta,\alpha_i,\tau}]^{-1}\Sigma^{-\frac{1}{2}}_{\alpha_i,\tau}$
    $U^{(k)}_i = \Sigma^{-1}_{\alpha_i,\tau}\Sigma^{(k-1)}_\beta M^{(k)}_i + M^{(k)}_i\Sigma^{(k-1)}_\beta\Sigma^{-1}_{\alpha_i,\tau}$
    $\mathbf{G}^{(k)}_2 = {}^1\!/_n\sum^n_{i=1}\left[2\Sigma^{-1}_{\alpha_i,\tau} + \frac{1}{2}(U^{(k)}_i + \tau M^{(k)}_i)\right]$
    $V^{(k)}_i = \left[\mathsf{Id} + \tau\Sigma^{\frac{1}{2}}_{\alpha_i,\tau}[\Sigma^{(k-1)}_\beta]^{-1}\Sigma^{\frac{1}{2}}_{\alpha_i,\tau}\right]^{\frac{1}{2}}$
    $P^{(k)}_i = [\Sigma^{(k-1)}_\beta]^{-1}\Sigma^{\frac{1}{2}}_{\alpha_i,\tau}\big[\mathsf{Id} + V^{(k)}_i\big]^{-1}\Sigma^{\frac{1}{2}}_{\alpha_i,\tau}[\Sigma^{(k-1)}_\beta]^{-1}$
    $Q^{(k)}_i = \Sigma^{\frac{1}{2}}_{\alpha_i,\tau}\big[\mathsf{Id} + V^{(k)}_i\big]^{-1}\Sigma^{\frac{1}{2}}_{\alpha_i,\tau}[\Sigma^{(k-1)}_\beta]^{-2}$
    $\mathbf{G}^{(k)}_3 = {}^1\!/_n\sum^n_{i=1}(P^{(k)}_i + Q^{(k)}_i)$
    $\mathbf{G}^{(k)} = \eta\left(\mathbf{G}^{(k)}_1 - \mathbf{G}^{(k)}_2 + \frac{\tau}{2}\mathbf{G}^{(k)}_3\right)$
    $\Sigma^{(k)}_\beta = \mathbf{G}^{(k)}\Sigma^{(k-1)}_\beta\mathbf{G}^{(k)}$
    if $\left\|W^2_{2_{\text{SUOT}}}\left(\mathcal{P}, \Sigma^{(k-1)}_\beta\right) - W^2_{2_{\text{SUOT}}}\left(\mathcal{P}, \Sigma^{(k)}_\beta\right)\right\| \le \epsilon$ **then**
      Output: $\Sigma_\beta = \Sigma^{(k)}_\beta$ which is the solution of the barycenter problem.
    end if
  end for
  Output: $\Sigma_\beta = \Sigma_{\beta^{(T)}}$ which is the solution of the barycenter problem.

Remark 1: Our consideration focusing on centered Gaussians does not lose the generality. In fact, the update equation for the descent step decomposes into two parts: one for the mean and one for the covariance matrix. However, the updated equation for the mean is straightforwardly inferred from Theorem 4.2. Specifically, by denoting

$$M_i = \left(\left\|\Sigma^{-1}_{\alpha_i,\tau}\right\|^2_F + 1\right)\mathsf{Id} - 2\Sigma^{-1}_{\alpha_i,\tau} + \frac{\tau}{2}\Sigma^{-1}_{\alpha_i,\tau}\Sigma^{-1}_{\alpha_i}\Sigma^{-1}_{\alpha_i,\tau},$$

then taking the first derivative of $W^2_{2_{\text{SUOT}}}$ with respect to $\mathbf{b}$ is equivalent to taking the first derivative of $(\mathbf{a}_i - \mathbf{b})^T M_i(\mathbf{a}_i - \mathbf{b})$ with respect to $\mathbf{b}$, which is $-2M_i(\mathbf{a} - \mathbf{b})$. Summing over $(\alpha_i)^n_{i=1}$ yields

$$\mathbf{b} = \Big(\sum^n_{i=1} M_i\Big)^{-1}\Big(\sum^n_{i=1} M_i\mathbf{a}_i\Big).$$

Theorem 4.5. *Suppose we apply Exact Geodesic Bures-Wasserstein Gradient Descent Algorithm with starting points $\Sigma^{(0)}_\beta \in \mathcal{K}_{[1/\rho,\rho]} := \{\Sigma \in \mathbb{S}^d_{++} | \frac{1}{\rho} \le \lambda_i(\Sigma) \le \rho \quad \forall i = 1, \dots, d\}$ for fixed ρ with learning rate η and all the updated matrices lie in $\mathcal{K}_{[1/\rho,\rho]}$, then the algorithm converges to an optimal solution $\Sigma^\star_\beta$. Moreover, we have convergence guarantees at k-th iteration*

$$\mathcal{D}\left(\Sigma^{(k)}_\beta\right) \le \left(1 - \frac{8\tau^2\eta(1 - \frac{\eta}{2})}{\rho(\rho^2 + 2\tau\rho)^{\frac{3}{2}}}\right)^k \mathcal{D}\left(\Sigma^{(0)}_\beta\right),$$

where $\mathcal{D}\left(\Sigma^{(k)}_\beta\right) = L\left(\Sigma^{(k)}_\beta\right) - L\left(\Sigma^\star_\beta\right)$ is distance from objective function $L(\cdot)$ at k-th iteration to the optimal value.

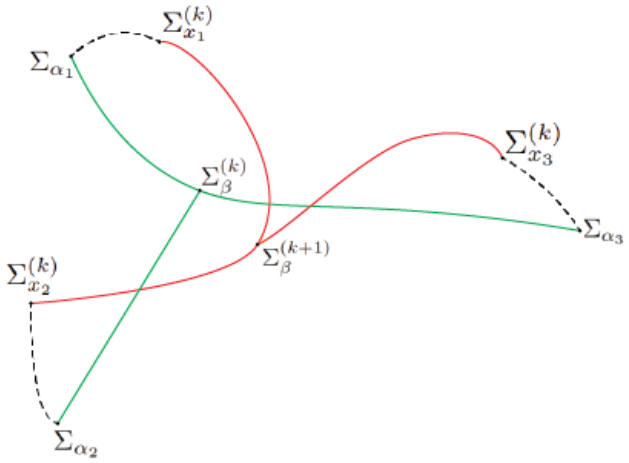

Figure 1: Overview of hybrid updated iteration

We give the full proof in Appendix D.2. The condition on the set $\mathcal{K}_{[1/\rho,\rho]}$ is natural in the context of finite distributions and is also employed in Altschuler et al. (2021). The proof hinges on showing that $W^2_{2_{\text{SUOT}}}(\alpha,\beta,\tau)$ is 1-smooth on the Bures manifold and $\frac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}$-strongly convex. Our Theorem 4.5 establishes the algorithm's geometric convergence, with the added novelty that the convergence rate is dimension-free, extending the conclusion in Altschuler et al. (2021). From the bound in the result, we see that a good choice of learning rate is $\eta \in (0,\delta]$ such that $\frac{8\tau^2\eta(1-\eta)}{\rho(\rho^2+2\tau\rho)^{3/2}} \in (0,1)$. It is worth emphasizing that a reasonable learning rate here does depend solely on the data condition number ρ and contribution of KL divergence τ, notably does not depend on the dimension d.

4.3 Hybrid Algorithm for Barycenter UOT

Although Algorithm **1** is precise and theoretically grounded, computing the exact gradient becomes cumbersome if we modify the objective for different modelling scenarios. Even small changes may require rederiving lengthy and delicate expressions. This motivated the development of Algorithm **2**, which is considerably more flexible. This idea is motivated by Chewi et al. (2020) within the context of detecting robust terms. Our algorithm has two alternate steps: one step is to find the minimizer under the regularization of the KL divergence, the other step is to apply the Riemannian Gradient Descent as in the work of Chewi et al. (2020). Figure 1 demonstrates the step in Algorithm **2**: lines 3–7 perform the extraction step, while lines 8–10 update the barycenter covariance using Chewi's method. Formally, given set of SPD matrices $\{\Sigma_{\alpha_i}\}_{i=1}^n$ and a starting point $\Sigma_\beta^{(0)}$; at k-th update $\Sigma_\beta^{(k)}$, we first find $\Sigma_{x_i}^{(k)}$ which are minimizers of $W_2^2(\Sigma_{x_i}^{(k)},\Sigma_\beta^{(k)}) + \tau\text{KL}(\Sigma_{x_i}^{(k)}\|\Sigma_{\alpha_i})$ through Theorem 4.2. Then given $\left(\Sigma_{x_i}^{(k)}\right)_{i=1}^n$, we find $\Sigma_\beta^{(k+1)}$ as Wasserstein Barycenter of them. Theoretically, the scheme of Hybrid Bures-Wassertein algorithm could be seen as a Block Coordinate Descent on SPD Manifolds Peng and Vidal (2023). In every iteration, the total objective function decreases, so we will arrive at a solution.

Algorithm 2 Hybrid Bures-Wasserstein Gradient Descent

Require: $\mathcal{P} = \left(\mathcal{N}(\mathbf{0},\Sigma_{\alpha_i})\right)_1^n, \mathcal{N}(\mathbf{0},\Sigma_\beta^{(0)}), T, \epsilon$
  for $k = 1,\dots,T$ **do**
    for $i = 1,\dots,n$ **do**
      $\Sigma_{\alpha_i,\tau,\beta}^{(k-1)} = \left(\Sigma_\beta^{(k-1)}\right)^{-\frac{1}{2}}\left(\mathsf{Id} + \frac{\tau}{2}\Sigma_{\alpha_i}^{-1}\right)\left(\Sigma_\beta^{(k-1)}\right)^{-\frac{1}{2}}$
      $\Sigma_{\gamma_i}^{(k)} = \frac{\tau}{2}\mathsf{Id} + \frac{1}{2}\left(\Sigma_{\alpha_i,\tau,\beta}^{(k-1)}\right)^{-1}\left[\mathsf{Id} + \left(\mathsf{Id} + 2\tau\Sigma_{\alpha_i,\tau,\beta}^{(k-1)}\right)^{\frac{1}{2}}\right]$
      $\Sigma_{x_i}^{(k)} = \left(\Sigma_\beta^{(k-1)}\right)^{-\frac{1}{2}}\left(\Sigma_{\alpha_i,\tau,\beta}^{(k-1)}\right)^{-1}\Sigma_{\gamma_i}^{(k)}\left(\Sigma_\beta^{(k-1)}\right)^{-\frac{1}{2}}$
    end for
    $\Sigma_{\beta,x_i}^{(k-1)} = \left[\left(\Sigma_\beta^{(k-1)}\right)^{\frac{1}{2}}\Sigma_{x_i}^{(k)}\left(\Sigma_\beta^{(k-1)}\right)^{\frac{1}{2}}\right]\forall i = 1,\dots,n$
    $\mathbf{S}^{(k)} = \frac{1}{n}\sum_{i=1}^n\left(\Sigma_\beta^{(k-1)}\right)^{-\frac{1}{2}}\left(\Sigma_{\beta,x_i}^{(k-1)}\right)^{\frac{1}{2}}\left(\Sigma_\beta^{(k-1)}\right)^{-\frac{1}{2}}$
    $\Sigma_\beta^{(k)} = \mathbf{S}^{(k)}\Sigma_\beta^{(k-1)}\mathbf{S}^{(k)}$
    if $\left\|W^2_{2_{\text{SUOT}}}\left(\mathcal{P},\Sigma_\beta^{(k-1)}\right) - W^2_{2_{\text{SUOT}}}\left(\mathcal{P},\Sigma_\beta^{(k)}\right)\right\| \le \epsilon$ **then**
      Output: $\Sigma_\beta = \Sigma_\beta^{(k)}$ which is the solution of the barycenter problem.
    end if
  end for
  Output: $\Sigma_\beta = \Sigma_{\beta(T)}$ which is the solution of the barycenter problem.

Note that our algorithms rely on closed-form computations, and even minor changes to the objective require significant derivations. theorem 4.3 makes these adaptations more accessible, allowing practitioners to replace the update of Σ_x in Algorithm **2** with the expression in theorem 4.3 to obtain an algorithm for Entropic SUOT.

We provide theorem 4.6 as a guarantee for finding the optimal solution of Algorithm **2**.

Theorem 4.6. *Suppose we apply Hybrid Bures-Wasserstein Gradient Descent Algorithm with starting points* $\Sigma_\beta^{(0)} \in \mathcal{K}_{[1/\rho,\rho]} := \{\Sigma \in \mathbb{S}^d_{++} | \frac{1}{\rho} \le \lambda_i(\Sigma) \le \rho \quad \forall i = 1,\dots,d\}$ *for fixed* ρ *and all the updated matrices lie in* $\mathcal{K}_{[1/\rho,\rho]}$*, then the algorithm converges to an optimal solution.*

The full proof is given in Appendix D.3, employing a flow similar to Block Coordinate Descent. This result is meaningful when considering a scenario where we have contaminated Gaussian data, denoted as Σ_αs, and assume the true underlying data follows a Gaussian distribution, Σ_xs. Intuitively, it is preferable to recover the true data before performing calculations on it.

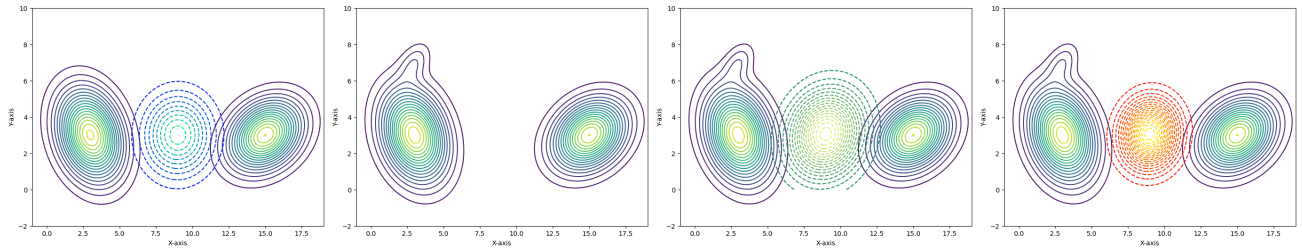

Figure 2: 2D Contour Plot Gaussian Mixture Distribution. From left to right: two Gaussians with their barycenter (blue); noise is added to one Gaussian on the left, creating a mixture of Gaussians; normal Wasserstein Gaussian Barycenter (green); SUOT-based Gaussian Barycenter (red).

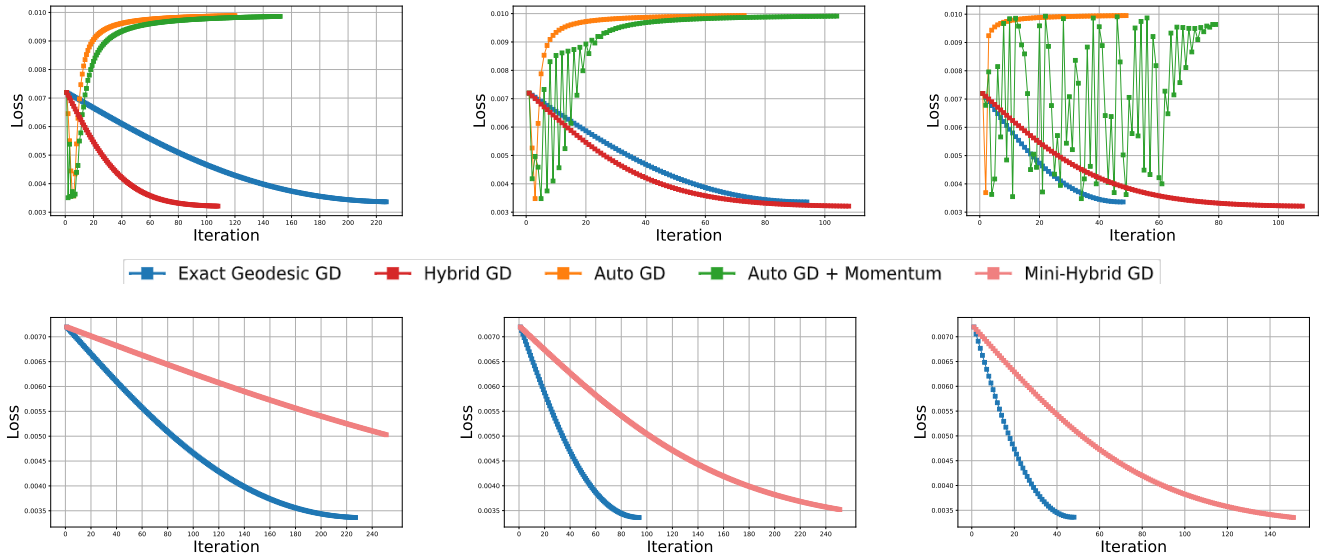

Figure 3: Loss on $L(\Sigma_\beta)$ through iterations with different step sizes (0.1; 0.25 and 0.5). (Top): red: Hybrid Gradient Descent, blue: Exact Geodesic Gradient Descent, orange: Auto-Geodesic Gradient Descent, green: Auto-Geodesic Gradient Descent with Momentum. (Bottom): pink: Mini-Hybrid Gradient Descent, blue: Exact Geodesic Gradient Descent.

5 NUMERICAL EXPERIMENTS

In this section, we provide numerical evidence regarding our presented SUOT-based Barycenter with normal Wasserstein Barycenter, as well as some ablation studies on optimization methods. To generate a sample of the SPD matrix, we suggest the strategy in Chewi et al. (2020). Let A_i be a matrix where entries are i.i.d. samples from $\mathcal{N}(0, \sigma^2)$. Our random sample on the Bures manifold is then given by taking a symmetric version of A_i as $\frac{A_i + A_i^\top}{2}$, then applying matrix exponential function $\mathbf{expm}(A) : \mathbb{S}^d \to \mathbb{S}^d_{++}$ to this symmetric version. Particularly, if $P\,\mathbf{diag}(\lambda_i)P^\top$ is the SVD decomposition of A, then

$$\mathbf{expm}(A) = P\,\mathbf{diag}(e^{\lambda_i})P^\top.$$

Figure 2a compares the standard Wasserstein Barycenter with our SUOT-based Barycenter in the presence of noise. The experiment uses two 2D Gaussian distributions. The first subfigure shows the clean Gaussian distributions and their barycenter. In the second subfigure, noise is introduced to the left Gaussian by computing a weighted sum of its PDF with an outlier Gaussian, resulting in a mixture of Gaussians. We approximate this mixture of Gaussians by a Gaussian with an estimated mean and covariance matrix for the experiments. We plot the barycenter of the contaminated measure on the left and the clean measure on the right using both the SUOT-based method and the standard Wasserstein method from Chewi et al. (2020). The third subfigure depicts the standard Wasserstein Barycenter, which diverges from the true barycenter shown earlier. In

contrast, the fourth subfigure presents the SUOT-based Barycenter, which closely aligns with the true barycenter despite the noise. The results demonstrate that the SUOT-based method is less affected by noise, yielding a barycenter that better represents the original one. Quantitatively, the Wasserstein distances from the true barycenter are 0.2673 for the standard method and 0.06 for the SUOT-based approach.

For the optimization study, we implemented five Riemannian Gradient Descent methods and examined the behavior of the loss objective over iterations. These methods include: *Exact Geodesic Gradient Descent (Algorithm 1)*; *Hybrid Gradient Descent (Algorithm 2)*; *Mini-Hybrid Gradient Descent*, which modifies the Hybrid Gradient approach by updating Σ_β with a one-step update as described in Chewi et al. (2020), rather than performing a complete barycenter calculation; *Auto-Geodesic Gradient Descent*, in which the Euclidean gradient $\nabla F(\Sigma)$ is automatically derived using the PyTorch library, while the Riemannian gradient is approximated by $2\left(\nabla L(\Sigma)\Sigma + (\nabla L(\Sigma)\Sigma)^T\right)$ Han et al. (2021); and *Auto-Geodesic Gradient Descent with Momentum* which incorporates momentum. We generated a dataset comprising 50 centered Gaussians in $\mathbb{R}^5$, using diagonal SPD matrices for the covariance matrices to ensure efficiency and tractability. Figure 3a illustrates the convergence of Gradient Descent on the Bures manifold. The step sizes selected were 0.1, 0.25, and 0.5 for the various methods; however, for the Hybrid method, the step size was set to 1, as each iteration involves finding a barycenter through the complete process outlined in Chewi et al. (2020), indicating that step sizes do not influence convergence. Each run began with a fixed $\Sigma_{\beta^{(0)}} \in \mathbb{S}^d_{++}$ and employed the same stopping condition. All the optimal solutions in this part are computed using the **Pytorch** library. All the experiments are conducted on a server with 8 GPU Tesla v100-sxm2-32GB RAM.

We divide the figures into two sequences for the following reasons: At the top, we aim to demonstrate the consistent convergence rates of our Algorithms 1 and 2 in practice, as well as highlight the importance of computing exact Riemannian gradient in Theorem 4.4. At the bottom, we compare the reduction in the loss function achieved by a single downhill step using our gradient in Algorithm 2 against the gradient from Chewi et al. (2020). From **Figure 3**, we make the following observations: **(1)** For different step sizes, the objective values of our methods (Exact, Hybrid, and Mini-Hybrid Geodesic Gradient Descent) consistently identify a descent direction and converge quickly to a solution. This behavior is not observed with the standard Riemannian gradient descent using auto-approximated Wasserstein gradients, as they initially approach a near minimizer but subsequently diverge. Note that these methods do not arise from our algorithms; rather, they reflect what occurs in the absence of theorem 4.2 and theorem 4.4, which guarantee a correct Riemannian gradient. **(2)** The Exact method, when the step size is increased to 0.5, converges more rapidly in terms of iterations compared to the Hybrid method, and the overall runtime is significantly faster (5.10 seconds per iteration on average compared to 170.03 seconds per iteration on average). **(3)** Across various step sizes, the downhill step taken by our Exact method results in a more substantial decrease in the loss function compared to the vanilla approach.

For practical use, we recommend Algorithm **1** when computational efficiency is the priority, as it offers faster runtime. Algorithm **2**, by contrast, provides a more flexible framework for modifying the objective for different modelling purposes. And to avoid the sensitivity to the chosen step size with Descent methods, we suggest using an adaptive step-size strategy; which means that we start with a large value of step size to accelerate convergence in the early stages, then whenever the loss function tends to increase, the step size is gradually reduced by multiplying it with a factor less than 1 (e.g. 0.8). This adjustment helps maintain stability while ensuring efficient convergence. Additionally, we present the performance of the Stochastic Gradient Descent version of these methods and include an ablation study examining the relationship between the SUOT-based Barycenter and the parameter τ in Appendix E.

6 CONCLUSION

This paper introduces a new approach to barycenter computation using Semi-Unbalanced Optimal Transport (SUOT) for contaminated Gaussian distributions. Building on the closed-form expression for the SUOT distance, we propose two barycenter computation methods on the SPD manifold, which are shown to converge consistently through theoretical analysis and experiments. Our work contributes a novel closed-form distance and its derivative on the SPD manifold, enabling manifold-based tools for broader applications beyond barycenter computation. However, our experiments do have certain limitations related to synthetic datasets, which we plan to address in future work. There are several potential directions for applying the SUOT-based barycenter framework, as in generative modelling Gazdieva et al. (2024), where it enables conditional sampling from the barycenter, allowing for the generation of images that reflect specific attributes or features based on input conditions, and handle imperfect data and class imbalance at a large scale.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press.

Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Altschuler, J., Chewi, S., Gerber, P. R., and Stromme, A. (2021). Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34:22132–22145.

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2016). A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138.

Bhatia, R. (2009). Positive definite matrices. In *Positive Definite Matrices*. Princeton university press.

Blondel, M., Seguy, V., and Rolet, A. (2018). Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pages 880–889. PMLR.

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

Bunne, C., Hsieh, Y.-P., Cuturi, M., and Krause, A. (2023). The schrödinger bridge between gaussian measures has a closed form. In *International Conference on Artificial Intelligence and Statistics*, pages 5802–5833. PMLR.

Chapel, L., Flamary, R., Wu, H., Févotte, C., and Gasso, G. (2021). Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282.

Cheng, K., Aeron, S., Hughes, M. C., and Miller, E. L. (2021). Dynamical Wasserstein barycenters for time-series modeling. *Advances in Neural Information Processing Systems*, 34:27991–28003.

Chewi, S., Maunu, T., Rigollet, P., and Stromme, A. J. (2020). Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR.

Chi, J., Yang, Z., Li, X., Ouyang, J., and Guan, R. (2023). Variational wasserstein barycenters with c-cyclical monotonicity regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7157–7165.

Chizat, L. (2023). Doubly regularized entropic Wasserstein barycenters. *arXiv preprint arXiv:2303.11844*.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2016). Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*.

Cuesta-Albertos, J. A., Matrán-Bea, C., and Tuero-Diaz, A. (1996). On lower bounds for the l 2-wasserstein metric in a hilbert space. *Journal of Theoretical Probability*, 9(2):263–283.

Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882.

Gao, Z., Wu, Y., Jia, Y., and Harandi, M. (2020). Learning to optimize on SPD manifolds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7700–7709.

Gazdieva, M., Choi, J., Kolesov, A., Choi, J., Mokrov, P., and Korotin, A. (2024). Robust barycenter estimation using semi-unbalanced neural optimal transport. *arXiv preprint arXiv:2410.03974*.

Gutman, D. H. and Ho-Nguyen, N. (2023). Coordinate descent without coordinates: Tangent subspace descent on riemannian manifolds. *Mathematics of Operations Research*, 48(1):127–159.

Han, A., Mishra, B., Jawanpuria, P. K., and Gao, J. (2021). On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. *Advances in Neural Information Processing Systems*, 34:8940–8953.

Herath, S., Harandi, M., and Porikli, F. (2017). Learning an invariant hilbert space for domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3845–3854.

Janati, H., Muzellec, B., Peyré, G., and Cuturi, M. (2020). Entropic optimal transport between unbalanced Gaussian measures has a closed form. *Advances in neural information processing systems*, 33:10468–10479.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Kobler, R., Hirayama, J.-i., Zhao, Q., and Kawanabe, M. (2022). Spd domain-specific batch normalization to crack interpretable unsupervised domain adaptation in eeg. *Advances in Neural Information Processing Systems*, 35:6219–6235.

Korotin, A., Li, L., Solomon, J., and Burnaev, E. (2021). Continuous wasserstein-2 barycenter estimation without minimax optimization. *arXiv preprint arXiv:2102.01752*.

Korotin, A., Selikhanovych, D., and Burnaev, E. (2022). Neural optimal transport. *arXiv preprint arXiv:2201.12220*.

Le, K., Le, D. Q., Nguyen, H., Do, D., Pham, T., and Ho, N. (2022). Entropic Gromov-Wasserstein between Gaussian distributions. In *International Conference on Machine Learning*, pages 12164–12203. PMLR.

Liero, M., Mielke, A., and Savaré, G. (2018). Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117.

Mallasto, A., Gerolin, A., and Minh, H. Q. (2022). Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Information Geometry*, 5(1):289–323.

Mignon, S., Galerne, B., Hidane, M., Louchet, C., and Mille, J. (2023). Semi-unbalanced regularized optimal transport for image restoration. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 466–470. IEEE.

Mokrov, P., Korotin, A., Kolesov, A., Gushchin, N., and Burnaev, E. (2023). Energy-guided entropic neural optimal transport. *arXiv preprint arXiv:2304.06094*.

Montesuma, E. F. and Mboula, F. M. N. (2021). Wasserstein barycenter for multi-source domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16785–16793.

Naas, J., Nies, G., Li, H., Stoldt, S., Schmitzer, B., Jakobs, S., and Munk, A. (2024). Multi-match: geometry-informed colocalization in multi-color super-resolution microscopy. *Communications Biology*, 7(1):1139.

Nguyen, D., Diep, N., Nguyen, T., Le, H., Nguyen, T., Nguyen, T., Nguyen, T., Ho, N., Xie, P., Wattenhofer, R., Zhou, J., Sonntag, D., and Niepert, M. (2024). LoGra-Med: Long context multi-graph alignment for medical vision-language model. *arXiv preprint arXiv:2410.02615*.

Nguyen, H., Le, K., Nguyen, Q., Pham, T., Bui, H., and Ho, N. (2021). On robust optimal transport: Computational complexity and barycenter computation. In *Advances in NeurIPS*.

Noble, M., De Bortoli, V., Doucet, A., and Durmus, A. (2024). Tree-based diffusion schrödinger bridge with applications to wasserstein barycenters. *Advances in Neural Information Processing Systems*, 36.

Otto, F. and Villani, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400.

Peng, L. and Vidal, R. (2023). Block coordinate descent on smooth manifolds: Convergence theory and twenty-one examples. *arXiv preprint arXiv:2305.14744*.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. (2020). On unbalanced optimal transport: An analysis of Sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR.

Séjourné, T., Peyré, G., and Vialard, F.-X. (2023). Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471.

Simon, D. and Aberdam, A. (2020). Barycenters of natural images constrained Wasserstein barycenters for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7910–7919.

Simou, E. and Frossard, P. (2019). Graph signal representation with Wasserstein barycenters. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5386–5390. IEEE.

Simou, E., Thanou, D., and Frossard, P. (2020). node2coords: Graph representation learning with Wasserstein barycenters. *IEEE Transactions on Signal and Information Processing over Networks*, 7:17–29.

Takatsu, A. (2011). Wasserstein geometry of gaussian measures.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. (2021). Semi-relaxed gromov-wasserstein divergence with applications on graphs. *arXiv preprint arXiv:2110.02753*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] The paper includes clear descriptions for mathematical setting, assumptions, algorithm, and models.

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] The paper include convergence rate analysis for the main algorithms.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The source code of the paper is anonymized.

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes] The statements of the theoretical results include the necessary assumptions.

(b) Complete proofs of all theoretical results. [Yes] All the important proofs in the paper are accomplished

(c) Clear explanations of any assumptions. [Yes] We give the explanations for the assumptions in the paper.

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The code for reproducing is included.

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] We give the detail for training process in Section 5 "Numerical Experiments".

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We provide that we run experiments through random seed for multiple times.

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] We provide the information about the GPUs in the Section 5 "Numerical Experiments".

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include: [No] Our paper does not use existing assets or release new assets.

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include: [No] Our paper does not involve crowdsourcing nor research with human subjects.

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

On Barycenter Computation: Analyzing Semi-Unbalanced Optimal Transport-based Method on Bures-Wasserstein manifold Supplementary Materials

Contents

1 INTRODUCTION **1**

2 BACKGROUND **2**

3 SEMI-UNBALANCED OPTIMAL TRANSPORT-BASED BARYCENTER **3**

4 ANALYSIS ON BURES-WASSERSTEIN MANIFOLD **4**

5 NUMERICAL EXPERIMENTS **8**

6 CONCLUSION **9**

A Proof for Theorem 4.1 **13**

B Proofs for Theorem 4.2 and Theorem 4.3 **15**

C Proof for Theorem 4.4 **23**

D Proof for Theorem 4.5 and 4.6 **28**

E Additional Experiments **34**

Overall, Appendix A gives the proof for Theorem 4.1 about the Gaussian form of the SUOT-based Barycenter. Appendix B contains the proofs for key theoretical results about closed form in Theorem 4.2 and Theorem 4.3, alongside supporting propositions and lemmas. The proof for the closed-form of the Wasserstein gradient (Theorem 4.4) is given in Appendix C. Appendix D presents convergence guarantees under specific conditions, as Appendix D.2 provides a detailed derivation of convergence for the Exact Geodesic Gradient Descent algorithm and Appendix D.3 discusses the Hybrid Bures-Wasserstein, which are central to the numerical methods proposed in the paper on the Bures-Wasserstein manifold. Appendix E includes the ablation study on the impact of the parameter τ on the SUOT-based Barycenter, as well as comparisons with standard Wasserstein Barycenter.

A Proof for Theorem 4.1

First, we have the following Proposition.

Proposition A.1. *Let Σ_1 and Σ_2 be two positive definite matrices in $\mathbb{R}^d$. Consider the problem*

$$\inf_{\mu\in\mathcal{P}(\Sigma_1),\nu\in\mathcal{P}(\Sigma_2)} W_2^2(\mu,\nu),$$

where $\mathcal{P}(\Sigma_1), \mathcal{P}(\Sigma_2)$ be the set of zero-mean probability distribution in $\mathbb{R}^d$ having covariance matrix Σ_1, Σ_2, respectively. Then, the result of this optimization of this problem is $\mathsf{tr}\left(\Sigma_1 + \Sigma_2 - (\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right)$, and the problem admits optimal solution.

Proof. The Wasserstein distance can be written as

$$W_2^2(\mu,\nu) = \inf_{(X,Y)\in\Pi(\mu,\nu)} \mathbb{E}[\|X\|^2] + \mathbb{E}[\|Y\|^2] - 2\mathbb{E}[X^\top Y].$$

We have $\mathbb{E}\left(\|X\|^2\right) = \mathsf{tr}\left(\Sigma_1\right)$ and $\mathbb{E}\left(\|Y\|^2\right) = \mathsf{tr}\left(\Sigma_2\right)$. To find the supermum for $\mathbb{E}[X^\top Y]$, let $U\Sigma V^\top$ be the SVD decomposition of $\Sigma_1^{\frac{1}{2}}\Sigma_2^{\frac{1}{2}}$. Let $\bar{X} = U^\top\Sigma_1^{-1/2}X, \bar{Y} = V^\top\Sigma_2^{-1/2}Y$, then $X = \Sigma_1^{1/2}U\bar{X}, Y = \Sigma_2^{1/2}V\bar{Y}$. The covariance matrices of $\tilde{X}$ is

$$\begin{aligned}\mathbb{E}[\tilde{X}\tilde{X}^\top] &= \mathbb{E}[U^\top\Sigma_1^{-1/2}XX^\top\Sigma_1^{-1/2}U] = \mathbb{E}[U^\top\Sigma_1^{-1/2}\Sigma_1^\top\Sigma_1^{-1/2}U] \\ &= \mathbb{E}[U^\top U] = \mathsf{Id},\end{aligned}$$

and similarly, $\mathbb{E}[\tilde{Y}\tilde{Y}^\top] = \mathsf{Id}$. On the other hand,

$$\mathbb{E}[X^\top Y] = \mathbb{E}[\tilde{X}^\top U^\top\Sigma_1^{1/2}\Sigma_2^{1/2}V\tilde{Y}] = \mathbb{E}[\tilde{X}^\top\Sigma\tilde{Y}] = \sum_{i=1}^d \lambda_i(\Sigma)\mathbb{E}[\tilde{X}_i\tilde{Y}_i].$$

The third equality arises from the fact that we can express $\tilde{X}^\top\Sigma\tilde{Y} = \sum_{i=1}^d \tilde{X}^\top\tilde{\Sigma}_i\tilde{Y}$, where $\tilde{\Sigma}_i$ is a matrix filled with zeros except for the ii-th element, which is λ_i. By Cauchy-Schwarz inequality, we have

$$\sum_{i=1}^d \lambda_i(\Sigma)\mathbb{E}[\tilde{X}_i\tilde{Y}_i] \le \sum_{i=1}^d \lambda_i(\Sigma)\left(\mathbb{E}[\tilde{X}_i^2]\mathbb{E}[\tilde{Y}_i^2]\right)^{1/2} = \mathsf{tr}(\Sigma) = \mathsf{tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right).$$

Since $\mathbb{E}[\tilde{X}\tilde{X}^\top] = (E[X_iX_j])_{i\times j}$ and $\mathbb{E}[\tilde{X}\tilde{X}^\top] = \mathsf{Id}$, this implies $\mathbb{E}[\tilde{X}_i^2] = 1$ for all i (similarly, $\mathbb{E}[\tilde{Y}_i^2] = 1$), leading to the second equality. The final equality happens due to $\mathsf{tr}\left((\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2}\right) = \mathsf{tr}\left(\left[\Sigma_1^{1/2}\Sigma_2^{1/2}(\Sigma_1^{1/2}\Sigma_2^{1/2})^\top\right]^{1/2}\right)$ which equals to the sum of singular values of $\Sigma_1^{1/2}\Sigma_2^{1/2}$. The equality for Cauchy-Schwartz occurs if and only if $\tilde{X}_i = \tilde{Y}_i$ (a.s), that mean $\tilde{X} = \tilde{Y}$, or $U^\top\Sigma_1^{-1/2}X = V^\top\Sigma_2^{-1/2}Y$. Particularly, in the case of Gaussian, i.e. μ follows Gaussian distribution, the value of ν that minimizes the optimization problem is also a Gaussian distribution. □

This finding is consistent with the results presented in Cuesta-Albertos et al. (1996). Now the proof for Theorem 4.1 is given as follows

Proof. Suppose that $\tilde{\beta}$ be an optimal solution of Equation (2), then from the form of SUOT distance, there exists probability measures $(\tilde{x}_i)_{i=1}^n$ such that

$$\tilde{\beta}, (\tilde{x}_i)_{i=1}^n = \arg\min \frac{1}{n}\sum_{i=1}^n W_2^2(\beta, x_i) + \tau\mathrm{KL}(x_i\|\alpha_i).$$

Let $\beta^*, (x_i^*)_{i=1}^n$ be the Gaussian probability measure having the same mean and variance as $\tilde{\beta}, (\tilde{x}_i)_{i=1}^n$, respectively. Then, using Lemma 3.3 Le et al. (2022) about KL divergence minimum and Proposition A.1, we achieve

$$W_2^2(\beta^*, x_i^*) \le W_2^2(\tilde{\beta}, \tilde{x}_i), \quad \text{and} \quad \mathrm{KL}(x_i^*\|\alpha_i) \le \mathrm{KL}(\tilde{x}_i\|\alpha_i).$$

The equalities hold if and only if $\tilde{x}_i$ is Gaussian measure, which implies that $\tilde{\beta}$ is also a Gaussian measure. Thus, we can conclude that the optimal solution of Equation (2) is Gaussian measure. □

B Proofs for Theorem 4.2 and Theorem 4.3

B.1 Proof for Theorem 4.2

Before giving full proof for Theorem 4.2, we state some necessary lemmas:

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

$$\mathrm{KL}(\alpha \| \beta) = m_\alpha \mathrm{KL}(\bar{\alpha} \| \bar{\beta}) + \mathrm{KL}\left(m_\alpha \| m_\beta\right).$$

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

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

Moreover, the expression for $\mathrm{KL}(\bar{\alpha} \| \bar{\beta})$ is given by

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

□

Lemma B.2. *For positive constants a, τ and Υ, the function*

$$f(x) = \Upsilon x + \tau \mathrm{KL}(x \| a)$$

attains its minimum at $x^* = a \exp\left\{\frac{-\Upsilon}{\tau}\right\}$.

Proof. By taking the derivative of $f(x)$, we have

$$f'(x) = \Upsilon + \tau\{\log(x) - \log(a)\}.$$

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

$$x^* = \exp\left\{\frac{\tau \log(a) - \Upsilon}{\tau}\right\} = a \exp\left\{\frac{-\Upsilon}{\tau}\right\}.$$

□

Now we are ready to give the proof for Theorem 4.2.

Proof of Theorem 4.2. Recall that π is a positive measure such that $\pi = m_\pi \overline{\pi}$ with $\overline{\pi}$ is a probability measure with mean $(\mathbf{a}_x, \mathbf{b})$ and covariance matrix

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

Here the two marginals of $\overline{\pi}$ are denoted by $\overline{\pi}_x$ and $\overline{\pi}_y$ where $\overline{\pi}_x$ has mean $\mathbf{a}_x$ and covariance matrix Σ_x while $\overline{\pi}_y$ has mean $\mathbf{b}$ and covariance matrix $\Sigma_y = \Sigma_\beta$. Let π_x and π_y be two marginals of π, then $\pi_x = m_\pi \overline{\pi}_x$ and $\pi_y = m_\pi \overline{\pi}_y$. Note that by our Proposition A.1 and Lemma 3.3 in Le et al. (2022), π needs to be Gaussian. We also have

$$\mathbb{E}_\pi \|X - Y\|_2^2 = m_\pi \left\{\|\mathbf{a}_x - \mathbf{b}\|_2^2 + \mathsf{tr}\left(\Sigma_x\right) + \mathsf{tr}\left(\Sigma_\beta\right) - 2\mathsf{tr}\left(K_{x\beta}\right)\right\}.$$

According to Lemma B.1,

$$\mathrm{KL}\left(\pi_x \| \alpha\right) = m_\pi \mathrm{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) + \mathrm{KL}\left(m_\pi \| m_\alpha\right).$$

Combining the above results, the objective function between α and β reads as

$$W^2_{2_{\mathrm{SUOT}}}(\alpha, \beta; \tau) := \min_{\pi \in \mathcal{M}^+(\mathbb{R}^d \times \mathbb{R}^d)} m_\pi \left\{ \|\mathbf{a}_x - \mathbf{b}\|_2^2 + \mathsf{tr}\left(\Sigma_x\right) + \mathsf{tr}\left(\Sigma_\beta\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \tau \mathrm{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) \right\} + \tau \mathrm{KL}\left(m_\pi \| m_\alpha\right).$$

Denote

$$\Upsilon = \|\mathbf{a}_x - \mathbf{b}\|_2^2 + \mathsf{tr}\left(\Sigma_x\right) + \mathsf{tr}\left(\Sigma_\beta\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \tau \mathrm{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right).$$

Due to the independence of Υ with m_π, we could minimize Υ then find m_π.
Minimization of Υ: For the KL term, due to Lemma B.1 we have

$$\tau \mathrm{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) = \frac{\tau}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) - d + \log\left(\frac{\det\left(\Sigma_\alpha\right)}{\det\left(\Sigma_x\right)}\right)\right] + \frac{\tau}{2}\left(\mathbf{a}_x - \mathbf{a}\right)^\top \Sigma_\alpha^{-1}\left(\mathbf{a}_x - \mathbf{a}\right).$$

Now Υ reads

$$\begin{aligned}\Upsilon = \mathsf{tr}\left(\Sigma_x\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \frac{\tau}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) + \log\left(\frac{\det\left(\Sigma_\alpha\right)}{\det\left(\Sigma_x\right)}\right)\right] + \|\mathbf{a}_x - \mathbf{b}\|^2 + \\ \frac{\tau}{2}\left(\mathbf{a}_x - \mathbf{a}\right)^\top \Sigma_\alpha^{-1}\left(\mathbf{a}_x - \mathbf{a}\right) + \mathsf{tr}\left(\Sigma_\beta\right) - \frac{\tau d}{2}.\end{aligned}$$

Means part: We first work with the terms involving $\mathbf{a}_x$. Let $\mathbf{a}_x - \mathbf{a} = \widetilde{\mathbf{a}}_x$, then

$$\begin{aligned}\|\mathbf{a}_x - \mathbf{b}\|_2^2 &= \|\widetilde{\mathbf{a}}_x + \mathbf{a} - \mathbf{b}\|_2^2 \\ &= \|\widetilde{\mathbf{a}}_x\|^2 + 2\widetilde{\mathbf{a}}_x^\top(\mathbf{a} - \mathbf{b}) + \|\mathbf{a} - \mathbf{b}\|_2^2.\end{aligned}$$

Hence, sum of all terms which include $\widetilde{\mathbf{a}}_x$ is equal to

$$\Upsilon_{\mathbf{a},\mathbf{b}} = \|\widetilde{\mathbf{a}}_x\|_2^2 + 2\widetilde{\mathbf{a}}_x^\top(\mathbf{a} - \mathbf{b}) + \|\mathbf{a} - \mathbf{b}\|_2^2 + \frac{\tau}{2}\widetilde{\mathbf{a}}_x^\top \Sigma_\alpha^{-1}\widetilde{\mathbf{a}}_x.$$

Taking derivative according to $\widetilde{\mathbf{a}}_x$ and set equation to $\mathbf{0}$, we have

$$2\widetilde{\mathbf{a}}_x + 2(\mathbf{a} - \mathbf{b}) + \tau\Sigma_\alpha^{-1}\widetilde{\mathbf{a}}_x = 0,$$

which turns into

$$\left(\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\right)\widetilde{\mathbf{a}}_x = \mathbf{b} - \mathbf{a}.$$

Denote $\Sigma_{\alpha,\tau} = \mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}$, we obtain

$$\widetilde{\mathbf{a}}_x = \Sigma_{\alpha,\tau}^{-1}(\mathbf{b} - \mathbf{a}),$$

then $\mathbf{a}_x = \Sigma_{\alpha,\tau}^{-1}(\mathbf{b} - \mathbf{a}) + \mathbf{a}$. Pull it back to $\Upsilon_{\mathbf{a},\mathbf{b}}$ gives

$$\Upsilon_{\mathbf{a}^*,\mathbf{b}^*} = (\mathbf{a} - \mathbf{b})^\top \left\{ \left(\left\|\Sigma_{\alpha,\tau}^{-1}\right\|_F^2 + 1 \right) \mathsf{Id} - 2\Sigma_{\alpha,\tau}^{-1} + \frac{\tau}{2}\Sigma_{\alpha,\tau}^{-1}\Sigma_\alpha^{-1}\Sigma_{\alpha,\tau}^{-1} \right\}(\mathbf{a} - \mathbf{b}).$$

Covariance matrix part: The second part is to group all factors of Σ_x and $K_{x\beta}$, which is

$$\begin{aligned}\Upsilon_\Sigma &= \mathsf{tr}\left(\Sigma_x\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \frac{\tau}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) - \log\left(\det\left(\Sigma_x\right)\right)\right] \\ &= \mathsf{tr}\left(\Sigma_x\left(\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\right)\right) - 2\mathsf{tr}(K_{x\beta}) - \frac{\tau}{2}\log\left(\det\left(\Sigma_x\right)\right) \\ &= \mathsf{tr}\left(\Sigma_x\Sigma_{\alpha,\tau}\right) - 2\mathsf{tr}(K_{x\beta}) - \frac{\tau}{2}\log\left(\det\left(\Sigma_x\right)\right).\end{aligned}$$

First, we embark on the task of maximizing $\mathsf{tr}(K_{x\beta})$ while fixing Σ_x. Note that matrices $\Sigma_x, K_{x\beta}$ must satisfy condition that covariance block matrix $\begin{pmatrix} \Sigma_x & K_{x\beta} \\ K_{x\beta}^\top & \Sigma_\beta \end{pmatrix}$ is SPD. This can be equivalently expressed as

$$\begin{aligned} & \Sigma_x - K_{x\beta}\Sigma_\beta^{-1}K_{x\beta}^\top \succeq 0 \\ \Leftrightarrow \quad & \mathsf{Id} - \left(\Sigma_x^{-\frac{1}{2}} K_{x\beta}\Sigma_\beta^{-\frac{1}{2}}\right)\left(\Sigma_\beta^{-\frac{1}{2}} K_{x\beta}^T \Sigma_x^{-\frac{1}{2}}\right) \succeq 0. \end{aligned}$$

By denoting $\Sigma_x^{-\frac{1}{2}} K_{x\beta}\Sigma_\beta^{-\frac{1}{2}}$ as H, this condition can be reformulated as

$$\mathsf{Id} - HH^\top \succeq 0.$$

Noting that $K_{x\beta} = \Sigma_x^{\frac{1}{2}} H \Sigma_\beta^{\frac{1}{2}}$, it follows that $\mathsf{tr}\left(K_{x\beta}\right) = \mathsf{tr}\left(H\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)$. Let $\left(\lambda_i(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}})\right)_{i=1}^d$ and $(\lambda_i(H))_{i=1}^d$ represent the singular values of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$ and H in descending order, respectively. Applying von Neuman inequality, we get

$$\mathsf{tr}(K_{x\beta}) \leq \sum_{i=1}^d \lambda_i\big(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\big)\lambda_i(H) \leq \sum_{i=1}^d \lambda_i\big(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\big).$$

The second inequality is a consequence of the Lemma 3.4 in Le et al. (2022), which states that all singular values of $\Sigma_x^{-\frac{1}{2}} K_{x\beta}\Sigma_\beta^{-\frac{1}{2}}$ lie between 0 and 1. Additionally, if we consider the singular value decomposition $U\Lambda V$ of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$, we can choose $H = V^\top U^\top$, satisfying the condition for H mentioned earlier and

$$\mathsf{tr}\left(H\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) = \mathsf{tr}(\Lambda) = \sum_i^d \lambda_i\big(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\big).$$

It confirms that when Υ_Σ achieves its maximum value, $\mathsf{tr}(K_{x\beta})$ corresponds to the sum of all singular values of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$. In such a scenario, we obtain

$$\Upsilon_\Sigma = \mathsf{tr}\left(\Sigma_x\Sigma_{\alpha,\tau}\right) - 2\mathsf{tr}\Big(\big[\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\big]^{\frac{1}{2}}\Big) - \frac{\tau}{2}\log\left(\det\left(\Sigma_x\right)\right).$$

Let us define $\widetilde{\Sigma} = \left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\right)^{\frac{1}{2}}$, which can be expressed as $\Sigma_x = \Sigma_\beta^{-\frac{1}{2}}\widetilde{\Sigma}^2\Sigma_\beta^{-\frac{1}{2}}$. Consequently, we obtain

$$\mathsf{tr}(\Sigma_x\Sigma_{\alpha,\tau}) = \mathsf{tr}\left(\widetilde{\Sigma}^2\Sigma_\beta^{-\frac{1}{2}}\Sigma_{\alpha,\tau}\Sigma_\beta^{-\frac{1}{2}}\right) = \mathsf{tr}\big(\widetilde{\Sigma}^2\Sigma_{\alpha,\tau,\beta}\big).$$

Additionally, we have

$$\det(\Sigma_x) = \frac{\det(\widetilde{\Sigma})^2}{\det(\Sigma_\beta)}.$$

Now Υ_Σ turns into

$$\Upsilon_\Sigma = \mathsf{tr}\left(\widetilde{\Sigma}^2\Sigma_{\alpha,\tau,\beta}\right) - 2\mathsf{tr}(\widetilde{\Sigma}) - \frac{\tau}{2}\log\big(\frac{\det(\widetilde{\Sigma})^2}{\det(\Sigma_\beta)}\big).$$

Note that $\widetilde{\Sigma}$ is symmetric, so there exists a diagonal matrix Λ that $\widetilde{\Sigma}$ is similar to Λ (which implies $\widetilde{\Sigma}^2$ is similar to Λ^2. Assume that $\Lambda = \mathrm{diag}(\lambda_i(\widetilde{\Sigma}))_{i=1}^d$ which is decreasingly ordered and $(\lambda_i(\Sigma_{\alpha,\tau,\beta}))_{i=1}^d$ are eigenvalues of $\Sigma_{\alpha,\beta,\tau}$ in ascending order, by Ruhe's trace inequality

$$\mathsf{tr}\left(\widetilde{\Sigma}^2\Sigma_{\alpha,\tau,\beta}\right) \geq \sum_{i=1}^d \lambda_i(\widetilde{\Sigma})^2\lambda_{d-i+1}(\Sigma_{\alpha,\tau,\beta}),$$

where the equality holds when $\widetilde{\Sigma}^2$ and $\Sigma_{\alpha,\beta,\tau}$ are commuting. The optimization part now is calculated on eigenvalues of Λ, because

$$\begin{aligned}\Upsilon_\Sigma &\geq \sum_{i=1}^{d} \lambda_i^2(\widetilde{\Sigma})\lambda_{n-i+1}(\Sigma_{\alpha,\tau,\beta}) - 2\sum_{r=1}^{d}\lambda_i(\widetilde{\Sigma}) - \tau\log\left(\prod_{i=1}^{d}\lambda_i(\widetilde{\Sigma})\right)\\ &= \sum_{i=1}^{d}\Big(\lambda_{n-i+1}(\Sigma_{\alpha,\tau,\beta})\lambda_i^2(\widetilde{\Sigma}) - 2\lambda_i(\widetilde{\Sigma}) - \tau\log\big(\lambda_i(\widetilde{\Sigma})\big)\Big).\end{aligned}$$

Consider the function

$$f(v) = uv^2 - 2v - \tau\log(v).$$

Take the derivative and set it to 0, we get

$$\frac{2uv^2 - 2v - \tau}{v} = 0,$$

which has unique positive solution $v^* = \frac{1+\sqrt{1+2u\tau}}{2u}$. It verifies that, to attain the minimization of Υ_Σ,

$$\lambda_i(\widetilde{\Sigma}) = \frac{1+\sqrt{1+2\tau\lambda_{n-i+1}(\Sigma_{\alpha,\tau,\beta})}}{2\lambda_{n-i+1}(\Sigma_{\alpha,\tau,\beta})}.$$

This results in the following expression

$$\widetilde{\Sigma} \quad \text{is similar to} \quad \overline{\Lambda} := \operatorname{diag}\left(\frac{1+\sqrt{1+2\tau\lambda_{n-i+1}(\Sigma_{\alpha,\tau,\beta})}}{2\lambda_{n-i+1}(\Sigma_{\alpha,\tau,\beta})}\right).$$

Since $\widetilde{\Sigma}$ and $\Sigma_{\alpha,\tau,\beta}$ are commuting, the eigenvalue of $\widetilde{\Sigma}$ can be computed from the eigenvalues of $\Sigma_{\alpha,\tau,\beta}$, we get

$$\widetilde{\Sigma}^* = \frac{1}{2}\Sigma_{\alpha,\tau,\beta}^{-1} + \frac{1}{2}\Big[\Sigma_{\alpha,\tau,\beta}^{-2} + 2\tau\Sigma_{\alpha,\tau,\beta}^{-1}\Big]^{\frac{1}{2}}.$$

The equation $2u(v^*)^2 - 2v^* - \tau = 0$ deduces that $(v^*)^2 = \frac{v^*}{u} + \frac{\tau}{2u}$. Hence, we yield

$$\big[\widetilde{\Sigma}^*\big]^2 = \frac{\tau}{2}\Sigma_{\alpha,\tau,\beta}^{-1} + \frac{1}{2}\Sigma_{\alpha,\tau,\beta}^{-2}\Big[\mathsf{Id} + \big(\mathsf{Id} + 2\tau\Sigma_{\alpha,\tau,\beta}\big)^{\frac{1}{2}}\Big].$$

That leads to the formula for Σ_x

$$\Sigma_x = \Sigma_\beta^{-\frac{1}{2}}\left[\frac{\tau}{2}\Sigma_{\alpha,\tau,\beta}^{-1} + \frac{1}{2}\Sigma_{\alpha,\tau,\beta}^{-2}\Big[\mathsf{Id} + \big(\mathsf{Id} + 2\tau\Sigma_{\alpha,\tau,\beta}\big)^{\frac{1}{2}}\Big]\right]\Sigma_\beta^{-\frac{1}{2}}.$$

For the shake of simplicity, we denote

$$\Sigma_\gamma = \frac{\tau}{2}\mathsf{Id} + \frac{1}{2}\Sigma_{\alpha,\tau,\beta}^{-1}\Big[\mathsf{Id} + \big(\mathsf{Id} + 2\tau\Sigma_{\alpha,\tau,\beta}\big)^{\frac{1}{2}}\Big],$$

then

$$\Sigma_x = \Sigma_\beta^{-\frac{1}{2}}\Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\gamma\Sigma_\beta^{-\frac{1}{2}}.$$

At this optimum, the function Υ_Σ takes the value

$$\Upsilon_\Sigma = \mathsf{tr}(\Sigma_\gamma) - \mathsf{tr}\left(\Big[\Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\gamma\Big]^{\frac{1}{2}}\right) - \frac{\tau}{2}\log\left(\det\Big[\Sigma_\gamma\Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\beta^{-1}\Big]\right),$$

and

$$\begin{aligned}\Upsilon = \quad & \mathsf{tr}(\Sigma_\gamma) + \mathsf{tr}(\Sigma_\beta) - \mathsf{tr}\left(\Big[\Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\gamma\Big]^{\frac{1}{2}}\right) - \frac{\tau}{2}\log\left(\det\Big[\Sigma_\gamma\Sigma_{\alpha,\tau,\beta}^{-1}\Sigma_\beta^{-1}\Sigma_\alpha^{-1}\Big]\right)\\ & + (\mathbf{a}-\mathbf{b})^\top\left\{\left(\big\|\Sigma_{\alpha,\tau}^{-1}\big\|_2^2 + 1\right)\mathsf{Id} - 2\Sigma_{\alpha,\tau}^{-1} + \frac{\tau}{2}\Sigma_{\alpha,\tau}^{-1}\Sigma_\alpha^{-1}\Sigma_{\alpha,\tau}^{-1}\right\}(\mathbf{a}-\mathbf{b}) - \frac{\tau d}{2}.\end{aligned}$$

Calculation of m_π: Recall that in order to find m_π, we minimize

$$m_\pi \Upsilon + \tau \mathrm{KL}\left(m_\pi \| m_\alpha\right).$$

Considering the mentioned Υ. As per Lemma B.2, we obtain the value of the optimizer m_π as

$$m_\pi = m_\alpha \exp\left\{\frac{-\Upsilon}{\tau}\right\}$$

and final objective function value

$$W^2_{2_{\mathrm{SUOT}}}(\alpha, \beta, \tau) = \tau m_\alpha \left(1 - \exp\left\{\frac{-\Upsilon}{\tau}\right\}\right).$$

Hence, we have thus proved our claims. In the preceding proof, the optimization of the trace operator using $K_{x\beta}$ is not constrained to a single choice. An alternative option involves utilizing $\Sigma_x^{\frac{1}{2}}\Sigma_\beta^{\frac{1}{2}}$ or the geometric mean of Σ_x and Σ_β. Additional details are in Lemma B.3.

Lemma B.3. *Given SPD matrices $A, B \in \mathbb{S}^d_{++}$. Moreover, A and B have the same unitary matric in SVD decomposition i.e. $A = S\Lambda_1^2 S^\top, B = S\Lambda_2^2 S^\top$. Then we have*

$$\mathsf{tr}(\left[A^{\frac{1}{2}} B A^{\frac{1}{2}}\right]^{\frac{1}{2}}) = \mathsf{tr}(B^{\frac{1}{2}} A^{\frac{1}{2}}).$$

Proof. We note that

$$\mathsf{tr}\left(\left[A^{\frac{1}{2}} B A^{\frac{1}{2}}\right]^{\frac{1}{2}}\right) = \mathsf{tr}\left(\left[AB\right]^{\frac{1}{2}}\right).$$

It follows that we need to prove

$$\mathsf{tr}\left(A^{\frac{1}{2}} B^{\frac{1}{2}}\right) = \mathsf{tr}\left(\left[AB\right]^{\frac{1}{2}}\right).$$

We have

$$A^{\frac{1}{2}} = S\Lambda_1 S^\top, \quad B^{\frac{1}{2}} = S\Lambda_2 S^T.$$

Thus, we get

$$\mathsf{tr}\left(A^{\frac{1}{2}} B^{\frac{1}{2}}\right) = \mathsf{tr}\left(\Lambda_1 \Lambda_2\right).$$

Furthermore, we have AB and $\Lambda_1^2\Lambda_2^2$ have the same set of eigenvalues. It gives

$$\mathsf{tr}\left(\left[AB\right]^{\frac{1}{2}}\right) = \sum_{i=1}^{d} \lambda_i\left(\left[AB\right]^{\frac{1}{2}}\right) = \sum_{i=1}^{d} \lambda_i\left(\left[\Lambda_1^2\Lambda_2^2\right]^{\frac{1}{2}}\right) = \mathsf{tr}\left(\Lambda_1\Lambda_2\right).$$

Then we deduce that

$$\mathsf{tr}\left(\left[A^{\frac{1}{2}} B A^{\frac{1}{2}}\right]^{\frac{1}{2}}\right) = \mathsf{tr}\left(A^{\frac{1}{2}} B^{\frac{1}{2}}\right).$$

□

Now to see the final expression, from Altschuler et al. (2021), we have Wasserstein distance between Gaussians $\alpha = \mathcal{N}(\mathbf{a}, \Sigma_\alpha), \beta = \mathcal{N}(\mathbf{b}, \Sigma_\beta)$ as

$$W_2^2\left(\alpha, \beta\right) = \|\mathbf{a} - \mathbf{b}\|_2^2 + \mathrm{tr}\left(\Sigma_\alpha + \Sigma_\beta - 2\left[\Sigma_\alpha^{\frac{1}{2}}\Sigma_\beta\Sigma_\alpha^{\frac{1}{2}}\right]^{\frac{1}{2}}\right).$$

Next, with the assumption that $m_\alpha = m_\beta$, the SUOT distance reads

$$W^2_{2_{\text{SUOT}}}(\alpha, \beta; \tau) := \min \|\mathbf{a_x} - \mathbf{b}\|_2^2 + \text{tr}(\Sigma_x) + \text{tr}(\Sigma_\beta) - 2\text{tr}(K_{x\beta}) + \tau \text{KL}(\pi_x \| \alpha).$$

From Lemma B.3, at the optimal solution, the term $\text{tr}\left(K_{x\beta}\right)$ takes the value $\text{tr}\Big(\big[\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\big]^{\frac{1}{2}}\Big)$. This leads to

$$W^2_{2_{\text{SUOT}}}(\alpha, \beta, \tau) = W_2^2(\pi_x, \beta) + \tau \text{KL}(\pi_x \| \Sigma_\alpha).$$

As a consequence, we obtain the full conclusion of the theorem. □

Upon to this formula, we have the results that our UOT distance is bounded by Wasserstein distance.

Proposition B.4. *Giving two centered Gaussians α, β. There exists a positive constant ξ_τ that we have*

$$\xi_\tau W_2^2(\Sigma_\alpha, \Sigma_\beta) \leq W^2_{2_{\text{SUOT}}}(\Sigma_\alpha, \Sigma_\beta) \leq W_2^2(\Sigma_\alpha, \Sigma_\beta).$$

Proof. The second inequality is straight forward while the first inequality happens due to Talagrand inequality Otto and Villani (2000), we have $\text{KL}(\Sigma_x \| \Sigma_\alpha) \geq \frac{\xi}{2} W_2^2(\Sigma_x, \Sigma_\alpha)$ for a constant ξ. It leads to

$$\begin{aligned}
&W^2_{2_{\text{SUOT}}}(\Sigma_\alpha, \Sigma_\beta) \\
&= W_2^2(\Sigma_x, \Sigma_\beta) + \tau \text{KL}(\Sigma_x \| \Sigma_\alpha) \\
&\geq W_2^2(\Sigma_x, \Sigma_\beta) + \frac{\xi\tau}{2} W_2^2(\Sigma_x, \Sigma_\alpha) \\
&\geq \min\{1, \frac{\xi\tau}{2}\} \left(W_2^2(\Sigma_x, \Sigma_\beta) + W_2^2(\Sigma_x, \Sigma_\alpha)\right) \\
&\geq \frac{\min\{1, \frac{\xi\tau}{2}\}}{2} \left(W_2\left(\Sigma_x, \Sigma_\beta\right) + W_2\left(\Sigma_x, \Sigma_\alpha\right)\right)^2 \\
&\geq \underbrace{\frac{\min\{1, \frac{\xi\tau}{2}\}}{2}}_{\xi_\tau} W_2^2(\Sigma_\alpha, \Sigma_\beta).
\end{aligned}$$

□

B.2 Proof for Theorem 4.3

Proof. We follow the proof of Theorem (4.2) to obtain the explicit form of the minimizer and objective function. Here the two marginals of $\overline{\pi}$ are denoted by $\overline{\pi}_x$ and $\overline{\pi}_y$ where $\overline{\pi}_x$ has mean $\mathbf{0}$ and covariance matrix Σ_x while $\overline{\pi}_y$ has mean $\mathbf{0}$ and covariance matrix $\Sigma_y = \Sigma_\beta$. Let π_x and π_y be two marginals of π, then $\pi_x = m_\pi \overline{\pi}_x$ and $\pi_y = m_\pi \overline{\pi}_y$. We also have

$$\mathbb{E}_\pi \|X - Y\|_2^2 = m_\pi \Big\{ \mathsf{tr}\left(\Sigma_x\right) + \mathsf{tr}\left(\Sigma_\beta\right) - 2\mathsf{tr}\left(K_{x\beta}\right) \Big\}.$$

According to Lemma B.1,

$$\begin{aligned}
\text{KL}\left(\pi_x \| \alpha\right) &= m_\pi \text{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) + \text{KL}\left(m_\pi \| m_\alpha\right) \\
\text{KL}\left(\pi_x \| \alpha \otimes \beta\right) &= m_\pi \text{KL}\left(\overline{\pi}_x \| \overline{\alpha} \otimes \bar{\beta}\right) + \text{KL}\left(m_\pi \| m_\alpha m_\beta\right).
\end{aligned}$$

Combining the above results, the entropic objective function between α and β reads as

$$\begin{aligned}
W^2_{2_{\text{SUOT}},\delta}(\alpha, \beta; \tau) := \min_{\pi \in \mathcal{M}^+(\mathbb{R}^d \times \mathbb{R}^d)} & m_\pi \left\{ \mathsf{tr}\left(\Sigma_x\right) + \mathsf{tr}\left(\Sigma_\beta\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \tau \text{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) + \delta \text{KL}\left(\overline{\pi} \| \overline{\alpha} \otimes \overline{\beta}\right) \right\} \\
&+ \tau \text{KL}\left(m_\pi \| m_\alpha\right) + \delta \text{KL}\left(m_\pi \| m_\alpha m_\beta\right).
\end{aligned}$$

Denote

$$\Upsilon = \mathsf{tr}\left(\Sigma_x\right) + \mathsf{tr}\left(\Sigma_\beta\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \tau \text{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) + \delta \text{KL}\left(\overline{\pi} \| \overline{\alpha} \otimes \overline{\beta}\right).$$

Due to the independence of Υ with m_π, we could minimize Υ then find m_π. For a fixed Υ, take derivative according to m_π of

$$m_\pi \Upsilon + \tau \mathrm{KL}\left(m_\pi \| m_\alpha\right) + \delta \mathrm{KL}\left(m_\pi \| m_\alpha m_\beta\right).$$

and set equation to 0, we obtain the value of the optimizer m_π as

$$m_\pi = m_\alpha m_\beta^{\frac{\delta}{\tau+\delta}} \exp\left(\frac{-\Upsilon}{\tau+\delta}\right).$$

Minimization of Υ: For the KL term, due to Lemma B.1 we have

$$\begin{aligned}
\mathrm{KL}\left(\overline{\pi}_x \| \overline{\alpha}\right) &= \frac{1}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) - d + \log\left(\frac{\det\left(\Sigma_\alpha\right)}{\det\left(\Sigma_x\right)}\right)\right] \\
\mathrm{KL}\left(\overline{\pi} \| \overline{\alpha} \otimes \overline{\beta}\right) &= \frac{1}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) - d + \log\left(\frac{\det\left(\Sigma_\alpha\right)}{\det\left(\Sigma_x\right)}\right)\right] - \frac{1}{2}\sum_{i=1}^{d}\log\left(1-\lambda_i(H)\right),
\end{aligned}$$

where $(\lambda_i(H))$ is the i-th largest singular value of $H := \Sigma_x^{-\frac{1}{2}} K_{x\beta} \Sigma_\beta^{-\frac{1}{2}}$ for all $i \in [d]$. Now Υ reads

$$\Upsilon = \mathsf{tr}\left(\Sigma_x\right) - 2\mathsf{tr}\left(K_{x\beta}\right) + \frac{(\tau+\delta)}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) + \log\left(\frac{\det\left(\Sigma_\alpha\right)}{\det\left(\Sigma_x\right)}\right)\right] - \frac{\delta}{2}\sum_{i=1}^{d}\log\left(1-\lambda_i(H)\right) + \mathsf{tr}(\Sigma_\beta) - \frac{(\tau+\delta)d}{2}.$$

First, we embark on the task of maximizing $\mathsf{tr}(K_{x\beta})$ while fixing Σ_x. Noting that $K_{x\beta} = \Sigma_x^{\frac{1}{2}} H \Sigma_\beta^{\frac{1}{2}}$, it follows that $\mathsf{tr}\left(K_{x\beta}\right) = \mathsf{tr}\left(H\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)$. Let $\left(\lambda_i(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}})\right)_{i=1}^{d}$ represent the singular values of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$ in descending order. Applying von Neuman inequality, we get

$$\mathsf{tr}(K_{x\beta}) \le \sum_{i=1}^{d} \lambda_i\big(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\big)\lambda_i(H).$$

Now our task turns into maximizing

$$\sum_{i=1}^{d}\Big[2\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)\lambda_i(H) + \frac{\delta}{2}\log\left(1-\lambda_i(H)\right)\Big].$$

Consider the function $f(\lambda_i(H)) = 2\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)\lambda_i(H) + \frac{\delta}{2}\log\left(1-\lambda_i(H)\right)$. Taking derivative and setting it to 0 we get

$$\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) - \frac{\delta}{4}\frac{1}{1-\lambda_i(H)} = 0 \qquad \Leftrightarrow \qquad \lambda_i(H) = 1 - \frac{\delta}{4\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)}.$$

Note that due to Lemma 3.4 Le et al. (2022), all values of $\lambda_i(H)$ should be in interval $[0,1]$. Hence,

- If $\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) \ge \frac{\delta}{4}$ (equivalent to $1 - \frac{\delta}{4\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)} \ge 0$), maximum value attains at

$$\lambda_i^*(H) = 1 - \frac{\delta}{4\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)}.$$

- If $\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) < \frac{\delta}{4}$ (equivalent to $1 - \frac{\delta}{4\lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right)} < 0$), maximum value attains at

$$\lambda_i^*(H) = 0.$$

In conclusion

$$\lambda_i^*(H) = \begin{cases} 1 - \frac{\delta}{4}\lambda_i^{-1}\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) & \text{if } \lambda_i\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) \geq \frac{\delta}{4} \\ 0 & \text{otherwise} \end{cases}.$$

In this proof, we assume that δ is small enough that $\lambda_i^*(H) = 1 - \frac{\delta}{4}\lambda_i^{-1}\left(\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}\right) \forall i$. Since the equality of von Neuman inequality holds when H and $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$ are commuting, the eigenvalues of H could be calculated from eigenvalues of $\Sigma_\beta^{\frac{1}{2}}\Sigma_x^{\frac{1}{2}}$; we obtain

$$H = \mathsf{Id} - \frac{\delta}{4}\Sigma_x^{-\frac{1}{2}}\Sigma_\beta^{-\frac{1}{2}},$$

which gives

$$K_{x\beta} = \Sigma_x^{\frac{1}{2}}\Sigma_\beta^{\frac{1}{2}} - \frac{\delta}{4}\mathsf{Id},$$

and

$$\begin{aligned}\sum_{i=1}^{d}\log\left(1-\lambda_i(H)\right) &= \log\left(\Pi_{i=1}^{d}\left(1-\lambda_i\left(H\right)\right)\right) \\ &= \log\left(\det\left(\mathsf{Id}-H\right)\right) \\ &= \log\left((\frac{\delta}{4})^d\det\left(\Sigma_x^{-\frac{1}{2}}\Sigma_\beta^{-\frac{1}{2}}\right)\right) \\ &= d\log(\frac{\delta}{4}) + \log\left(\det\left(\Sigma_x^{-\frac{1}{2}}\right)\right) + \log\left(\det\left(\Sigma_\beta^{-\frac{1}{2}}\right)\right).\end{aligned}$$

In such a scenario, we care about minimizing

$$\begin{aligned}\Upsilon = \quad & \mathsf{tr}\left(\Sigma_x\right) - 2\mathsf{tr}\left(\Sigma_x^{\frac{1}{2}}\Sigma_\beta^{\frac{1}{2}} - \frac{\delta}{4}\mathsf{Id}\right) + \frac{(\tau+\delta)}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) + \log\left(\frac{\det\left(\Sigma_\alpha\right)}{\det\left(\Sigma_x\right)}\right)\right] \\ & - \frac{\delta}{2}\left[d\log(\frac{\delta}{4}) + \log\left(\det\left(\Sigma_x^{-\frac{1}{2}}\right)\right) + \log\left(\det\left(\Sigma_\beta^{-\frac{1}{2}}\right)\right)\right] + \mathsf{tr}(\Sigma_\beta) - \frac{(\tau+\delta)d}{2}. \\ = \quad & \mathsf{tr}\left(\Sigma_x\right) - 2\mathsf{tr}\left(\Sigma_x^{\frac{1}{2}}\Sigma_\beta^{\frac{1}{2}}\right) + \frac{(\tau+\delta)}{2}\left[\mathsf{tr}\left(\Sigma_\alpha^{-1}\Sigma_x\right) - \log\left(\det\left(\Sigma_x\right)\right)\right] - \frac{\delta}{2}\log\left(\det\left(\Sigma_x^{-\frac{1}{2}}\right)\right) \\ & + \frac{\tau+\delta}{2}\log\left(\det(\Sigma_\alpha)\right) - \frac{\delta}{2}\log\left(\det\left(\Sigma_\beta^{-\frac{1}{2}}\right)\right) + \mathsf{tr}(\Sigma_\beta) - \frac{\tau d}{2} - \frac{\delta d}{2}\log(\frac{\delta}{4}).\end{aligned}$$

Letting $\Sigma_x^{\frac{1}{2}}\Sigma_\beta^{\frac{1}{2}} = \widetilde{\Sigma}$, $\Sigma_{\alpha,\beta,\tau,\delta} = \Sigma_\beta^{-\frac{1}{2}}\left(\mathsf{Id} + \frac{\tau+\delta}{2}\Sigma_\alpha^{-1}\right)\Sigma_\beta^{-\frac{1}{2}}$, then

$$\begin{aligned}\Upsilon =& \mathsf{tr}\left(\widetilde{\Sigma}^2\Sigma_{\alpha,\beta,\tau,\delta}\right) - 2\mathsf{tr}\left(\widetilde{\Sigma}\right) - (\tau+\frac{3}{2}\delta)\log\left(\det\left(\widetilde{\Sigma}\right)\right) - \delta\log\left(\det\left(\Sigma_\beta^{\frac{1}{2}}\right)\right) \\ & + \frac{\tau+\delta}{2}\left[\log\left(\det(\Sigma_\alpha)\right) + \log\left(\det\left(\Sigma_\beta\right)\right)\right] + \mathsf{tr}(\Sigma_\beta) - \frac{\tau d}{2} - \frac{\delta d}{2}\log(\frac{\delta}{4}).\end{aligned}$$

The only left part now is minimizing

$$\mathsf{tr}\left(\widetilde{\Sigma}^2\Sigma_{\alpha,\beta,\tau,\delta}\right) - 2\mathsf{tr}\left(\widetilde{\Sigma}\right) - (\tau+\frac{3}{2}\delta)\log\left(\det\left(\widetilde{\Sigma}\right)\right).$$

Note that $\widetilde{\Sigma}$ is symmetric, so there exists a diagonal matrix Λ that $\widetilde{\Sigma}$ is similar to Λ (which implies $\widetilde{\Sigma}^2$ is similar to Λ^2. Assume that $\Lambda = \mathrm{diag}\left(\lambda_i(\widetilde{\Sigma})\right)_{i=1}^d$ which is decreasingly ordered and $\left(\lambda_i(\Sigma_{\alpha,\beta,\tau,\delta})\right)_{i=1}^d$ are eigenvalues of $\Sigma_{\alpha,\beta,\tau,\delta}$ in ascending order, by Ruhe's trace inequality

$$\mathsf{tr}\left(\widetilde{\Sigma}^2\Sigma_{\alpha,\beta,\tau,\delta}\right) \geq \sum_{i=1}^{d}\lambda_i(\widetilde{\Sigma})^2\lambda_{d-i+1}(\Sigma_{\alpha,\beta,\tau,\delta}),$$

where the equality holds when $\widetilde{\Sigma}^2$ and $\Sigma_{\alpha,\beta,\tau}$ are commuting. The optimization part now is calculated on eigenvalues of Λ

$$\sum_{i=1}^{d} \lambda_i^2(\widetilde{\Sigma})\lambda_{n-i+1}(\Sigma_{\alpha,\beta,\tau,\delta}) - 2\sum_{r=1}^{d} \lambda_i(\widetilde{\Sigma}) - (\tau + \frac{3}{2}\delta)\log\left(\prod_{i=1}^{d} \lambda_i(\widetilde{\Sigma})\right)$$
$$= \sum_{i=1}^{d} \left(\lambda_{n-i+1}(\Sigma_{\alpha,\beta,\tau,\delta})\lambda_i^2(\widetilde{\Sigma}) - 2\lambda_i(\widetilde{\Sigma}) - (\tau + \frac{3}{2}\delta)\log\left(\lambda_i\big(\widetilde{\Sigma}\big)\right)\right).$$

Consider the function

$$f(v) = uv^2 - 2v - \theta\log(v).$$

Take the derivative and set it to 0, we get

$$\frac{2uv^2 - 2v - \theta}{v} = 0,$$

which has unique positive solution $v^* = \frac{1+\sqrt{1+2u\theta}}{2u}$. It verifies that, to attain the minimization of Υ_Σ,

$$\lambda_i(\widetilde{\Sigma}) = \frac{1 + \sqrt{1 + (2\tau + 3\delta)\lambda_{n-i+1}(\Sigma_{\alpha,\beta,\tau,\delta})}}{2\lambda_{n-i+1}(\Sigma_{\alpha,\beta,\tau,\delta})}.$$

This results in the following expression

$$\widetilde{\Sigma} \quad \text{is similar to} \quad \overline{\Lambda} := \operatorname{diag}\left(\frac{1 + \sqrt{1 + (2\tau + 3\delta)\lambda_{n-i+1}(\Sigma_{\alpha,\beta,\tau,\delta})}}{2\lambda_{n-i+1}(\Sigma_{\alpha,\beta,\tau,\delta})}\right).$$

Since $\widetilde{\Sigma}$ and $\Sigma_{\alpha,\beta,\tau,\delta}$ are commuting, the eigenvalue of $\widetilde{\Sigma}$ can be computed from the eigenvalues of $\Sigma_{\alpha,\beta,\tau,\delta}$, we get

$$\widetilde{\Sigma}^* = \frac{1}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-1} + \frac{1}{2}\Big[\Sigma_{\alpha,\beta,\tau,\delta}^{-2} + (2\tau + 3\delta)\Sigma_{\alpha,\beta,\tau,\delta}^{-1}\Big]^{\frac{1}{2}}.$$

The equation $2u(v^*)^2 - 2v^* - \theta = 0$ deduces that $(v^*)^2 = \frac{v^*}{u} + \frac{\theta}{2u}$. Hence, we have

$$\big[\widetilde{\Sigma}^*\big]^2 = \frac{\tau}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-1} + \frac{1}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-2}\Big[\mathsf{Id} + \big(\mathsf{Id} + (2\tau + 3\delta)\Sigma_{\alpha,\beta,\tau,\delta}\big)^{\frac{1}{2}}\Big].$$

That leads to the formula for Σ_x

$$\Sigma_x = \Sigma_\beta^{-\frac{1}{2}}\left[\frac{\tau}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-1} + \frac{1}{2}\Sigma_{\alpha,\beta,\tau,\delta}^{-2}\Big[\mathsf{Id} + \big(\mathsf{Id} + (2\tau + 3\delta)\Sigma_{\alpha,\beta,\tau,\delta}\big)^{\frac{1}{2}}\Big]\right]\Sigma_\beta^{-\frac{1}{2}}.$$

Hence we complete the proof. □

C Proof for Theorem 4.4

First we need these below lemmas

Lemma C.1 (Trace)**.** *Let A and B be SPD matrices of the same size. Then*

1. $\mathsf{tr}\big(\big[AB\big]^{\frac{1}{2}}\big) = \mathsf{tr}\big(\big[BA\big]^{\frac{1}{2}}\big) = \mathsf{tr}\big(\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^{\frac{1}{2}}\big) = \mathsf{tr}\big(\big[B^{\frac{1}{2}}AB^{\frac{1}{2}}\big]^{\frac{1}{2}}\big)$.
2. $\mathsf{tr}\Big(\Big[(AB)^2 + \tau(AB)\Big]^{\frac{1}{2}}\Big) = \mathsf{tr}\Big(\Big\{\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^2 + \tau\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]\Big\}^{\frac{1}{2}}\Big)$.

Proof. For part (**1**), first we note that

$$\big[AB\big]^{\frac{1}{2}} = A^{\frac{1}{2}}\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^{\frac{1}{2}}A^{-\frac{1}{2}}.$$

Thus, we get

$$\mathsf{tr}\big(\big[AB\big]^{\frac{1}{2}}\big) = \mathsf{tr}\big(\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^{\frac{1}{2}}\big),$$
$$\mathsf{tr}\big(\big[BA\big]^{\frac{1}{2}}\big) = \mathsf{tr}\big(\big[B^{\frac{1}{2}}AB^{\frac{1}{2}}\big]^{\frac{1}{2}}\big).$$

The above equations also means that $\big[AB\big]^{\frac{1}{2}}$ and $\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^{\frac{1}{2}}$ are similar matrices. That follows that their eigenvalues sets are coincidence. Note that the eigenvalues of $\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^{\frac{1}{2}}$ are square root of eigenvalues of $\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]$. Furthermore,

$$A^{\frac{1}{2}}BA^{\frac{1}{2}} = A^{\frac{1}{2}}B^{\frac{1}{2}}B^{\frac{1}{2}}A^{\frac{1}{2}}.$$

Note that matrices UV and VU have the same set of eigenvalues. Hence, $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ and $B^{\frac{1}{2}}AB^{\frac{1}{2}}$ have the same set of eigenvalues. Finally, all four matrices have the same sets of eigenvalues.

For part (**2**), since the formula in part (a) between $[AB]^{\frac{1}{2}}$ and $\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^{\frac{1}{2}}$, we have

$$\begin{aligned}(AB)^2 + \tau(AB) &= A^{\frac{1}{2}}\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^2A^{-\frac{1}{2}} + \tau A^{\frac{1}{2}}\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]A^{-\frac{1}{2}}\\ &= A^{\frac{1}{2}}\Big\{\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^2 + \tau\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]\Big\}A^{-\frac{1}{2}}.\end{aligned}$$

It follows that

$$\Big[(AB)^2 + \tau(AB)\Big]^{\frac{1}{2}} = A^{\frac{1}{2}}\Big\{\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^2 + \tau\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]\Big\}^{\frac{1}{2}}A^{-\frac{1}{2}}.$$

We deduce that

$$\mathsf{tr}\Big(\Big[(AB)^2 + \tau(AB)\Big]^{\frac{1}{2}}\Big) = \mathsf{tr}\Big(\Big\{\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]^2 + \tau\big[A^{\frac{1}{2}}BA^{\frac{1}{2}}\big]\Big\}^{\frac{1}{2}}\Big).$$

□

Lemma C.2. *Let A and B be two symmetric matrices of the same size. Then*

$$\frac{\partial \log\det(A+tB)}{\partial t}\Big|_{t=0} = \mathsf{tr}\big(A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big).$$

Proof. Note that $A + tB = A^{\frac{1}{2}}\big[\mathsf{Id} + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big]A^{\frac{1}{2}}$. Then

$$\begin{aligned}\log\det(A+tB) &= \log\det(A) + \log\det\big(\mathsf{Id} + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big)\\ &= \log\det(A) + \sum_{i=1}^{d}\log\Big(1 + t\lambda_i\big(A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big)\Big).\end{aligned}$$

Taking derivative with respect to t of both sides

$$\begin{aligned}\frac{\partial \log\det(A+tB)}{\partial t}\Big|_{t=0} &= \sum_{i=1}^{d}\frac{\lambda_i\big(A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big)}{1 + t\lambda_i\big(A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big)}\Big|_{t=0} = \sum_{i=1}^{d}\lambda_i\big(A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big)\\ &= \mathsf{tr}\big(A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\big).\end{aligned}$$

□

The next lemma is about the Taylor expansion for trace of square root matrix.

Lemma C.3. *Given $B \in \mathbb{S}^d_{++}$, we define the Lyapunov's operator $\mathcal{L}$ as:*

$$\begin{aligned}\mathcal{L}: \quad \mathbb{S}^d_{++} &\to \mathbb{S}^d_{++}\\ A &\mapsto \mathcal{L}_B[A]\end{aligned}$$

where $\mathcal{L}_B[A]$ is the matric satisfying $\mathcal{L}_B[A]B + B\mathcal{L}_B[A] = A$. Then let Σ_0 and A be SPD matrices in $\mathbb{S}^d_{++}$,

$$\mathsf{tr}\big(\big[\Sigma_0 + tA\big]^{\frac{1}{2}}\big) = \mathsf{tr}\big(\Sigma_0^{\frac{1}{2}}\big) + \frac{1}{2}t\mathsf{tr}\big(\mathcal{L}_{\Sigma_0^{\frac{1}{2}}}[A]\big) + o(t).$$

Moreover, we have

$$\mathsf{tr}\big(\mathcal{L}_{\Sigma_0^{\frac{1}{2}}}[A]\big) = \mathsf{tr}\big(\Sigma_0^{-\frac{1}{2}}A\big).$$

Proof. Assume that $\big[\Sigma_0 + tA\big]^{\frac{1}{2}} = \Sigma_0^{\frac{1}{2}} + tX$, then

$$\begin{aligned} &\Sigma_0 + tA = \big[\Sigma_0^{\frac{1}{2}} + tX\big]^2 = \Sigma_0 + t\big[\Sigma_0^{\frac{1}{2}}X + X\Sigma_0^{\frac{1}{2}}\big] + t^2X^2 \\ \Rightarrow\quad &A = \Sigma_0^{\frac{1}{2}}X + X\Sigma_0^{\frac{1}{2}}. \end{aligned}$$

Then the Lyapunov's operator produces X from Σ_0 and A is equal to $X = \mathcal{L}_{\Sigma_0^{\frac{1}{2}}}[A]$. We have

$$\Sigma_0^{-\frac{1}{2}}A = X + \Sigma_0^{-\frac{1}{2}}X\Sigma_0^{\frac{1}{2}},$$

that leads to $\frac{1}{2}\mathsf{tr}(\Sigma_0^{-\frac{1}{2}}A) = \mathsf{tr}(X)$. □

Now we are ready to give the full proof for Theorem 4.4.

Proof. Use the notation Σ_x as an optimal solution to problem 4.2. Denote

$$\Big[\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\Big]^{\frac{1}{2}} = \Sigma_{x,\beta}.$$

Since $\Sigma_{x,\beta}$ and $\Sigma_{\alpha,\tau,\beta}$ share the same set of eigenvectors, we have

$$\Sigma_{x,\beta} = \frac{1}{2}\Sigma^{-1}_{\alpha,\tau,\beta} + \frac{1}{2}\Big[\Sigma^{-2}_{\alpha,\tau,\beta} + 2\tau\Sigma^{-1}_{\alpha,\tau,\beta}\Big]^{\frac{1}{2}}.$$

The objective function $W^2_{2_{\mathrm{SUOT}}}(\alpha,\beta,\tau)$ is equal to

$$\mathsf{tr}\big(\Sigma_\beta\big) - \mathsf{tr}\big(\Sigma_{x,\beta}\big) - \frac{\tau}{4}\log\det(\Sigma_x) + \mathsf{constant}.$$

Let $\gamma(t) = \Sigma_{\beta\to z,t}$ for $t\in[0,1]$ be the geodesic on Bures-Wasserstein manifold from Σ_β to Σ_z, then

$$\begin{aligned} \Sigma_{\beta\to z,t} &= \Big[\mathsf{Id} + t\big(T_{\Sigma_\beta\to\Sigma_z} - \mathsf{Id}\big)\Big]\Sigma_\beta\Big[\mathsf{Id} + t\big(T_{\Sigma_\beta\to\Sigma_z} - \mathsf{Id}\big)\Big] \\ &= \Sigma_\beta + t\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big) + t^2T_{\beta z}\Sigma_\beta T_{\beta z}, \end{aligned}$$

where $T_{\beta z} = T_{\Sigma_\beta\to\Sigma_z} - \mathsf{Id}$. We consider the derivative of the loss function on the geodesic $\gamma(t)$.

Derivative of $\mathsf{tr}(\Sigma_\beta)$: By the above formula,

$$\frac{\partial\mathsf{tr}(\Sigma_{\beta\to z,t})}{\partial t}\Big|_{t=0} = 2\mathsf{tr}(\Sigma_\beta T_{\beta z}) = \big\langle 2\mathsf{Id}, T_{\beta z}\big\rangle_{\Sigma_\beta}.$$

Derivative of $\mathsf{tr}\big(\Sigma_{x,\beta}\big)$: We recall the formula of $\Sigma_{x,\beta}$ as

$$\mathsf{tr}\big(\Sigma_{x,\beta}\big) = \mathsf{tr}\bigg(\frac{1}{2}\Big\{\Sigma^{-1}_{\alpha,\tau,\beta} + \Big[\Sigma^{-2}_{\alpha,\tau,\beta} + \tau\Sigma^{-1}_{\alpha,\tau,\beta}\Big]^{\frac{1}{2}}\Big\}\bigg).$$

We deal with each term separately. We start with $\Sigma_{\alpha,\tau,\beta}$.

$$\begin{aligned}
&\Sigma_{\alpha,\tau,\beta} = \Sigma_\beta^{-\frac{1}{2}}\Big(\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\Big)\Sigma_\beta^{-\frac{1}{2}} \\
\Rightarrow\quad &\Sigma_{\alpha,\tau,\beta}^{-1} = \Sigma_\beta^{\frac{1}{2}}\Big(\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\Big)^{-1}\Sigma_\beta^{\frac{1}{2}} \\
\Rightarrow\quad &\mathsf{tr}\big(\Sigma_{\alpha,\tau,\beta}^{-1}\big) = \mathsf{tr}\Big(\Sigma_\beta\big[\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\big]^{-1}\Big) = \mathsf{tr}\big(\Sigma_\beta\Sigma_{\alpha,\tau}^{-1}\big).
\end{aligned}$$

It follows that

$$\frac{\partial \mathsf{tr}(\Sigma_{\alpha,\tau,\beta\to z,t}^{-1})}{\partial t}\Big|_{t=0} = \big\langle 2\Sigma_{\alpha,\tau}^{-1}, T_{\beta z}\big\rangle_{\Sigma_\beta}.$$

We move to the next term. We first recall that

$$\begin{aligned}
\mathsf{tr}(\Sigma_{x,\beta}) &= \mathsf{tr}\Big(\Big[\Sigma_\beta^{\frac{1}{2}}\Sigma_x\Sigma_\beta^{\frac{1}{2}}\Big]^{\frac{1}{2}}\Big) = \mathsf{tr}\Big(\Big[\Sigma_x^{\frac{1}{2}}\Sigma_\beta\Sigma_x^{\frac{1}{2}}\Big]^{\frac{1}{2}}\Big) \\
\Sigma_{\alpha,\tau,\beta}^{-1} &= \Sigma_\beta^{\frac{1}{2}}\Big(\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\Big)^{-1}\Sigma_\beta^{\frac{1}{2}} = \Sigma_\beta^{\frac{1}{2}}\Sigma_{\alpha,\tau}\Sigma_\beta^{\frac{1}{2}}.
\end{aligned}$$

If we define

$$\Sigma_{\alpha,\beta,\tau}^{-1} = \Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}},$$

then $\Sigma_{\alpha,\tau,\beta}^{-1}$ and $\Sigma_{\alpha,\beta,\tau}^{-1}$ share the same set of eigenvalues. Hence,

$$\mathsf{tr}\Big(\Big[\Sigma_{\alpha,\tau,\beta}^{-2} + \tau\Sigma_{\alpha,\tau,\beta}^{-1}\Big]^{\frac{1}{2}}\Big) = \mathsf{tr}\Big(\Big[\Sigma_{\alpha,\beta,\tau}^{-2} + \tau\Sigma_{\alpha,\beta,\tau}^{-1}\Big]^{\frac{1}{2}}\Big).$$

Instead of working with the LHS, we work with the RHS. Then, we have

$$\mathsf{tr}\Big(\Big[\Sigma_{\alpha,\beta,\tau}^{-2} + \tau\Sigma_{\alpha,\beta,\tau}^{-1}\Big]^{\frac{1}{2}}\Big) = \mathsf{tr}\Big(\Big\{\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]^2 + \tau\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]\Big\}^{\frac{1}{2}}\Big).$$

Replace Σ_β by $\Sigma_{\beta\to z,t} = \Sigma_\beta + t\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big) + t^2 T_{\beta z}\Sigma_\beta T_{\beta z}$, we have

$$\begin{aligned}
\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_{\beta\to z,t}\Sigma_{\alpha,\tau}^{-\frac{1}{2}} =& \Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}} + t\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}^{-\frac{1}{2}} + t^2\Sigma_{\alpha,\tau}^{-\frac{1}{2}}T_{\beta z}\Sigma_\beta T_{\beta z}\Sigma_{\alpha,\tau}^{-\frac{1}{2}} \\
\Big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_{\beta\to z,t}\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Big]^2 =& \Big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Big]^2 + \mathcal{O}(t^3) + \\
& t\Big\{\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big] + \big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]\Big\} + \\
& t^2\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Big[T_{\beta z}\Sigma_\beta T_{\beta z}\Sigma_{\alpha,\tau}^{-1}\Sigma_\beta + \Sigma_\beta\Sigma_{\alpha,\tau}^{-1}T_{\beta z}\Sigma_\beta T_{\beta z} + \big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}^{-1}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Big]\Sigma_{\alpha,\tau}^{-\frac{1}{2}}.
\end{aligned}$$

Denote

$$\Big\{\Big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Big]^2 + \tau\Big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Big]\Big\}^{\frac{1}{2}} = \widetilde{\Sigma}_{\beta,\alpha,\tau}.$$

Taking derivative gives

$$\begin{aligned}
&2\frac{\partial \mathsf{tr}\Big(\Big\{\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_{\beta\to z,t}\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]^2 + \tau\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_{\beta\to z,t}\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]\Big\}^{\frac{1}{2}}\Big)}{\partial t}\Big|_{t=0} \\
&= \tau\mathsf{tr}\Big(\widetilde{\Sigma}_{\beta,\alpha,\tau}^{-1}\big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big]\Big). \\
&\quad + \mathsf{tr}\Big(\widetilde{\Sigma}_{\beta,\alpha,\tau}^{-1}\Big[\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}\Sigma_\beta\Sigma_{\alpha,\tau}^{-\frac{1}{2}} + \Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Sigma_\beta\Sigma_{\alpha,\tau}\big(T_{\beta z}\Sigma_\beta + \Sigma_\beta T_{\beta z}\big)\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\Big]\Big)
\end{aligned}$$

The first term of the RHS is equal to

$$\mathsf{tr}\Big(\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\widetilde{\Sigma}_{\beta,\alpha,\tau}^{-1}\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\big[T_{\beta z}\Sigma_\beta+\Sigma_\beta T_{\beta z}\big]\Big)=\Big\langle 2M,T_{\beta z}\Big\rangle_{\Sigma_\beta},$$

where we denote

$$M=\Sigma_{\alpha,\tau}^{-\frac{1}{2}}\widetilde{\Sigma}_{\beta,\alpha,\tau}^{-1}\Sigma_{\alpha,\tau}^{-\frac{1}{2}}.$$

Similarly, the second term of the RHS is equal to

$$\mathsf{tr}\Big(\big[\Sigma_{\alpha,\tau}^{-1}\Sigma_\beta M+M\Sigma_\beta\Sigma_{\alpha,\tau}^{-1}\big]\big[T_{\beta z}\Sigma_\beta+\Sigma_\beta T_{\beta z}\big]\Big)=\Big\langle 2U,T_{\beta z}\Big\rangle_{\Sigma_\beta},$$

where we denote

$$U=\Sigma_{\alpha,\tau}^{-1}\Sigma_\beta M+M\Sigma_\beta\Sigma_{\alpha,\tau}^{-1}.$$

Hence, the derivative of RHS is equal to

$$\big\langle \tau M+U,T_{\beta z}\big\rangle_{\Sigma_\beta}$$

and final derivative of $\mathsf{tr}(\Sigma_{x,\beta})$ is

$$\big\langle \Sigma_{\alpha,\tau}^{-1}+\frac{1}{2}\big[\tau M+U\big],T_{\beta z}\big\rangle_{\Sigma_\beta}.$$

Derivative of the log det: We recall some formulas

$$\det(\Sigma_x)=\det(\Sigma_{\beta\to z,t})^{-1}\det\left(\Sigma_{x^*,\beta\to z,t}\right)^2$$

Taking the logarithm of both sides

$$\log\det(\Sigma_x)=-\log\det(\Sigma_{\beta\to z,t})+2\log\det(\Sigma_{x^*,\beta\to z,t}).$$

For the second term,

$$\Sigma_{x^*,\beta\to z,t}=\Sigma_{\alpha,\tau,\beta\to z,t}^{-1}\Big\{\mathsf{Id}+\big[\mathsf{Id}+\tau\Sigma_{\alpha,\tau,\beta\to z,t}\big]^{\frac{1}{2}}\Big\}$$

then

$$\log\det(\Sigma_{x^*,\beta\to z,t})=-\log\det\left(\Sigma_{\alpha,\tau,\beta\to z,t}\right)+\log\det\left(\mathsf{Id}+\big[\mathsf{Id}+\tau\Sigma_{\alpha,\tau,\beta\to z,t}\big]^{\frac{1}{2}}\right).$$

For the first sub-terms

$$\log\det\left(\Sigma_{\alpha,\tau,\beta\to z,t}\right)=\log\det\left(\Sigma_{\beta\to z,t}\right)+\det(\Sigma_{\alpha,\tau}^{-1}),$$

then we have

$$\log\det(\Sigma_x)=-3\log\det(\Sigma_{\beta\to z,t})-2\det(\Sigma_{\alpha,\tau}^{-1})+2\log\det\left(\mathsf{Id}+\big[\mathsf{Id}+\tau\Sigma_{\alpha,\tau,\beta\to z,t}\big]^{\frac{1}{2}}\right).$$

For the first term RHS of the new expression,

$$\log\det(\Sigma_{\beta\to z,t})=\log\det(\Sigma_\beta)+2\log\det(\mathsf{Id}+tT_{\beta z}).$$

We also have

$$\frac{\partial\log\det(\mathsf{Id}+tT_{\beta x})}{\partial t}\bigg|_{t=0}=\sum_{i=1}^{d}\frac{\lambda_i(T_{\beta z})}{1+t\lambda_i(T_{\beta z})}\bigg|_{t=0}=\mathsf{tr}(T_{\beta z})=\big\langle\Sigma_\beta^{-1},T_{\beta z}\big\rangle_{\Sigma_\beta}.$$

then $\frac{\partial \log\det(\Sigma_{\beta\to z,t})}{\partial t}\Big|_{t=0} = \langle 2\Sigma_\beta^{-1}, T_{\beta z}\rangle_{\Sigma_\beta}$. For the second sub-terms, observe that

$$\begin{aligned}\Big[(\mathsf{Id}+tT_{\beta z})\Sigma_\beta(\mathsf{Id}+tT_{\beta z})\Big]^{-1} &= (\mathsf{Id}+tT_{\beta z})^{-1}\Sigma_\beta^{-1}(\mathsf{Id}+tT_{\beta z})^{-1}\\ &= \big(\mathsf{Id}-tT_{\beta z}+t^2T_{\beta z}^2\big)\Sigma_\beta^{-1}\big(\mathsf{Id}-tT_{\beta z}+t^2T_{\beta z}^2\big)+O(t^3)\\ &= \Sigma_\beta^{-1}-t\big(T_{\beta z}\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}\big)+t^2\Big[T_{\beta z}^2\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}^2+T_{\beta z}\Sigma_\beta^{-1}T_{\beta z}\Big]+\mathcal{O}(t^3).\end{aligned}$$

Hence

$$\begin{aligned}\Sigma_{\alpha,\tau,\beta\to z,t} &= \Sigma_{\alpha,\tau}^{\frac12}\Sigma_\beta^{-1}\Sigma_{\alpha,\tau}^{\frac12}-t\Sigma_{\alpha,\tau}^{\frac12}\big(T_{\beta z}\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}\big)\Sigma_{\alpha,\tau}^{\frac12}\\ &= \Sigma_{\alpha,\tau,\beta}-t\Sigma_{\alpha,\tau}^{\frac12}\big(T_{\beta z}\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}\big)\Sigma_{\alpha,\tau}^{\frac12}.\end{aligned}$$

Put it into the form

$$\mathsf{Id}+\tau\Sigma_{\alpha,\tau,\beta\to z,t} = \mathsf{Id}+\tau\Sigma_{\alpha,\tau}^{\frac12}\Sigma_\beta^{-1}\Sigma_{\alpha,\tau}^{\frac12}-t\tau\Sigma_{\alpha,\tau}^{\frac12}\big(T_{\beta z}\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}\big)\Sigma_{\alpha,\tau}^{\frac12}.$$

Then by Lemma C.3

$$\Big[\mathsf{Id}+\tau\Sigma_{\alpha,\tau,\beta\to z,t}\Big]^{\frac12} = V-t\mathcal{L}_V\Big[\tau\Sigma_{\alpha,\tau}^{\frac12}\big(T_{\beta z}\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}\big)\Sigma_{\alpha,\tau}^{\frac12}\Big]+o(t).$$

where $\Big[\mathsf{Id}+\tau\Sigma_{\alpha,\tau}^{\frac12}\Sigma_\beta^{-1}\Sigma_{\alpha,\tau}^{\frac12}\Big]^{\frac12} = V$. Therefore, applying Lemma C.2, we get the derivative of the last term as

$$\begin{aligned}&-\mathsf{tr}\Big(\big[\mathsf{Id}+V\big]^{-1}\Big[\tau\Sigma_{\alpha,\tau}^{\frac12}\big(T_{\beta z}\Sigma_\beta^{-1}+\Sigma_\beta^{-1}T_{\beta z}\big)\Sigma_{\alpha,\tau}^{\frac12}\Big]\Big)\\ =&-\tau\big\langle P+Q, T_{\beta z}\big\rangle_{\Sigma_\beta},\end{aligned}$$

where

$$\begin{aligned}P &= \Sigma_\beta^{-1}\Sigma_{\alpha,\tau}^{\frac12}\big[\mathsf{Id}+V\big]^{-1}\Sigma_{\alpha,\tau}^{\frac12}\Sigma_\beta^{-1}\\ Q &= \Sigma_{\alpha,\tau}^{\frac12}\big[\mathsf{Id}+V\big]^{-1}\Sigma_{\alpha,\tau}^{\frac12}\Sigma_\beta^{-2}.\end{aligned}$$

and the derivative of log det term

$$\big\langle -6\Sigma_\beta^{-1}-2\tau(P+Q), T_{\beta z}\big\rangle_{\Sigma_\beta}.$$

Hence, in conclusion, we get the first order Wasserstein gradient according to Σ_β of the objective function

$$2\mathsf{Id}-\Big(2\Sigma_{\alpha,\tau}^{-1}+\frac12(U+\tau M)\Big)+\frac32\tau\Sigma_\beta^{-1}+\frac{\tau^2}{2}(P+Q).$$

□

D Proof for Theorem 4.5 and 4.6

D.1 Proof for Convexity

Lemma D.1. *Given two SPD matrices $A, B \in \mathbb{S}_{++}^d$ Recall that $\{\lambda_i(\Sigma)\}_{i=1}^d$ is the eigenvalues of a matrix Σ in descending order. Then we have $\mathsf{tr}(AB) \geq \lambda_d(A)\mathsf{tr}(B)$.*

Proof. Note that this is equivalent to proving that $\mathsf{tr}\Big(\big[A-\lambda_d(A)\mathsf{Id}\big]B\Big) \geq 0$. Let $A-\lambda_d(A)\mathsf{Id} = U\Lambda U^\top$ be its spectral decomposition, where $\Lambda = \mathbf{diag}(a_i)$. Let $C = [c_{ij}] = U^\top BU$. Then

$$\mathsf{tr}\Big(\big[A-\lambda_d(A)\mathsf{Id}\big]B\Big) = \mathsf{tr}\big(\Lambda U^\top BU\big) = \mathsf{tr}(\Lambda C) = \sum_i a_i c_{ii} \geq 0,$$

since both Λ and C are symmetric positive semi-definite matrices. □

Lemma D.2. *Consider the function*

$$f : \mathbb{S}^d_{++} \to \mathbb{R}$$
$$\Sigma \mapsto -\mathsf{tr}\left(\left[\Sigma^2 + 2\tau\Sigma\right]^{\frac{1}{2}}\right).$$

Assume that $\Sigma \in \mathcal{K}_{[1/\rho,\rho]}$. Then we have $f''(\Sigma) : \mathbb{S}^d_{++} \times \mathbb{S}^d_{++} \to \mathbb{R}$ is positive definite. Specifically, f is strongly convex with coefficient $\dfrac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}$.

Proof. The square root function and $x \mapsto x^2 + 2\tau x$ respectively be the analytic function in $(0,\infty)$ and $\mathbb{R}$, thus function $x \mapsto -\sqrt{x^2+2\tau x}$ is an analytic function in $(0,\infty)$. Thus, we can define the function f in the set of of SPD matrix using the Taylor expansion. Thus, the derivative of $f(\sigma)$ can be calculated as

$$\frac{df}{d\Sigma} = f'(\Sigma) = (\Sigma + \tau\mathsf{Id})\left[\Sigma^2 + 2\tau\Sigma\right]^{-\frac{1}{2}},$$

or in the operator viewpoint,

$$\frac{df}{d\Sigma} : \Sigma_1 \mapsto -\mathsf{tr}\left((\Sigma + \tau\mathsf{Id})\left[\Sigma^2 + 2\tau\Sigma\right]^{-\frac{1}{2}}\Sigma_1\right).$$

For the second derivative, we have to take the derivative respect to Σ of the function $\dfrac{df}{d\Sigma}(\Sigma_1)$. In the interval $(-1,1)$, consider the well-known Taylor expansion

$$\sqrt{1-x} = 1 - \sum_{k=0}^{\infty} \frac{2}{4^{k+1}(k+1)}\binom{2k}{k}x^{k+1}.$$

Taking the first and second derivatives of both sides, we achieve for $x \in (-1,1)$

$$\frac{1}{\sqrt{1-x}} = \sum_{k=0}^{\infty} \frac{1}{4^k}\binom{2k}{k}x^k,$$
$$\frac{1}{(1-x)\sqrt{1-x}} = \sum_{k=1}^{\infty} \frac{2}{4^k}\binom{2k}{k}kx^{k-1}.$$

Let $g(x) = \frac{x+\tau}{x^2+2\tau x}$, then $g(x) = \dfrac{1}{\sqrt{1-\frac{\tau^2}{(x+\tau)^2}}}$. Using the Taylor expansion formula, the calculation of the second derivative of f can be implemented as

$$\begin{aligned}
&\frac{f'\left(\Sigma + t\Sigma_2\right)\Sigma_1 - f'(\Sigma)\Sigma_1}{t} \\
&= \lim_{t\to 0}\frac{-1}{t}\mathsf{tr}\left(\{g(\Sigma + t\Sigma_2) - g(\Sigma)\}\Sigma_1\right) \\
&= \lim_{t\to 0}\frac{1}{t}\mathsf{tr}\left(\sum_{k=0}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}\left(\frac{1}{(\Sigma+\tau\mathsf{Id})^{2k}} - \frac{1}{(\Sigma+t\Sigma_2+\tau\mathsf{Id})^{2k}}\right)\Sigma_1\right) \\
&= \lim_{t\to 0}\frac{1}{t}\mathsf{tr}\left(\sum_{k=0}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}\left((\Sigma+\tau\mathsf{Id})^{-2k} - (\Sigma+t\Sigma_2+\tau\mathsf{Id})^{-2k}\right)\Sigma_1\right) \\
&= \lim_{t\to 0}\frac{1}{t}\mathsf{tr}\left(\sum_{k=0}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}(\Sigma+t\Sigma_2+\tau\mathsf{Id})^{-2k}\left((\Sigma+t\Sigma_2+\tau\mathsf{Id})^{2k} - (\Sigma+\tau\mathsf{Id})^{2k}\right)(\Sigma+\tau\mathsf{Id})^{-2k}\Sigma_1\right) \\
&= \lim_{t\to 0}\frac{1}{t}\mathsf{tr}\left(\sum_{k=1}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}(\Sigma+\tau\mathsf{Id})^{-2k}\left(\sum_{q=0}^{2k-1}(\Sigma+\tau\mathsf{Id})^{q}(t\Sigma_2)(\Sigma+\tau\mathsf{Id})^{2k-1-q}\right)(\Sigma+\tau\mathsf{Id})^{-2k}\Sigma_1\right) \\
&= \mathsf{tr}\left(\sum_{k=1}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}\left(\sum_{q=0}^{2k-1}(\Sigma+\tau\mathsf{Id})^{q-2k}\Sigma_2(\Sigma+\tau\mathsf{Id})^{-1-q}\Sigma_1\right)\right).
\end{aligned}$$

Now we prove that this bilinear form is positive definite by showing that for arbitrary $\Sigma_1 = \Sigma_2$ be symmetric matrices, it has a positive lower bound. Firstly, we can verify that $\lambda_d((\Sigma + \tau\mathsf{Id})^{-k}) \geq (\rho + \tau)^{-k}$ for $k < 0$ and $\Sigma \in \mathcal{K}_{[1/\rho,\rho]}$. In the formula of second derivative, let $\Sigma_1 = \Sigma_2$, according to Lemma D.1, we have for $0 \leq q \leq 2k-1$

$$\begin{aligned}\mathsf{tr}\left((\Sigma + \tau\mathsf{Id})^{q-2k}\Sigma_1(\Sigma + \tau\mathsf{Id})^{-1-q}\Sigma_1\right) &\geq (\rho+\tau)^{q-2k}\mathsf{tr}(\Sigma_1(\Sigma + \tau\mathsf{Id})^{-1-q}\Sigma_1)\\ &\geq (\rho+\tau)^{q-2k}\lambda^{-1-q}\mathsf{tr}(\Sigma_1\Sigma_1)\\ &= (\rho+\tau)^{-1-2k}\|\Sigma_1\|_F^2.\end{aligned}$$

Thus, we have the lower bound for the coefficient α-convexity of the second derivative is

$$\sum_{k=1}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}2k(\rho+\tau)^{-1-2k}.$$

On the other hand, by substituting $x = \frac{\tau^2}{(\rho+\tau)^2}$ in the Taylor expansion formula of $\frac{1}{(1-x)\sqrt{1-x}}$, we have

$$\frac{1}{\left(1-\frac{\tau^2}{(\rho+\tau)^2}\right)\sqrt{1-\frac{\tau^2}{(\rho+\tau)^2}}} = \sum_{k=1}^{\infty}\frac{2}{4^k}\binom{2k}{k}k\left(\frac{\tau^2}{(\rho+\tau)^2}\right)^{k-1} \Leftrightarrow \frac{(\rho+\tau)^3}{(\rho^2+2\tau\rho)^{3/2}} = \sum_{k=1}^{\infty}\frac{2k}{4^k}\binom{2k}{k}\left(\frac{\tau}{\rho+\tau}\right)^{2(k-1)}.$$

In other words, the lower bound can be expressed as

$$\sum_{k=1}^{\infty}\frac{\tau^{2k}}{4^k}\binom{2k}{k}2k(\rho+\tau)^{-1-2k} = \frac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}.$$

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

□

Proposition D.3. *$W^2_{2_{\mathrm{SUOT}}}(\Sigma_\alpha, \Sigma_\beta, \tau)$ is Euclidean convex with respect to Σ_β. Moreover, if $\Sigma_\beta \in \mathcal{K}_{[\frac{1}{\rho},\rho]}$, then it is $\frac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}$-strongly convex.*

Proof. From the proof of Theorem 4.4 with the same notation, we need to prove the Euclidean convexity of

$$\mathsf{tr}\big(\Sigma_\beta\big) - \mathsf{tr}\big(\Sigma_{x,\beta}\big) - \frac{\tau}{4}\log\det(\Sigma_x).$$

The Euclidean convexity of $\mathsf{tr}(\Sigma_\beta)$ is obvious. We move to the next term $-\mathsf{tr}\big(\Sigma_{x,\beta}\big)$, which is known to be reformulated as

$$\begin{aligned}&-\mathsf{tr}\Big(\frac{1}{2}\Sigma^{-1}_{\alpha,\beta,\tau} + \frac{1}{2}\Big[\Sigma^{-2}_{\alpha,\beta,\tau} + 2\tau\Sigma^{-1}_{\alpha,\beta,\tau}\Big]^{\frac{1}{2}}\Big)\\ &= -\mathsf{tr}\Big(\frac{1}{2}\Sigma^{-1}_{\alpha,\beta,\tau}\Big) - \mathsf{tr}\Big(\frac{1}{2}\Big[\Sigma^{-2}_{\alpha,\beta,\tau} + 2\tau\Sigma^{-1}_{\alpha,\beta,\tau}\Big]^{\frac{1}{2}}\Big).\end{aligned}$$

Note that $\Sigma^{-1}_{\alpha,\beta,\tau} = \Sigma^{-\frac{1}{2}}_{\alpha,\tau}\Sigma_\beta\Sigma^{-\frac{1}{2}}_{\alpha,\tau}$ is linear transformation of Σ_β, then the first term is obviously convex. For the second term, we must prove the convexity of $-\mathsf{tr}(\big[\Sigma^2 + 2\tau\Sigma\big]^{\frac{1}{2}})$. Consider the function

$$\begin{aligned}f : \mathbb{S}^d_{++} &\to \mathbb{R}\\ \Sigma &\mapsto -\mathsf{tr}\left(\big[\Sigma^2 + 2\tau\Sigma\big]^{\frac{1}{2}}\right).\end{aligned}$$

Due to Lemma D.2, f is $\frac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}$-strongly convex. Subsequently, we assert the claim. For the last term $-\log\det(\Sigma_x)$, we know that

$$-\log\det(\Sigma_x) = \log\det(\Sigma_\beta) - 2\log\det(\Sigma_{x,\beta}).$$

Similarly, the transformation $\Sigma_\beta \mapsto \Sigma_{\alpha,\beta,\tau}^{-1}$ is linear, then we can convert our problem into proving the convexity of

$$\begin{aligned}\log\det(\Sigma) - 2\log\det\left(\frac{1}{2}\left(\Sigma + \left[\Sigma^2 + 2\tau\Sigma\right]^{\frac{1}{2}}\right)\right) &= -2\log\det\left(\frac{1}{2}\Sigma^{-\frac{1}{2}}\left(\Sigma + \left[\Sigma^2 + (2\tau\Sigma)\right]^{\frac{1}{2}}\right)\right)\\ &= -2\log\det\left(\frac{1}{2}\left(\Sigma^{\frac{1}{2}} + \left[\Sigma + 2\tau\mathsf{Id}\right]^{\frac{1}{2}}\right)\right).\end{aligned}$$

It is equivalent to prove the concavity of

$$\Sigma \mapsto \log\det\left(\Sigma^{\frac{1}{2}} + \left[\Sigma + 2\tau\mathsf{Id}\right]^{\frac{1}{2}}\right) := f(\Sigma).$$

Consider two matrices Σ_1, Σ_2, by the concavity of $\log\det$ function

$$\frac{f(\Sigma_1) + f(\Sigma_2)}{2} \leq \log\det\left(\frac{\Sigma_1^{\frac{1}{2}} + \Sigma_2^{\frac{1}{2}}}{2} + \frac{(\Sigma_1 + 2\tau\mathsf{Id})^{\frac{1}{2}} + (\Sigma_2 + 2\tau\mathsf{Id})^{\frac{1}{2}}}{2}\right).$$

By the fact that two SPD matrices A, B satisfy $A^2 \succeq B^2$ then $A \succeq B$, we have inequality $\frac{\Sigma_1^{\frac{1}{2}} + \Sigma_2^{\frac{1}{2}}}{2} \preceq \left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{\frac{1}{2}}$. Indeed, we have $\left(\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}}\right)^2 \succeq 0$, which is equivalent to $\Sigma_1 + \Sigma_2 + \Sigma_1^{\frac{1}{2}}\Sigma_2^{\frac{1}{2}} + \Sigma_2^{\frac{1}{2}}\Sigma_1^{\frac{1}{2}} \preceq 2(\Sigma_1 + \Sigma_2)$. Then we have $\frac{\Sigma_1^{\frac{1}{2}} + \Sigma_2^{\frac{1}{2}}}{2} \preceq \left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{\frac{1}{2}}$. Hence, we yield

$$\begin{aligned}\frac{f(\Sigma_1) + f(\Sigma_2)}{2} &\leq \log\det\left(\frac{\Sigma_1^{\frac{1}{2}} + \Sigma_2^{\frac{1}{2}}}{2} + \frac{(\Sigma_1 + 2\tau\mathsf{Id})^{\frac{1}{2}} + (\Sigma_2 + 2\tau\mathsf{Id})^{\frac{1}{2}}}{2}\right)\\ &\leq \log\det\left(\left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{\frac{1}{2}} + \left[\frac{\Sigma_1 + \Sigma_2}{2} + 2\tau\mathsf{Id}\right]^{\frac{1}{2}}\right)\\ &= f\left(\frac{\Sigma_1 + \Sigma_2}{2}\right).\end{aligned}$$

Upon achieving the required concavity, we conclude the convexity. Hence we finish the proof. □

D.2 Proof of Theorem 4.5

Proof. First, we must show the α-smoothness of $W_{2_{\mathrm{SUOT}}}^2(\alpha,\beta,\tau)$ on Bures-manifold. Consider the functional $\mathcal{F} : \mathcal{P}_{2,ac}\left(\mathbb{R}^d \times \mathbb{R}^d\right) \times \mathcal{P}_{2,ac}\left(\mathbb{R}^d \times \mathbb{R}^d\right) \to \mathbb{R}$ defined as

$$\mathcal{F}(\Sigma_\beta, \Sigma_x) = W_2^2(\Sigma_x, \Sigma_\beta) + \tau\mathrm{KL}(\Sigma_x \| \Sigma_\alpha).$$

In fact, $W_{2_{\mathrm{SUOT}}}^2(\alpha,\beta,\tau) = \min_{\Sigma_x}\mathcal{F}(\Sigma_\beta, \Sigma_x)$, and for each Σ_β let

$$\Sigma_{\beta,x} = \arg\min_{\Sigma_x}\mathcal{F}(\Sigma_\beta, \Sigma_x).$$

Due to inequality 3.3 in Altschuler et al. (2021), we know that $\mathcal{F}$ is 1-smooth in the first variable.

Thus for $W_{2_{\mathrm{SUOT}}}^2(\alpha,\beta,\tau)$, consider two points $\Sigma_{\beta_0}, \Sigma_{\beta_1}$ in $\mathbb{S}_{++}^d$. Let $\Sigma_{\beta_s} (0 \leq s \leq 1)$ be geodesic connecting them. Then we have

$$\begin{aligned}&(1-s)\mathcal{F}(\Sigma_{\beta_0}, \Sigma_{\beta_0,x}) + s\mathcal{F}(\Sigma_{\beta_1}, \Sigma_{\beta_1,x}) - s(1-s)W_2^2(\Sigma_{\beta_0}, \Sigma_{\beta_1})\\ &\quad\leq (1-s)\mathcal{F}(\Sigma_{\beta_0}, \Sigma_{\beta_s,x}) + s\mathcal{F}(\Sigma_{\beta_1}, \Sigma_{\beta_s,x}) - s(1-s)W_2^2(\Sigma_{\beta_0}, \Sigma_{\beta_1})\\ &\quad\leq \mathcal{F}(\Sigma_{\beta_s}, \Sigma_{\beta_s,x}),\end{aligned}$$

where the first inequality happens due to $\Sigma_{\beta_0,x} = \arg\min_{\Sigma_x}\mathcal{F}(\Sigma_{\beta_0}, \Sigma_x), \Sigma_{\beta_1,x} = \arg\min_{\Sigma_x}\mathcal{F}(\Sigma_{\beta_1}, \Sigma_x)$ and the second inequality happens due to 1-smoothness of $\mathcal{F}$ according to Σ_β Then $W_{2_{\mathrm{SUOT}}}^2(\alpha,\beta,\tau)$ is 1-smooth in $\mathcal{P}_{2,ac}\left(\mathbb{R}^d\right)$

with Bures-Wasserstein metric. Thus,

$$L\Big(\Sigma_\beta\Big) = \frac{1}{n}\sum_{i=1}^{n} W^2_{2_{\mathrm{SUOT}}}(\Sigma_{\alpha_i}, \Sigma_\beta, \tau)$$

is also 1-smooth. That means if $\Sigma_\beta^{(k)}, \Sigma_\beta^{(k+1)}$ are two SPD matrices in update process, then from the 1-smoothness of the barycenter functional, we obtain the descent step

$$L\left(\Sigma_\beta^{(k+1)}\right) - L\left(\Sigma_\beta^{(k)}\right) \leq -\eta\left(1-\frac{\eta}{2}\right)\left\|\mathrm{grad}\, L\left(\Sigma_\beta^{(k)}\right)\right\|^2_{\Sigma_\beta^{(k)}}.$$

Summing up gives

$$L\left(\Sigma_\beta^{(0)}\right) - L\left(\Sigma_\beta^{(k+1)}\right) \geq \eta\left(1-\frac{\eta}{2}\right)\sum_{t=1}^{k} \| \mathrm{grad}\, L\left(\Sigma_\beta^{(k)}\right) \|^2_{\Sigma_\beta^{(k)}}.$$

It is equivalent that $\sum_{t=1}^{k} \| \mathrm{grad}\, L\left(\Sigma_\beta^{(k)}\right) \|^2_{\Sigma_\beta^{(k)}}$ has the upper bound as $L\left(\Sigma_\beta^{(0)}\right)$, it is also non-decreasing sequences then $\lim_{k\to\infty}\left\|\mathrm{grad}\, L\left(\Sigma_\beta^{(k)}\right)\right\|^2_{\Sigma_\beta^{(k)}} = 0$. It leads to that $\Sigma_\beta^{(k)}$ converges to the optimal point Σ_β^*.

Now to see the convergence rate, first we prove that if $\Sigma_\beta \in \mathcal{K}_{[1/\rho,\rho]}$, then

$$L(\Sigma_\beta) - L\left(\Sigma_\beta^*\right) \leq \frac{\rho}{8\tau^2}(\rho^2+2\tau\rho)^{\frac{3}{2}} \left\|\mathrm{grad}\, L(\Sigma_\beta)\right\|^2_{\Sigma_\beta}.$$

Indeed, from the second claim in Proposition D.3, and since $\mathcal{K}_{[1/\rho,\rho]}$ is convex with respect to Euclidean geodesics, we see that for $\Sigma \in \mathcal{K}_{[1/\rho,\rho]}$

$$\begin{aligned} L(\Sigma_\beta) - L\left(\Sigma_\beta^\star\right) &\leq \left\langle \nabla L(\Sigma_\beta), \Sigma_\beta - \Sigma_\beta^\star\right\rangle - \frac{1}{2}\frac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|^2_{\mathrm{F}} \\ &= \frac{1}{2}\left\langle\ \mathrm{grad}\, L(\Sigma_\beta), \Sigma_\beta - \Sigma_\beta^\star\right\rangle_{\Sigma_\beta} - \frac{1}{2}\frac{\tau^2}{(\rho^2+2\tau\rho)^{3/2}}\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|^2_{\mathrm{F}}, \end{aligned}$$

where the last line uses Appendix A.5 of Altschuler et al. (2021). Next we observe that by combining Cauchy-Schwarz with Young's inequality we get that for all $r > 0$,

$$\begin{aligned} \frac{1}{2}\left\langle \mathrm{grad}\, L(\Sigma_\beta), \Sigma_\beta - \Sigma_\beta^\star\right\rangle_{\Sigma_\beta} &\leq \frac{1}{2}\left\|\mathrm{grad}\, L(\Sigma_\beta)\right\|_{\Sigma_\beta}\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|_{\Sigma_\beta^{-1}} \\ &\leq \frac{r}{16}\left\|\mathrm{grad}\, L(\Sigma_\beta)\right\|^2_{\Sigma_\beta} + \frac{1}{r}\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|^2_{\Sigma_\beta^{-1}} \\ &\leq \frac{r}{16}\left\|\mathrm{grad}\, L(\Sigma_\beta)\right\|^2_{\Sigma_\beta} + \frac{\rho}{r}\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|^2_{\mathrm{F}}, \end{aligned}$$

where in the last inequality, we note that $\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|^2_{\Sigma_\beta^{-1}} = \mathsf{tr}\left((\Sigma_\beta - \Sigma_\beta^\star)^T\Sigma_\beta^{-1}(\Sigma_\beta - \Sigma_\beta^\star)\right) = \mathsf{tr}\left((\Sigma_\beta - \Sigma_\beta^\star)(\Sigma_\beta - \Sigma_\beta^\star)^T\Sigma_\beta^{-1}\right) \leq \lambda_1(\Sigma_\beta^{-1})\mathsf{tr}\left((\Sigma_\beta - \Sigma_\beta^\star)(\Sigma_\beta - \Sigma_\beta^\star)^T\right) \leq \rho\left\|\Sigma_\beta - \Sigma_\beta^\star\right\|^2_{\mathrm{F}}$. Putting $r = \frac{2}{\tau^2}\rho(\rho^2 + 2\tau\rho)^{\frac{3}{2}}$ yields the result. Then, with the assumption that all updated covariance matrices lie in $\mathcal{K}_{[1/\rho,\rho]}$ throughout the optimization trajectory, we have

$$\begin{aligned} L\left(\Sigma_\beta^{(k+1)}\right) - L\left(\Sigma_\beta^\star\right) &= L\left(\Sigma_\beta^{(k+1)}\right) - L\left(\Sigma_\beta^{(k)}\right) + L\left(\Sigma_\beta^{(k)}\right) - L\left(\Sigma_\beta^\star\right) \\ &\leq -\eta\left(1-\frac{\eta}{2}\right)\left\|\mathrm{grad}\, L\left(\Sigma_\beta^{(k)}\right)\right\|^2_{\Sigma_\beta^{(k)}} + L\left(\Sigma_\beta^{(k)}\right) - L\left(\Sigma_\beta^\star\right) \\ &\leq \left(1 - \eta(1-\frac{\eta}{2})\frac{8\tau^2}{\rho(\rho^2+2\tau\rho)^{\frac{3}{2}}}\right)\left\{L\left(\Sigma_\beta^{(k)}\right) - L\left(\Sigma_\beta^\star\right)\right\}. \end{aligned}$$

Then integrating gives

$$L\left(\Sigma_\beta^{(k)}\right) - L\left(\Sigma_\beta^\star\right) \leq \left(1 - \frac{8\tau^2\eta(1-\frac{\eta}{2})}{\rho(d\rho^2+2\tau\rho)^{\frac{3}{2}}}\right)^k \left\{L\left(\Sigma_\beta^{(0)}\right) - L\left(\Sigma_\beta^\star\right)\right\}.$$

As a consequence, we obtain the conclusion of the theorem. For the uniqueness of the optimal solution, we recall that the optimization is carried out over the set $\mathcal{K}_{[1/\rho,\rho]}$. The Euclidean and Bure-Wasserstein metrics induce different geometries on the same underlying object $\mathcal{K}_{[1/\rho,\rho]}$, but they do not alter the feasible region itself. Therefore, the optimal solution is intrinsic to this set and is independent of the chosen geometry. In particular, the optimum is identical whether the problem is viewed under the Euclidean or Bures–Wasserstein metric. □

D.3 Proof of Theorem 4.6

Proof. We rewrite the SUOT objective function as

$$\min_{\Sigma_\beta, \Sigma_{x_i(i=1,\dots,n)}} \sum_{i=1}^n W_2^2(\Sigma_{x_i}, \Sigma_\beta) + \tau \mathrm{KL}(\Sigma_{x_i} \| \Sigma_{\alpha_i}).$$

First we prove its Euclidean convexity in $n+1$ variable. Indeed, convexity of $\mathrm{KL}\left(\Sigma_{x_i} \| \Sigma_{\alpha_i}\right)$ according to Σ_{x_i} is well known. We only need to prove the convexity in two variable of $W_2^2\left(\Sigma_{x_i}, \Sigma_\beta\right)$. Consider $(\Sigma_1, \Sigma_1'), (\Sigma_2, \Sigma_2') \in \mathbb{S}_{++}(\mathbb{R}^d) \times \mathbb{S}_{++}(\mathbb{R}^d)$. Taking samples $(X_1, X_1'), (X_2, X_2')$ such that

$$\begin{aligned} X_1 &\sim \mathcal{N}(0, \Sigma_1), X_1' \sim \mathcal{N}(0, \Sigma_1') \\ X_2 &\sim \mathcal{N}(0, \Sigma_2), X_2' \sim \mathcal{N}(0, \Sigma_2'), \end{aligned}$$

where (X_1, X_1') and (X_2, X_2') satisfy the coupling minimizing; (X_1, X_1') and (X_2, X_2') are independent. Then we have

$$\begin{aligned} & tW_2^2(\Sigma_1, \Sigma_1') + (1-t)W_2^2(\Sigma_2, \Sigma_2') \\ &= \mathbb{E}\left[t|X_1 - X_1'|^2 + (1-t)\|X_2 - X_2^2\|\right] \\ &= \mathbb{E}\left[\left(\sqrt{t}(X_1 - X_1') + \sqrt{(1-t)}(X_2 - X_2')\right)^2\right] \text{ (due to the independency).} \\ &\geq W_2^2(t\Sigma_1 + (1-t)\Sigma_2, t\Sigma_1' + (1-t)\Sigma_2'). \end{aligned}$$

Then it is Euclidean convex by the definition. Next, we see that the barycenter objective function F decreases during the iterations (suppose that all the updated matrices lie on a compact set $\mathcal{K}_{[\frac{1}{\rho},\rho]}$). Indeed, the scheme of Hybrid Bures-Wassertein algorithm could be seen as a Block Coordinate Descent on SPD Manifolds Gutman and Ho-Nguyen (2023); Peng and Vidal (2023). At each iteration, we fix Σ_β to update $\{\Sigma_{x_i}\}_{i=1}^n$, then fix $\{\Sigma_{x_i}\}_{i=1}^n$ to update Σ_β. Both these updates should be done on SPD manifolds. In detail, at iteration k-th,

- If we fix $\Sigma_\beta = \Sigma_\beta^{(k)}$, then $\Sigma_{x_i}^{(k)}$s which are minimizer of $\sum_{i=1}^n W_2^2(\Sigma_{x_i}, \Sigma_\beta^{(k)}) + \tau \mathrm{KL}(\Sigma_{x_i} \| \Sigma_{\alpha_i})$ must satisfy the forms in Theorem 4.2 of our paper.
- If we fix $\Sigma_{x_i} = \Sigma_{x_i}^{(k)} \quad \forall i = 1, \dots, n$, we need to minimize $\sum_{i=1}^n W_2^2(\Sigma_{x_i}, \Sigma_\beta)$ according to Σ_β. The updates from Chewi et al. (2020) will give the minimizer for this problem as $\Sigma_\beta^{(k+1)}$ theoretically supported by Theorem 7 Chewi et al. (2020).

Now, the objective function decreases with corresponding solutions sequence $\left\{\Sigma_\beta^{(k)}\right\}_{k=1}^\infty$. Moreover, the sequences $\left\{\Sigma_\beta^{(k)}\right\}_{k=1}^\infty$ should not tend to infinity; otherwise, the objective function will also tend to infinity. Thus, we can extract a subsequence $\left\{\Sigma_\beta^{(k_n)}\right\}_{k=1}^\infty$ which converges to a limit $\tilde{\Sigma}_\beta$. $\Sigma_{x_i}^{(k_n)}$ will also converge to the corresponding $\tilde{\Sigma}_{x_i}$.

Now we leverage the fact that the derivative of L is continuous. Indeed, the function L is defined as the sum of Wasserstein distances and KL divergence. Both of them have closed form of expression as analytic functions (i.e. functions that we can take the derivative as many times we want, and the derivatives are helpful for Taylor's expansion). For Wasserstein distance: $W_2^2(\Sigma_1, \Sigma_2) = \mathrm{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}))$, for KL divergence: $KL(\Sigma_1||\Sigma_2) = \log(\frac{\det\Sigma_2}{\det\Sigma_1}) - d + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)$, where d is the dimension of the matrices Σ_2 and Σ_1. As a result, the continuity of the derivative of the loss function is automatically guaranteed.

Applying this for non-invariant time function $\partial L(\cdot)$ with the sequence $\{(\Sigma_\beta^{(k_n)}, \Sigma_{x_i}^{(k_n)})\}$ satisfying $\frac{\partial L}{\partial(\Sigma_\beta^{(k_n)}, \Sigma_{x_i}^{(k_n)})} = 0$ for all n and $\{(\Sigma_\beta^{(k_n)}, \Sigma_{x_i}^{(k_n)})\}$ converges to $(\tilde{\Sigma}_\beta, \tilde{\Sigma}_{x_i})$, it yields $\tilde{\Sigma}_\beta$ and $\tilde{\Sigma}_{x_i}$ satisfy that $\frac{\partial L}{\partial \tilde{\Sigma}_\beta} = 0$, $\frac{\partial L}{\partial \tilde{\Sigma}_{x_i}} = 0$, then they will be the stationary points. Following Appendix D.1, $W_{2\,\mathrm{SUOT}}^2$ is strictly Euclidean convex then these points are unique optimal points. Moreover, $\left\{L\left(\Sigma_\beta^{(k_n)}, \Sigma_{x_i}^{(k_n)}\right)\right\}_{k=1}^{\infty}$ and $\left\{L\left(\Sigma_\beta^{(k)}, \Sigma_{x_i}^{(k)}\right)\right\}_{k=1}^{\infty}$ are decreasing sequences then $\{L(\Sigma_\beta^{(k)}, \Sigma_{x_i}^{(k)})\}_{k=1}^{\infty}$ decreases to this optimal value, leading to $\{(\Sigma_\beta^{(k)}, \Sigma_{x_i}^{(k)})\}$ converges to the optimal points $(\tilde{\Sigma}_\beta, \tilde{\Sigma}_{x_i})$. □

E Additional Experiments

For the SGD study, we implemented *Exact Geodesic Stochastic Gradient Descent*; *Hybrid Stochastic Gradient Descent*; *Auto-Geodesic Stochastic Gradient Descent*, *Auto-Geodesic Stochastic Gradient Descent with Momentum*. At each iteration, we account for one covariance matrix from the samples. We run for 100 iterations and observe the behaviour of loss function.

The figure demonstrates that for SGD methods, the exact algorithm converges faster than the hybrid approach.

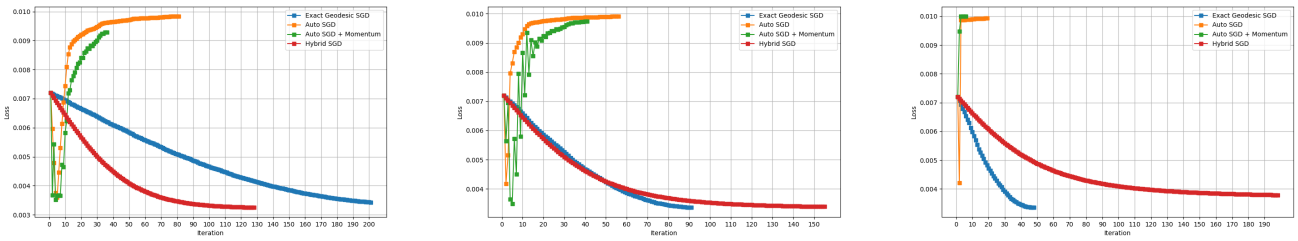

Figure 4: Loss on $L(\Sigma_\beta)$ through iterations with different step sizes (0.1; 0.25 and 0.5). red: Stochastic Hybrid Gradient Descent, blue: Stochastic Exact Geodesic Gradient Descent, orange: Stochastic Auto-Geodesic Gradient Descent, green: Stochastic Auto-Geodesic Gradient Descent with Momentum.

Next, we present the proposition regarding the relationship between Σ_x and the marginal matrices Σ_α and Σ_β.

Proposition E.1. *Considering the optimizer Σ_x of Theorem 4.2, the following convergences hold*

1. $\Sigma_x \overset{\tau\to 0}{\to} \Sigma_\beta$,
2. $\|\Sigma_x - \Sigma_\alpha\|_F \overset{\tau\to\infty}{\to} 0$.

Proof. The first part is trivial. When τ goes to zero, we could easily compute that $\Sigma_{\alpha,\tau}^{-1}$ goes to Id, then $\Sigma_{\alpha,\tau,\beta}^{-1}$ goes to Σ and Σ_x goes to Σ_β. For the second part, we note that

$$\left\|\Sigma_{\alpha,\tau}^{-1} - \frac{2}{\tau}\Sigma_\alpha\right\|_F \overset{\tau\to\infty}{\to} 0.$$

Actually, we have

$$\Sigma_{\alpha,\tau}^{-1} = \left(\mathsf{Id} + \frac{\tau}{2}\Sigma_\alpha^{-1}\right)^{-1} = \frac{2}{\tau}\Sigma_\alpha\left(\frac{2}{\tau}\Sigma_\alpha + \mathsf{Id}\right)^{-1},$$

so

$$\left\|\Sigma_{\alpha,\tau}^{-1}-\frac{2}{\tau}\Sigma_\alpha\right\|_F=\left\|\frac{2}{\tau}\Sigma_\alpha\left[\left(\frac{2}{\tau}\Sigma_\alpha+\mathsf{Id}\right)^{-1}-\mathsf{Id}\right]\right\|_F$$
$$\leq\frac{2}{\tau}\left\|\Sigma_\alpha\right\|_F\left\|\left(\frac{2}{\tau}\Sigma_\alpha+\mathsf{Id}\right)^{-1}-\mathsf{Id}\right\|_F.$$

When τ goes to infinity, $\frac{2}{\tau}\Sigma_\alpha+\mathsf{Id}$ goes to Id then limit of RHS is 0. It follows that our comment is true. Then we have

$$\left\|\frac{\tau}{2}\Sigma_{\alpha,\tau,\beta}^{-1}-\Sigma_\beta^{\frac{1}{2}}\Sigma_\alpha\Sigma_\beta^{\frac{1}{2}}\right\|_F\overset{\tau\to\infty}{\to}0$$

due to $\Sigma_{\alpha,\tau,\beta}^{-1}=\Sigma_\beta^{\frac{1}{2}}\Sigma_{\alpha,\tau}^{-1}\Sigma_\beta^{\frac{1}{2}}$. Hence, we have

$$\left\|\Sigma_x-\Sigma_\alpha\right\|_F\leq\left\|\Sigma_\beta^{-\frac{1}{2}}\right\|_F^2\left(\left\|\frac{\tau}{2}\Sigma_{\alpha,\tau,\beta}^{-1}-\Sigma_\beta^{\frac{1}{2}}\Sigma_\alpha\Sigma_\beta^{\frac{1}{2}}\right\|_F+\mathcal{O}\left(\tau^{-\frac{3}{2}}\right)\right).$$

It verifies our conclusion. □

The first figure visualizes our barycenter for various values of τ. The experimental settings are the same as those described in Section 5 but with the variable τ ranging from 0.005 to 100. The figure shows that when τ is large enough, the SUOT-based Barycenter produced by our method resembles the normal Wasserstein Barycenter (as shown in the last subfigure), thereby verifying our theoretical analysis in Proposition E.1.

To quantify this, we calculate the distances between the standard Wasserstein Barycenter learned by the method of Chewi et al. (2020) and our barycenters using the OT distance between Gaussians, as described in Proposition 3 of Appendix A.1. The second figure illustrates the distances under the effects of varying τ.

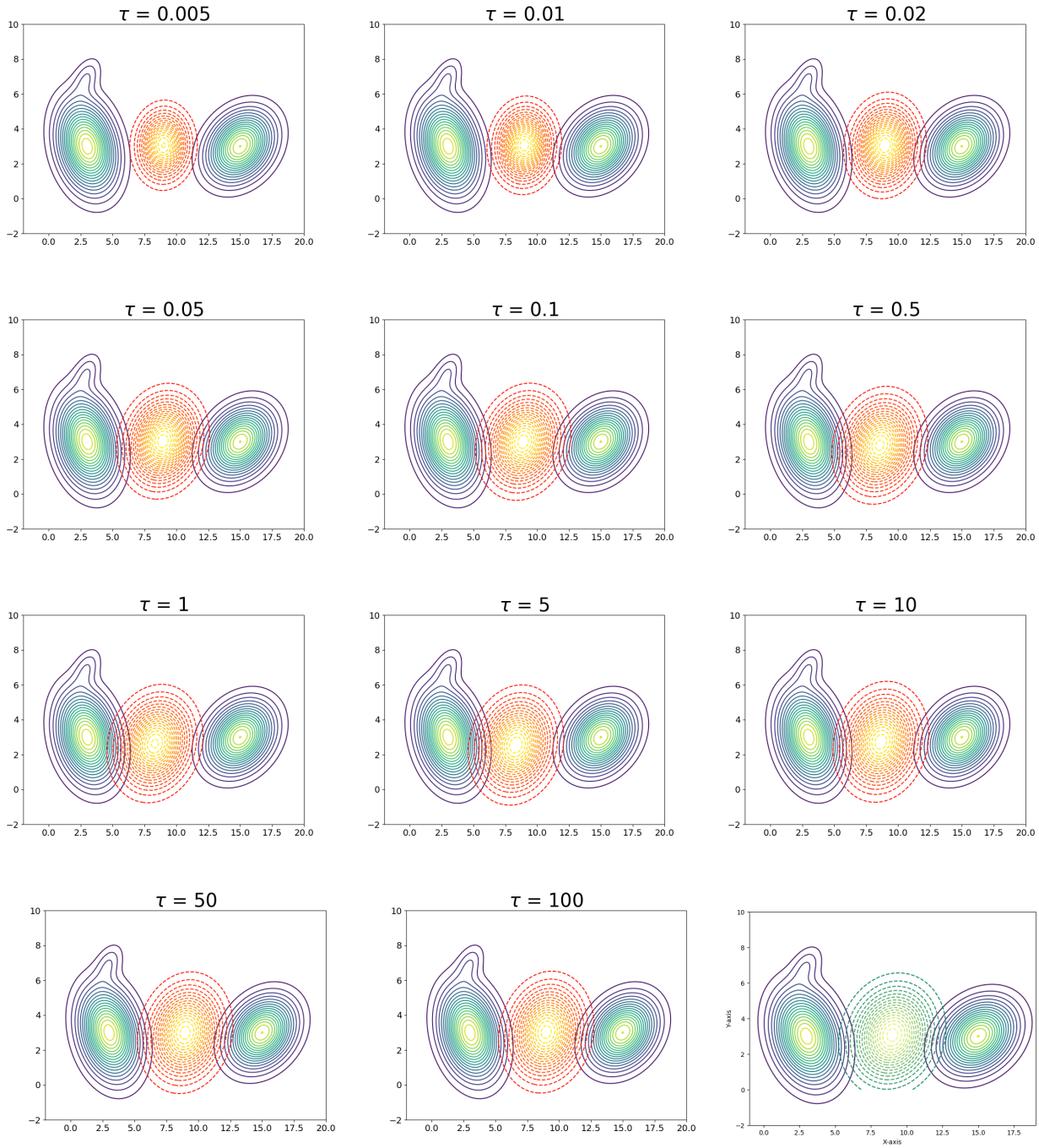

Figure 5: Ablation study of the dependence between the SUOT-based Barycenter and parameter τ. From top to bottom, left to right, we calculate the barycenter with values of τ as $0.005, 0.01, 0.02, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100$. The bottom right corner subfigure is normal Wasserstein Barycenter in Figure 2 of the main manuscript

Figure 6: Bures-Wasserstein Distance from SUOT-based Barycenters to the barycenter learnt by Chewi et al. (2020)