# Projection Robust Optimal Transport Between Unbalanced Distributions

**Anonymous Author**
Anonymous Institution

## Abstract

We propose a novel unbalanced optimal transport (UOT) formulation that has the potential to alleviate the curse of dimensionality. The key idea is borrowed from recent developments of the projection robust Wasserstein distance, which projects the sampled data onto lower-dimensional subspace and computes the Wasserstein distance between the projected data. Using the same idea, we propose the projection robust UOT, which is a max-min problem over Stiefel manifold. We propose two algorithms for solving this problem and analyze their complexity for obtaining an $\epsilon$-stationary point. Numerical experiments on both synthetic and real datasets are conducted to demonstrate the advantages of our new UOT formulation in high-dimensional cases.

## 1 Introduction

Optimal transport (OT) can be used to measure the distance of two probability distributions, which is called the Wasserstein distance. Recently, OT has drawn great attention due to its applications in model machine learning, such as the stationary Markov chain (O'Connor et al., 2022), representation learning (Ozair et al., 2019), manifold alignment (Demetci et al., 2022), graph attention model (Salimans et al., 2018) and domain adaptation (Damodaran et al., 2018). One issue about the OT is that it requires that the total mass of the two input measures should be equal. Thus the applicability of OT is greatly limited in scenarios where the measures have different masses or when they contain outliers. Many tasks have been solved efficiently by UOT, for example, domain adaptation (Fatras et al., 2021), large scale imaging regularization (Lee et al., 2019) and image translation (Zhan et al., 2021); or researchers have mentioned to develop a UOT architecture in their future work to handle the outliers, like multi-omics alignment (Demetci et al., 2022). These tasks are widely across computational biology, computational imaging, deep learning, and machine learning which reveals the significance of the UOT problem.

Motivated by the limitations of OT, the unbalanced optimal transport (UOT) has recently been proposed (Benamou, 2003) for computing the optimal transport between two measures of possibly different masses. It is a relaxation of the Kantorovich formulation which places penalty functions on the marginal distributions based on some given divergences (Liero et al., 2018).

A significant barrier to the direct application of OT or UOT lies in the estimation of high-dimensional Wasserstein distances. It is well known that the sample complexity of approximating Wasserstein distances between measures using only samples can grow exponentially in dimension (Dudley, 1969). Practitioners have long been aware of this issue of the curse of dimensionality in OT. There have been many attempts to mitigate the curse of dimensionality. Among them, one approach that was proposed recently is the sliced approximation of OT (Rabin et al., 2011). The sliced approximation of OT suggests projecting the sampled data to a given line and uses the Wasserstein distance between the projected data as an approximation to the Wasserstein distance between the two original distributions. In more recent papers, Niles-Weed and Rigollet (2019) and Paty and Cuturi (2019) propose to project the sampled data onto a lower-dimensional subspace and then compute the Wasserstein distance between the projected data. This model is called the projection robust Wasserstein (PRW) distance. As proved in Niles-Weed and Rigollet (2019), PRW indeed reduces the sample complexity and resolves the issue of the curse of dimensionality for a particular model named the spiked transport model.

Because it is reasonable to believe that the UOT inherits from OT the issue of the curse of dimensionality, encouraged by the success of PRW in OT, we propose to incorporate the projection robust idea to UOT, and this leads to our projection robust Wasserstein distance on unbalanced optimal transport (PRUOT) model. Our contributions in this paper can be summarized as follows.

- We propose the PRUOT model that can potentially mitigate the curse of dimensionality of UOT.
- We propose two algorithms for solving the PRUOT model: Riemannian Gradient Ascent with Sinkhorn (RGAS) and Riemannian Block Coordinate Descent (RBCD). We analyze the complexity of RGAS for obtaining an $\epsilon$-stationary point. We also discuss the difficulty of analyzing the convergence of RBCD.
- We conduct numerical experiments to evaluate our algorithms on both synthetic and real datasets. The results show that the proposed model and algorithms can indeed mitigate the curse of dimensionality of UOT, and the proposed model is more robust than the original UOT model.

The rest of this paper is organized as follows. In Section 2, we briefly review the projection robust OT problem and some necessary backgrounds of Riemannian optimization. In Section 3, we introduce our PRUOT and discuss its optimality condition. In Section 4, we analyze the complexity of RGAS for obtaining an $\epsilon$-stationary point of PRUOT. We also discuss the difficulty of analyzing the convergence of RBCD. In Section 5, we present numerical results on both synthetic and real datasets to demonstrate the advantages of PRUOT compared to UOT. We show that the PRUOT model indeed has the potential to mitigate the curse of dimensionality and is more robust to noise. Finally, we draw some conclusions in Section 6.

## 2 Projection Robust Optimal Transport

In this section, we review projection robust OT and basics of Riemannian optimization.

### 2.1 Projection Robust OT

Let $\{x_1, x_2 \ldots, x_n\}$ with $x_j \in \mathbb{R}^d, j = 1, \ldots, n$ and $\{y_1, y_2 \ldots, y_n\}$ with $y_j \in \mathbb{R}^d, j = 1, \ldots, n$ denote two sets of data points. We use $\Delta^n$ to denote the probability simplex set, i.e., $\Delta^n = \{x \in \mathbb{R}^n \mid \sum_i x_i = 1, x \geq 0\}$. We use $\mathbf{r} = (r_1, r_2, \ldots, r_n) \in \Delta^n$ and $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in \Delta^n$ to denote the two weight vectors. We denote $\sum_{i=1}^n r_i = \alpha$, $\sum_{j=1}^n c_i = \beta$. We define discrete probability measures $\mu_n := \sum_{i=1}^n r_i \delta_{x_i}$ and $\nu_n := \sum_{j=1}^n c_j \delta_{y_j}$. Here $\delta_x$ denotes the Dirac delta function at $x$. The Wasserstein distance between $\mu_n$ and $\nu_n$ is defined as:

$$\mathcal{W}^2(\mu_b, \nu_n) := \min_{\pi \in \Pi(\mu_n, \nu_n)} \langle C, \pi \rangle, \tag{1}$$

where the transportation polytope $\Pi(\mu_n, \nu_n) := \left\{\pi \in \mathbb{R}_+^{n \times n} | \pi \mathbf{1} = r, \pi^{\mathrm{T}} \mathbf{1} = c\right\}$, and $\mathbf{1}$ denotes the $n$-dimensional all-one vector. Throughout the paper, $C \in \mathbb{R}^{n \times n}$ denotes the matrix whose $(i, j)$-th component is $C_{ij} = \|x_i - y_j\|^2$.

The projection robust Wasserstein distance is defined as follows (see Paty and Cuturi (2019)):

$$\mathcal{P}_k^2(\mu_b, \nu_n) := \max_{U \in \mathcal{M}} \min_{\pi \in \Pi(\mu_n, \nu_n)} f(\pi, U), \tag{2}$$

where $U \in \mathbb{R}^{d \times k}$ denotes an orthonormal basis of the $k$-dimensional subspace, $f(\pi, U) := \sum_{i,j=1}^n \pi_{ij} \|U^T x_i - U^T y_j\|^2$, $\mathcal{M} := \{U \in \mathbb{R}^{d \times k} \mid U^\top U = I_k\}$ is the Stiefel manifold. Note that $\|U^T x_i - U^T y_j\|^2$ is the distance between the projected $x_i$ and $y_j$. It is very challenging to solve (2). In practice, it is suggested to solve the following entropy regularized version of (2) (see Paty and Cuturi (2019)):

$$\max_{U \in \mathcal{M}} \min_{\pi \in \Pi(\mu_n, \nu_n)} \sum_{i,j=1}^n \pi_{ij} \|U^T x_i - U^T y_j\|^2 - \eta H(\pi), \tag{3}$$

where $\eta > 0$ is a penalty parameter, $H(\pi) := -\langle \pi, \log \pi - \mathbf{1}\mathbf{1}^T \rangle$ is the entropy function. This is called the projection robust OT (PROT) problem. Recently, Lin et al. (2020) proposed a Riemannian gradient method to compute the (2). More specifically, they proposed the RGAS (Riemannian Gradient Ascent with Sinkhorn Iteration) algorithm for solving (3). RGAS requires to solve an entropy-regularized OT problem in every iteration, which can be time consuming. More recently, Huang et al. (2021) proposed the RBCD (Riemannian Block Coordinate Descent) algorithm which has a much lower per-iteration complexity compared to RGAS (Lin et al., 2020).

### 2.2 Basics of Riemannian Optimization

Now we review some important concepts in Riemannian optimization.

**Definition 2.1** *(Absil et al., 2009) The tangent space of $\mathcal{M}$ at $U \in \mathcal{M}$ is defined as*

$$\begin{aligned} \mathrm{T}_U\mathcal{M} = &\{\eta'(0) : \gamma \textit{ is a smooth curve with} \\ &\gamma(0) = U, \eta([-w, w] \subset \mathcal{M}, w > 0)\}. \end{aligned} \tag{4}$$

*The tangent bundle is defined as*

$$\mathrm{T}\mathcal{M} := \{(U, \xi) : U \in \mathcal{M}, \xi \in \mathrm{T}_U\mathcal{M}\}.$$

For the Stiefel manifold $\mathcal{M}$, its tangent space at point $U \in \mathcal{M}$ can be written as:

$$\mathrm{T}_U\mathcal{M} := \{\xi \in \mathbb{R}^{d \times k} : \xi^T U + U^T \xi = 0\}.$$

We consider $\mathcal{M}$ with Riemannian metric inherited from the Euclidean inner product. Therefore, for any $\xi, \eta \in \mathrm{T}_U\mathcal{M}$, we have $\langle \xi, \eta \rangle_U = \mathrm{Tr}(\xi^T \eta)$. Moreover, in this case, the Riemannian gradient of $f$ is the orthogonal projection of the Euclidean gradient onto the tangent space; i.e.,

$$\mathrm{grad} f(U) = \mathrm{Proj}_{\mathrm{T}_U\mathcal{M}} \nabla f(U).$$

**Definition 2.2** *(Absil et al., 2009) A retraction on $\mathcal{M}$ is a smooth mapping $\mathrm{Retr}(\cdot)$ from the tangent bundle $\mathrm{T}\mathcal{M}$ onto $\mathcal{M}$ satisfying the following two conditions:*

*(1) $\mathrm{Retr}_U(0) = U, \forall U \in \mathcal{M}$, where $0$ denotes the zero element of $\mathrm{T}_U\mathcal{M}$;*

*(2) For any $U \in \mathcal{M}$, it holds that*

$$\lim_{\xi \in \mathrm{T}_U\mathcal{M}, \xi \to 0} \frac{\|\mathrm{Retr}_U(\xi) - (U+\xi)\|_F}{\|\xi\|_F} = 0.$$

# 3 Projection Robust UOT

In this section, we introduce our projection robust UOT model. UOT is proposed to measure the distance between two measures of possibly different masses. It is a relaxation which replaces the marginal linear constraints with divergence penalty functions. Computing the Wasserstein distance between two measures $\mu$ and $\nu$ under UOT setting is equivalent to

$$\min_{\pi \in \mathbb{R}_+^{n\times n}} f(\pi)$$

where

$$f(\pi) := \langle C, \pi\rangle + \tau \mathbf{KL}(\pi\mathbf{1}\|\mathbf{r}) + \tau\mathbf{KL}(\pi^T\mathbf{1}\|\mathbf{c}). \quad (5)$$

Here, $\pi$ is a transportation plan and $\tau > 0$ is a given regularization parameter. The $\mathbf{KL}$-divergence between vector $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{KL}(\mathbf{x}\|\mathbf{y}) := \sum_{i=1}^{n} \mathbf{x}_i \log\left(\frac{\mathbf{x}_i}{\mathbf{y}_i}\right) - \mathbf{x}_i + \mathbf{y}_i.$$

When $\mathbf{r}^T\mathbf{1} = \mathbf{c}^T\mathbf{1}$ and $\tau \to \infty$, the UOT problem (5) reduces to the standard OT problem.

Because UOT ia a relaxation of OT, it is reasonable to expect that the UOT will inherit from OT the curse of dimensionality issue. To mitigate the curse of dimensionality, we propose the PRUOT model which can be formulated as follows:

$$\max_{U\in\mathcal{M}} \min_{\pi\in\mathbb{R}_+^{n\times n}} f(\pi, U) := \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{i,j}\|U^T x_i - U^T y_j\|^2 + \tau\mathbf{KL}(\pi\mathbf{1}_n\|\mathbf{r}) + \tau\mathbf{KL}(\pi^T\mathbf{1}_n\|\mathbf{c}). \quad (6)$$

Here similar to projection robust OT (3), the matrix $U$ denotes an orthonormal basis of the subspace which the sampled data are projected onto. Again, we can add an entropy regularization term to make the problem easier to solve. This leads to the following problem:

$$\max_{U\in\mathcal{M}} \min_{\pi\in\mathbb{R}_+^{n\times n}} g(\pi, U) := \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{i,j}\|U^T x_i - U^T y_j\|^2 + \tau\mathbf{KL}(\pi\mathbf{1}\|\mathbf{r}) + \tau\mathbf{KL}(\pi^T\mathbf{1}\|\mathbf{c}) - \eta H(\pi). \quad (7)$$

We refer to (7) as the entropy regularized projection robust UOT. We now define the $\epsilon$-optimal solution to UOT and $\epsilon$-stationary point for PRUOT.

**Definition 3.1** *For any $\epsilon > 0$, we call $\pi$ and $\epsilon$-approximation transportation plan if the following holds*

$$\langle C, \pi\rangle + \tau\mathbf{KL}(\pi\mathbf{1}\|\mathbf{r}) + \tau\mathbf{KL}(\pi^T\mathbf{1}\|\mathbf{c}) \le \langle C, \hat{\pi}\rangle + \tau\mathbf{KL}(\hat{\pi}\mathbf{1}\|\mathbf{r}) + \tau\mathbf{KL}(\hat{\pi}^T\mathbf{1}\|\mathbf{c}) + \epsilon,$$

*where $\hat{\pi}$ is an optimal solution to the UOT problem* (5)*.*

**Definition 3.2** *We call $(\hat{\pi}, \hat{U}) \in \mathbb{R}_+^{n\times n} \times \mathcal{M}$ an $(\epsilon_1, \epsilon_2)$-stationary point of the PRUOT problem* (6)*, if the following two inequalities hold:*

$$\|\mathrm{grad}_U f(\hat{\pi}, \hat{U})\|_F \le \epsilon_1,$$

$$f(\hat{\pi}, \hat{U}) - \min_{\pi\in\mathbb{R}_+^{n\times n}} f(\pi, \hat{U}) \le \epsilon_2.$$

## 3.1 Quantities

Table 1: Quantities mentioned in the lemmas.

| | |
|---|---|
| $\gamma$ | $\frac{1}{K\left((8L_1^2+16L_2)\|C\|_\infty + 16\eta^{-1}L_1^2\|C\|_\infty^2\right)}$ |
| $S$ | $\frac{1}{2}(\alpha+\beta) + \frac{1}{2} + \frac{1}{4\log(n)}$ |
| $D$ | $\left(\frac{\alpha+\beta}{2}\right)\left[\log\left(\frac{\alpha+\beta}{2}\right) + 2\log(n) - 1\right] + \log(n) + \frac{5}{2}$ |
| $J$ | $\max\left\{S + D, 2\epsilon_2, \frac{4\epsilon_2\log(n)}{\tau}, \frac{4\epsilon_2(\alpha+\beta)\log(n)}{\tau}\right\}$ |
| R | $\max\{\|\log(\mathbf{r})\|_\infty, \|\log(\mathbf{c})\|_\infty\} + \max\left\{\log(n), \frac{1}{\eta}\|C\|_\infty - \log(n)\right\}$ |

# 4 Riemannian Optimization Algorithms for Solving PRUOT

We discuss two algorithms for solving PRUOT (6): Riemannian Gradient Ascent with Sinkhorn's Iteration (RGAS) (Lin et al., 2020) and Riemannian Block Coordinate Descent (RBCD) (Huang et al., 2021).

## 4.1 RGAS

The RGAS was first proposed in Lin et al. (2020) for solving the entropy-regularized PROT (3). The RGAS for solving (3) can be described as follows. First, we denote

$$f_\eta(U) := \min_{\pi\in\Pi(\mu_n,\nu_n)} \sum_{i,j=1}^{n} \pi_{ij}\|U^T x_i - U^T y_j\|^2 - \eta H(\pi), \quad (8)$$

then the entropy-regularized PROT (3) is equivalent to

$$\max_{U\in\mathcal{M}} f_\eta(U).$$

This can be solved by Riemannian gradient ascent because $f_\eta$ is smooth with respect to $U$. However, to compute the Riemannian gradient of $f_\eta$ for a fixed $U$, one needs to solve the optimization problem in (8) using the Sinkhorn's algorithm, and this leads to the RGAS algorithm. A typical iteration of RGAS algorithm for solving entropy-regularized PROT (3) is given below.

(i) run Sinkhorn's algorithm to solve (8) with $U = U^t$ to obtain $\pi^*(U^t)$

(ii) $U^{t+1} := \mathrm{Retr}_{U^t}(\tau_t \mathrm{grad} f_\eta(U^t))$.

We can apply the same idea to solve the entropy-regularized PRUOT (7). For UOT problem, we can formulate the entropic regularized UOT by Fenchel-Legendra dual (Pham et al., 2020), which is given by

$$\max_{u,v\in\mathbb{R}^n} -F^*(-u) - G^*(-v) - \eta \sum_{i,j} \exp(\frac{u_i + v_i - C_{ij}}{\eta}),$$

where the functions $F^*(\bullet)$ and $G^*(\bullet)$ take the following forms:

$$F^*(u) = \max_{z\in\mathbb{R}^n} z^T u - \tau \mathbf{KL}(z\|\mathbf{r}) = \tau\langle e^{u/\tau}, \mathbf{r}\rangle - \mathbf{r}^T \mathbf{1}_n$$

$$G^*(u) = \max_{x\in\mathbb{R}^n} x^T v - \tau \mathbf{KL}(x\|\mathbf{c}) = \tau\langle e^{v/\tau}, \mathbf{c}\rangle - \mathbf{c}^T \mathbf{1}_n$$

With this formulation, we can solve UOT problem by Sinhorn iterations at each inner loop and run Riemannian gradient ascent at the outer loop. Therefore, we compute Eq. (7) by Algorithm 2 and the inner maximization problem is solved as shown in Algorithm 1.

**Algorithm 1** UOT-SINKHORN($C$, $\epsilon$)

**Input:** $h = 0$, $u^0 = v^0 = 0$ and $\eta = \frac{\epsilon}{J}$.
**while** $h \le (\frac{\tau J}{\epsilon} + 1)[\log(8\eta R) + \log(\tau(\tau+1)) + 3\log(\frac{J}{\epsilon})]$ **do**
  $a^h = \pi(u^h, v^h)\mathbf{1}_n$
  $b^h = \pi(u^h, v^h)^T \mathbf{1}_n$
  **if** $h$ is even **then**
    $u^{h+1} = [\frac{u^h}{\eta} + \log(\mathbf{r}) - \log(a^h)]\frac{\eta\tau}{\eta+\tau}$
    $v^{h+1} = v^h$
  **else**
    $v^{h+1} = [\frac{v^h}{\eta} + \log(\mathbf{c}) - \log(b^h)]\frac{\eta\tau}{\eta+\tau}$
    $u^{h+1} = u^h$
  **end if**
    $h = h + 1$
  **end while**
  **return** $\pi(u^h, v^h)$.

**Theorem 4.1** *Letting $\{(U_t, \pi_t)\}_{t\ge1}$ be the iterates generated by Algorithm 2, the number of iterations $t$ required for the outer loop to reach $\|\mathrm{grad}_{U_t} f(\pi_t, U_t)\|_F \le \epsilon_1$ satisfies that $t \le \frac{100\Delta_f}{\gamma(25\epsilon_1{}^2 - 3\epsilon_2{}^2)}$ where $\gamma$ is defined in Table 1. For the inner loop which updates $\pi$ by Sinkhorn algorithm, Pham et al. (2020) gave the results on its convergence. For inner iteration $k = (\frac{\tau J}{\epsilon} + 1)[\log(8\eta R) + \log(\tau(\tau+1)) + 3\log(\frac{J}{\epsilon})]$ and $\eta = \frac{\epsilon}{J}$, the update $\pi^k$ satisfies that $f(\hat{\pi}, \hat{U}) - f(\pi_k, \hat{U}) \le \epsilon_2$.*

**Algorithm 2** RGAS-UOT($C$, $\epsilon$)

**Input:** $\{(x_i, r_i)\}_{i\in[n]}$, $\{(y_i, c_i)\}_{j\in[n]}$, $k = \tilde{O}(1)$, $U_1 \in \mathcal{M}$, $\epsilon$, and $\alpha \in (0, 1)$.
**for** $t = 1, 2, \ldots$ **do**
  Compute $C^t$
  Compute $\pi_{t+1} \leftarrow$ UOT-SINKHORN($C^t, \epsilon$)
  Compute $\xi_{t+1} \leftarrow P_{T_{U_t}\mathrm{St}}(V_{\pi_{t+1}} U_t)$
  Compute $U_{t+1} \leftarrow \mathrm{Retr}_{U_t}(\gamma\xi_{t+1})$
**end for**

### 4.2 RBCD on UOT

RBCD solves the entropic regularized OT based on a new reformulation of the original problem. It formulates the problem as a minimization problem among the Sinkhorm updating vector $u$, $v$, and the manifold $U$, and updates the three block variables $(u, v, U)$ consecutively in each iteration. The pseudocode of the algorithm is shown in Algorithm 3. One thing different between RBCD and RBCD-UOT is that the projection of $\pi$ on the polytope $\Pi(\mu, \nu)$ is removed because we no longer consider the linear constraints for PRUOT model.

Compared to RGAS-UOT, RBCD-UOT also updates the inner minimization problem by the Sinkhorn algorithm but only once in each iteration instead of multiple times. Therefore, RBCD-UOT is more efficient in practice which will be shown in the late numerical experiments.

**Algorithm 3** RBCD-UOT($C$, $\epsilon$)

**Input:** $U^0 \in \mathcal{M}$, $u^0, v^0 \in \mathbb{R}^n$, and accuracy tolerance $\epsilon_1, \epsilon_2 > 0$.
**for** $t = 1, 2, \ldots$ **do**
  $\mathbf{a}^t = \pi(u^t, v^t, U^t)\mathbf{1}_n$
  $u^{t+1} = [\frac{u^t}{\eta} + \log(\mathbf{r}) - \log(\mathbf{a}^t)]\frac{\eta\tau}{\eta+\tau}$
  $\mathbf{b}^t = \pi(u^{t+1}, v^t, U^t)^T \mathbf{1}_n$
  $v^{t+1} = [\frac{v^t}{\eta} + \log(\mathbf{c}) - \log(\mathbf{b}^t)]\frac{\eta\tau}{\eta+\tau}$
  Compute $\pi^{t+1}(u^{t+1}, v^t, U^t)$ and $V_{\pi_{t+1}}$
  Compute $\xi_{t+1} \leftarrow P_{T_{U_t}\mathrm{St}}(V_{\pi_{t+1}} U_t)$
  Compute $U_{t+1} \leftarrow \mathrm{Retr}_{U_t}(\gamma\xi_{t+1})$
**end for**

**Remark 4.1** *Since we can formulate the entropic regularized UOT as the following Fenchel-Legendre dual form. It is equivalent to finding the optimal solution for the follow-*

*ing objective function:*

$$\min_{U\in\mathcal{M},u,v\in\mathbf{R}^n} \eta \sum_{i,j=1}^{n} \exp\left(\frac{u_i+v_j-\|U^Tx_i-U^Ty_j\|^2}{\eta}\right) + \tau\langle e^{\frac{v}{\tau}},\mathbf{c}\rangle + \tau\langle e^{\frac{u}{\tau}},\mathbf{r}\rangle. \tag{9}$$

*The main difficulty to prove the convergence of RBCD-UOT is to show that the updating of $U$ can decrease the Eq. 9. This part of the proof fails because the first term which contains the exponential of $U$ cannot converge in this formulation.*

## 5 Numerical Experiments

In this section, we first study the ability of the two proposed algorithms, RGAS-UOT and RBCD-UOT, to overcome the curse of dimensionality. We then evaluate the performance of the two algorithms on calculating the PRUOT distance for both synthetic and real datasets following the experiment setup in Paty and Cuturi (2019) and Lin et al. (2021). The baseline approach is the Sinkhorn algorithm that solves the entropic regularized Unbalanced Optimal Transport (UOT) problem proposed in Pham et al. (2020). All experiments in this section are implemented in Python 3.7 on a Linux server with a 6-core Intel Xeon CPU (E5-2630 v2 @ 2.60GHz).

### 5.1 Alleviating the curse of dimensionality

Our first experiment shows that our formulation of UOT can overcome the curse of dimensionality. We focus on the fragmented hypercube, which is adapted from Paty and Cuturi (2019); Lin et al. (2020). In particular, we consider a uniform distribution over a hypercube $\mu = \mathcal{U}([-1,1]^d)$ and a pushforward $\nu = T_{\#}\mu$ defined under the map $T(x) = x + 2sign(x) \odot (\sum_{k=1}^{k^*} e_k)$. Note that sign$(\cdot)$ is taken element-wise, $k^* \in [d]$ and $(e_1,\ldots,e_d)$ is the canonical basis of $\mathbb{R}^d$. By the definition, $T$ divides the hypercube into $2^{k^*}$ different hyper-rectangles, as well as serves as a subgradient of the convex function. This together with Brenier's theorem (Brenier, 1991) implies that $T$ is an optimal transport map between $\mu$ and $\nu = T_{\#}\mu$ with $\mathbb{W}_2^2(\mu,\nu) = 4k^*$. In this case the displacement vector $T(x)x$ lies in the $k^*$-dimensional subspace spanned by $\{e_j\}_{j\in[k^*]}$. Putting these pieces together yields that $\mathcal{P}_k^2(\mu,\nu) = 4k^*$ for any $k \geq k^*$. Moreover, in this case we have $U^* \in St(d,k^*)$ with $U^*(1:k^*,1:k^*) = I_{k^*}$. For all experiments below, unless specified, we set parameters as $\tau = 100$, $\eta = 0.2$, $\epsilon_{RGAS-UOT} = \epsilon_{RBCD-UOT} = 0.1$, $\gamma_{RGAS-UOT} = \gamma_{RBCD-UOT}/\eta$, and $\gamma_{RBCD-UOT} = 0.001$.

In this experiment, we set $k^* = 2$ and plot the number of data and computation time (in seconds) required by different algorithms to achieve a fixed mean estimation error for different total dimensions $d \in [2,13]$. We define the mean estimation error as $MEE = \mathcal{P}_k^2(\hat{\mu},\hat{\nu}) - 4k^*$ for the PRUOT distance and $MEE = \mathcal{W}_k^2(\hat{\mu},\hat{\nu}) - 4k^*$ for UOT distance. Figure 1 and Figure 2 show that the required sample size of ordinary UOT formulation grows exponentially with respect to the total dimension $d$ while our formulation of UOT does not display such behavior. We further plot Figure 2, which depicts the time needed by each algorithm to achieve an MEE of 1. In particular, we run each algorithm with increasing dimension $d$ using the required number of data points reported in Figure 1 for each dimension. We discover that for obtaining a fixed error rate, our formulation of UOT, especially with the RBCD-UOT algorithm, is significantly faster than the ordinary formulation of UOT. The results demonstrate that our formulation of UOT can significantly alleviate the curse of dimensionality.

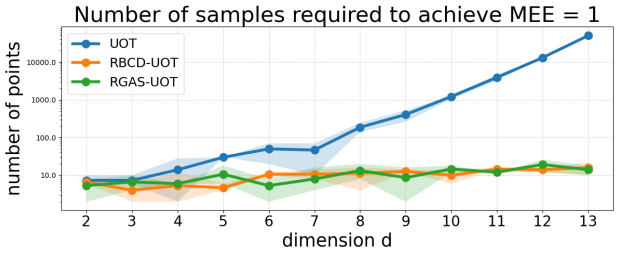


Figure 1: The number of samples required to achieve MEE = 1. The shaded areas represent the min and max quantiles over 3 repeated experiments. The required sample size of ordinary UOT formulation grows exponentially while our formulation of UOT does not.

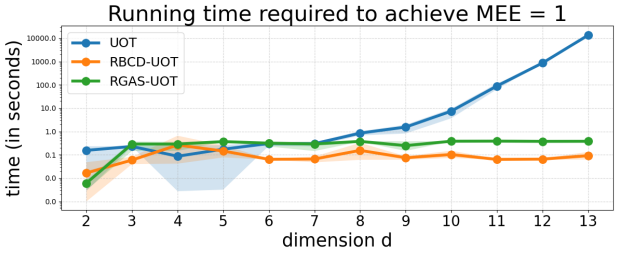


Figure 2: Computation time required to achieve MEE = 1. The shaded areas represent the min and max quantiles over 3 repeated experiments. For obtaining a fixed error rate, our formulation of UOT, especially with the RBCD-UOT algorithm, is significantly faster than the ordinary UOT.

### 5.2 Performance comparable to ordinary UOT

Using the fragmented hypercube setting in Experiment 5.1, we demonstrate that the performance of PRUOT is

comparable if not better than the ordinary UOT. Figure 3 and Figure 4 presents the mean estimation error (MEE) and mean subspace estimation error (MSE) as defined in Experiment 5.1 for a different choice of $n \in \{25, 50, 100, 250, 500, 1000\}$. We set $k^* = 2, d = 30$ and the subspace projection is calculated as $\Omega = \hat{U}\hat{U}^T$ in each run. The quality of solutions obtained by the RGAS-UOT and RAGAS-UOT algorithms are roughly the same and are better than the ordinary formulation of UOT.

In Figure 5 we plot the optimal transport plans between $(\hat{\mu}, \hat{\nu})$ generated by the UOT distance and the PRUOT distance calculated by the RGAS-UOT and RBCD-UOT algorithms. We considered the case when $k^* = 2, d = 30$ and $n \in 100, 250$. From Figure 5 we see that in both cases, our formulation of UOT can generate almost the same transport plans as the ordinary formulation of UOT, and the two proposed algorithms produce similar results.

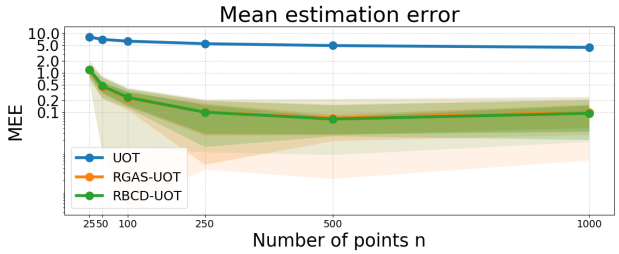


Figure 3: Mean estimation error with the varying number of points n. The shaded area shows the 10%-90% and 25%-75% quantiles over 10 repeated experiments. PRUOT has a lower error rate than UOT.

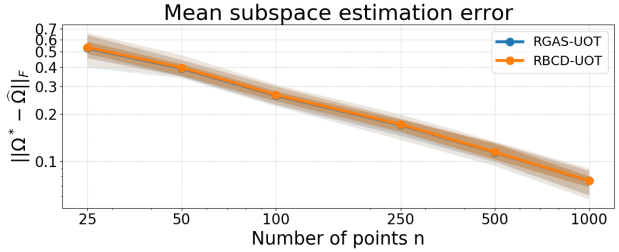


Figure 4: Mean subspace estimation error with the varying number of points n. The shaded area shows the 10%-90% and 25%-75% quantiles over 10 repeated experiments.

### 5.3 More robust to noise

In this experiment we consider $\mu = \mathcal{N}(0, \Sigma_1)$ and $\nu = \mathcal{N}(0, \Sigma_2)$ with $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ are positive semidefinite matrices of rank $k^*$. This implies that the support of $\mu$ and $\nu$ are $k^*$-dimensional subspace of $\mathbb{R}^d$. Although the supports of $\mu$ and $\nu$ can be different, their union is included in a $2k^*$-dimensional subspace. Therefore, for any $k \geq 2k^*$, $\mathcal{P}_k^2(\mu, \nu) = \mathbb{W}_2^2(\mu, \nu)$. In our experiment, we sample 10 independent couples of covariance matrices $(\Sigma_1, \Sigma_2)$ in dimension $d = 20$, where each has independently a Wishart distribution with $k^* = 5$ degrees of freedom. For each pair of matrices, we construct the empirical measures $\hat{\mu}, \hat{\nu}$ by drawing $n = 100$ points from $\mathcal{N}(0, \Sigma_1)$ and $\mathcal{N}(0, \Sigma_2)$.

Figure 6 presents the mean value of $\mathcal{P}_k^2(\mu, \nu)/\mathbb{W}_2^2(\mu, \nu)$ over 10 experiments with varying k. We set $\eta = 1.3$ and use RGAS-UOT for this experiment. We plot the curves for both noise-free and noisy data, where we add a white noise $(\mathcal{N}(0, I_d))$ to each data point. We calculate the PRUOT distance with the RGAS-UOT algorithm. With moderate noise, the data is approximately on the two 5-dimensional subspaces, and PRUOT distances do not vary too much. Our results are consistent with the result presented in Paty and Cuturi [Paty and Cuturi (2019), Figure 6], indicating that the PRUOT distance is also robust to random noise.

Figure 7 is the comparison of mean relative errors over 10 samples as the noise level varies. In particular, we consider 10 independent samples of couples $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$ similar to the above experiment and gradually add Gaussian noise $\sigma\mathcal{N}(0, I_d)$ to the points to construct the empirical measures $\hat{\mu_\sigma}$ and $\hat{\nu_\sigma}$. We set the regularization parameter as $\eta = 1.3$ when noise level $\leq 3$ and $\eta = 7$ otherwise. For the relative errors of the Wasserstein, UOT, and PRUOT distances, we follow the definition in Paty and Cuturi [Paty and Cuturi (2019), Section 6.3]. We see that when the noise has a moderate to high variance, the PRUOT distance is more robust to noise compared to the UOT distance.

### 5.4 Computational complexity

We conduct our third experiment on the fragmented hypercube with increasing dimension $d$ and fixed $k^* = 2$. We set $\epsilon = 0.1$, and $k = 2$ . For the PRUOT distances, we set the regularization parameter as $\eta = 0.2$ when $d \leq 250$ and $\eta = 0.5$ otherwise. We stop the RGAS-UOT and RAGAS-UOT algorithms when $\|grad_p(U^t)\|_F \leq \epsilon$

Figure 8 presents the mean computation time of the PRUOT distance with the RGAS-UOT and RBCD-UOT algorithms and the UOT distance with Sinkhorn algorithm (Pham et al., 2020). We see that for a fixed number of points $n$, the PRUOT algorithms are slower than UOT algorithm due to the computation of Rieman steps. However, recall that in Experiment 5.1, we show that for obtaining a fixed error rate, our formulation of UOT, especially with RBCD-UOT algorithm, is significantly faster than the ordinary formulation of UOT.

### 5.5 Experiments on real data

We consider a corpus of seven movie scripts that were used in Paty and Cuturi (2019). Each script is tokenized to a

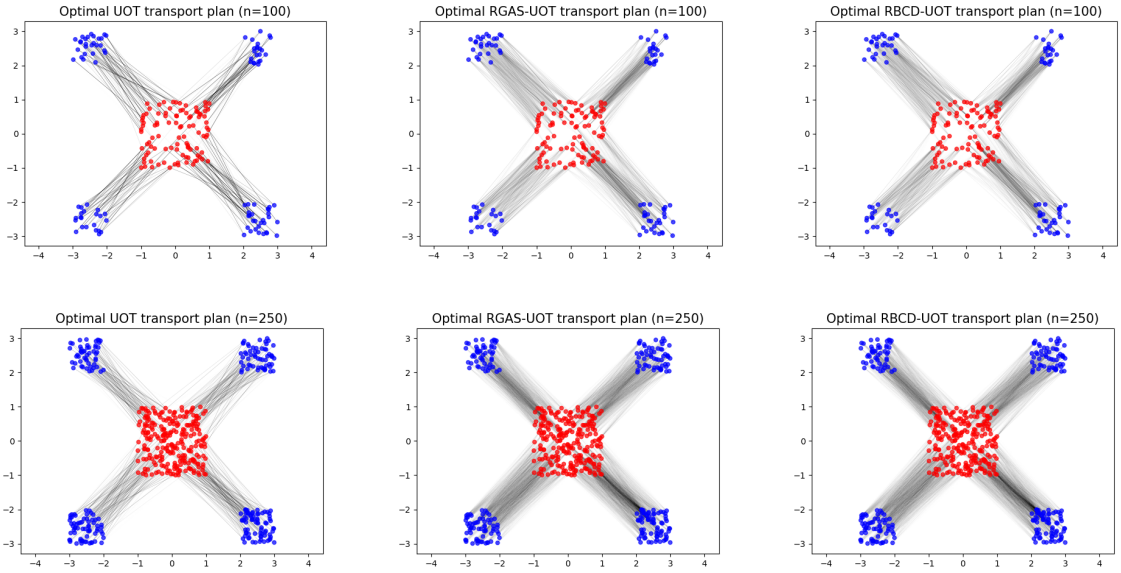


Figure 5: Fragmented hypercube with $(n, d) = (100, 30)$ (above) and $(n, d) = (250, 30)$ (bottom). Optimal mappings are calculated by UOT (left), PRW distance is calculated by RGAS-UOT (middle), and PRW is calculated by RBCD-UOT (right).

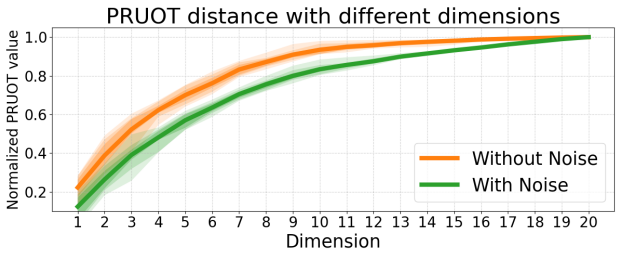


Figure 6: Mean normalized PRUOT distance depending on the dimension. The shaded area shows the 10%-90% and 25%-75% quantiles over the 10 runs.

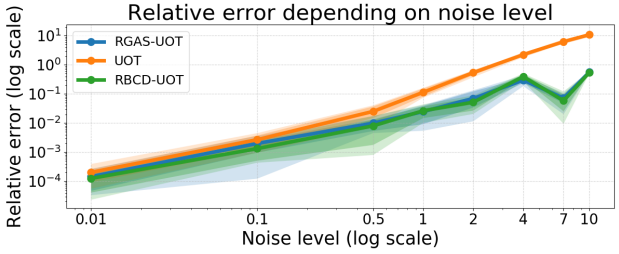


Figure 7: Comparison of mean relative errors over 10 runs depending on the noise level. The shaded area shows the 10%-90% and 25%-75% quantiles.

list of words, which is transformed to a measure over $\mathbb{R}^{300}$ using word2vec (Mikolov et al., 2018) where the weights

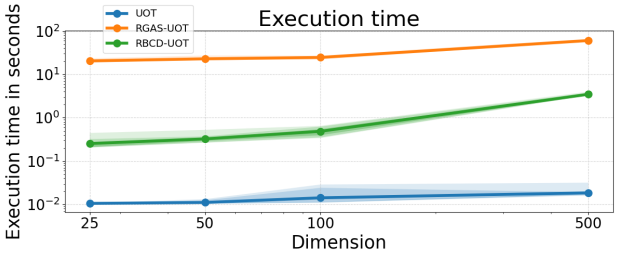


Figure 8: Comparisons of mean computation times on CPU with the number of points $n = 100$. The shaded areas show the minimum and maximum values over 10 runs.

correspond to word frequency. We set $k = 2, \eta = 0.1, \gamma = 0.08$ and compute the PRUOT distances between all pairs of movies as shown in Table 3. The results indeed reveal very useful information about the datasets. For example, from Table 3 we know that the movies "Dunkirk" and "Titanic" are close, and "Kill Bill Vol.1" and "Kill Bill Vol. 2" are close because their PRUOT distances are small.

## 6 Conclusion

In this paper, we propose a new model which solves unbalanced optimal transport by projection robust Wasserstein distance. We analyze the effectiveness of Riemannian gradient ascent with Sinkhorn (RGAS) and Riemannian Block Coordinate Descent (RBCD) under unbalanced optimal transport. Furthermore, we show the complexity

Table 2: Each entry is the PRUOT distance between different movie scripts. Bold values correspond to the minimum of each line. D = Dunkirk, G = Gravity, I= Interstellar, KB1 = Kill Bill Vol.1, KB2 = Kill Bill Vol.2, TM = The Martian, T = Titanic.

| | D | G | I | KB1 | KB2 | TM | T |
|---|---|---|---|---|---|---|---|
| D | 0.000 | 0.114 | 0.130 | 0.135 | 0.124 | 0.116 | **0.101** |
| G | 0.114 | 0.000 | 0.097 | 0.137 | 0.150 | **0.088** | 0.102 |
| I | 0.130 | 0.097 | 0.000 | 0.128 | 0.140 | **0.096** | 0.113 |
| KB1 | 0.135 | 0.137 | 0.128 | 0.000 | **0.083** | 0.107 | 0.107 |
| KB2 | 0.124 | 0.150 | 0.140 | **0.083** | 0.000 | 0.136 | 0.111 |
| TM | 0.116 | **0.088** | 0.096 | 0.107 | 0.136 | 0.000 | 0.096 |
| T | 0.101 | 0.102 | 0.113 | 0.107 | 0.111 | **0.096** | 0.000 |

of arithmetic operations for RGAS to obtain an $\epsilon$-stationary point. Numerical results on both synthetic and real datasets demonstrate that the new model strongly addresses the curse of dimensionality for the UOT problem and is more robust to noise.

## References


P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

J.-D. Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

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

P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29 (1):3–18, 2022.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1): 40–50, 1969.

K. Fatras, T. Séjourné, R. Flamary, and N. Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

M. Huang, S. Ma, and L. Lai. A riemannian block coordinate descent method for computing the projection robust wasserstein distance. In *International Conference on Machine Learning*, pages 4446–4455. PMLR, 2021.

J. Lee, N. P. Bertrand, and C. J. Rozell. Parallel unbalanced optimal transport regularization for large scale imaging problems. *arXiv preprint arXiv:1909.00149*, 2019.

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

T. Lin, C. Fan, N. Ho, M. Cuturi, and M. Jordan. Projection robust wasserstein distance and riemannian optimization. *Advances in neural information processing systems*, 33:9383–9397, 2020.

T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pages 262–270. PMLR, 2021.

T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1008`.

J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

K. O'Connor, K. McGoff, and A. B. Nobel. Optimal transport for stationary markov chains via policy iteration. *J. Mach. Learn. Res.*, 23:45–1, 2022.

S. Ozair, C. Lynch, Y. Bengio, A. Van den Oord, S. Levine, and P. Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

F.-P. Paty and M. Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

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

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

R. T. Rockafellar. *Convex analysis*, volume 36. Princeton university press, 1970.

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, and C. Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021.

# Projection Robust Optimal Transport Between Unbalanced Distributions

## 7 MISSING PROOFS

First we list the quantities that will be mentioned in the following proof.

$$R = \max\left\{\|\log(\mathbf{r})\|_\infty, \|\log(\mathbf{c})\|_\infty\right\} + \max\left\{\log(n), \frac{1}{\eta}\|C\|_\infty - \log(n)\right\},$$

$$H = \left(\frac{1}{2} + \frac{\eta\log(n)}{2\tau - 2\eta\log(n)}\right)(\alpha+\beta) + \frac{1}{6\log(n)},$$

$$\gamma = \frac{1}{H\|C\|_\infty(8L_1^2 + 16L_2) + 16(\eta+2\tau)^{-1}L_1^2\|C\|_\infty^2},$$

$$S = \frac{1}{2}(\alpha+\beta) + \frac{1}{2} + \frac{1}{4\log(n)},$$

$$D = \left(\frac{\alpha+\beta}{2}\right)\left[\log\left(\frac{\alpha+\beta}{2}\right) + 2\log(n) - 1\right] + \log(n) + \frac{5}{2},$$

$$J = \max\left\{S + D, 2\epsilon_2, \frac{4\epsilon_2\log(n)}{\tau}, \frac{4\epsilon_2(\alpha+\beta)\log(n)}{\tau}\right\}.$$

Define

$$\max_{U\in\mathcal{M}}\left\{f(U) := \min_{\pi\in\mathbb{R}_+^{n\times n}}\sum_{i,j=1}^n \pi_{i,j}\|U^T x_i - U^T y_j\|^2 + \tau\mathbf{KL}(\pi\mathbf{1}_n\|\mathbf{r}) + \tau\mathbf{KL}(\pi^T\mathbf{1}_n\|\mathbf{c})\right\} \tag{10}$$

and

$$\max_{U\in\mathcal{M}}\left\{f_{\tau,\eta}(U) := \min_{\pi\in\mathbb{R}_+^{n\times n}}\sum_{i,j=1}^n \pi_{i,j}\|U^T x_i - U^T y_j\|^2 + \tau\mathbf{KL}(\pi\mathbf{1}_n\|\mathbf{r}) + \tau\mathbf{KL}(\pi^T\mathbf{1}_n\|\mathbf{c}) - \eta H(\pi)\right\} \tag{11}$$

**Definition 7.1** *The correlation matrix between* $\mathbf{r} = \sum_{i=1}^n r_i\delta_{x_i}$ *and* $\mathbf{c} = \sum_{j=1}^n c_j\delta_{y_j}$ *is defined by* $V_\pi = \sum_{i=1}^n\sum_{j=1}^n \pi_{i,j}(x_i - y_j)(x_i - y_j)^T \in \mathbb{R}_{d\times d}$. The supplementary materials may contain detailed proofs of the results that are missing in the main paper.

**Proposition 7.1** *Boumal et al. (2019) For all* $Z \in \mathrm{St} \equiv \mathrm{St}(d,k)$ *and* $\xi \in \mathrm{T_Z St}$*, there exists constants* $L_1 > 0$ *and* $L_2 > 0$ *such that the following to inequalities hold:*

$$\|\mathrm{Retr}_Z - Z\|_F \le L_1\|\xi\|_F$$

$$\|\mathrm{Retr}_Z - (Z+\xi)\|_F \le L_2\|\xi\|_F^2$$

**Lemma 7.1** *The function* $f$ *is* $(\alpha+\beta)\cdot\|C\|_\infty$*-weakly concave.*

**Proof:** By Vial (Vial (1983), Proposition 4.3), it sufficies to show that the function $f(U) - \left(\frac{\alpha+\beta}{2}\right) \cdot \|C\|_\infty \|U\|_F^2$ is concave for any $U \in \mathcal{M}$. By the definition of $f$, we have

$$f(U) = \min_{\pi \in \mathbb{R}_+^{n \times n}} \mathrm{Trace}(U^T V_\pi U) + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}).$$

Define $\hat{\pi}(U) := \arg\min_{\pi \in \mathbb{R}_+^{n \times n}} \left\{ \mathrm{Trace}\left(U^T V_\pi U\right) + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}) \right\}$ and $\hat{x}(U) = \sum_{i=1}^n \sum_{j=1}^n \hat{\pi}_{i,j}(U)$ for some fixed $U$. Therefore,

$$f(U) = \mathrm{Trace}(U^T V_{\hat{\pi}(U)} U) + \tau \mathbf{KL}(\hat{\pi}(U) \cdot \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\hat{\pi}(U)^T \cdot \mathbf{1}_n \| \mathbf{c}).$$

By Corollary 2 in Pham et al. (2020), we have $\hat{x}(U) \leq \frac{\alpha+\beta}{2}$. Therefore, we define the space of $\pi$ that satisfies this constraint as $\Pi\left(\frac{\alpha+\beta}{2}\right)$ which is a semi-hypersphere of dimension $\mathbb{R}_+^{n \times n}$. Putting these pieces together with Jensen's inequality, we have

$$\|V_\pi\|_F \leq \sum_{i=1}^n \sum_{j=1}^n \pi_{i,j} \|(x_i - y_j)(x_i - y_j)^T\|_F \leq \max_{1 \leq i,j \leq n} \|x_i - y_j\|^2 \cdot \left(\frac{\alpha+\beta}{2}\right) = \left(\frac{\alpha+\beta}{2}\right) \cdot \|C\|_\infty$$

This implies that $U \to \mathrm{Trace}(U^T V_\pi U) + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}) - \left(\frac{\alpha+\beta}{2}\right) \cdot \|C\|_\infty \|U\|_F^2$ is concave for any $\pi \in \Pi\left(\frac{\alpha+\beta}{2}\right)$. Since $\Pi\left(\frac{\alpha+\beta}{2}\right)$ is compact, Danskin's theorem Rockafellar (1970) implies the desired result. ■

**Lemma 7.2** *Each element of the subdifferential $\partial f(U)$ is bounded by $(\alpha+\beta) \cdot \|C\|_\infty$ for all $U \in \mathrm{St}(d,k)$.*

**Proof:** By the definition of the subdifferential $\partial f$, it suffices to show that $\|V_\pi U\|_F \leq \left(\frac{\alpha+\beta}{2}\right) \cdot \|C\|_\infty$ for all $\pi \in \Pi\left(\frac{\alpha+\beta}{2}\right)$ and $U \in \mathrm{St}(d,k)$. Indeed, by the definition, $V_\pi$ is symmetric and positive semi-definite. Therefore, we have

$$\max_{U \in \mathrm{St}(d,k)} \|V_\pi U\|_F \leq \|V_\pi\|_F \leq \left(\frac{\alpha+\beta}{2}\right) \cdot \|C\|_\infty.$$

Putting these pieces together implies the desired result. ■

We present the *Riemannian gradient ascent with Sinkhorn on Unbalaned Optimal Transport* (RGAS-UOT) algorithm for solving Eq. (11). By the definition of $V_\pi$, we can rewrite

$$f_{\tau,\eta}(U) = \min_{\pi \in \mathbb{R}_+^{n \times n}} \left\{ \langle UU^T, V_\pi \rangle + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}) - \eta H(\pi) \right\}.$$

Fix $U \in \mathbb{R}^{d \times k}$, and define the mapping $\pi \to \langle UU^T, V_\pi \rangle + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}) - \eta H(\pi)$ with respect to $l_1$-norm. Danskin's theorem Rockafellar (1970) implies that $f_{\tau,\eta}(U)$ is smooth. Moreover, by the symmetry of $V_\pi$, we have

$$\nabla f_{\tau,\eta}(U) = 2 V_{\pi^*(U)} U \text{ for any } U \in \mathbb{R}^{d \times k},$$

where $\pi^*(U) := \arg\min_{\pi \in \mathbb{R}_+^{n \times n}} \left\{ \langle UU^T, V_\pi \rangle + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}) - \eta H(\pi) \right\}$. Define $x^*(U) = \sum_{i=1}^n \sum_{j=1}^n {\pi^*}_{i,j}(U)$ for some fixed $U$. This entropic regularized UOT is solved at each inner loop of the maximization and we use the output $\pi_{t+1} \approx \pi(U_t)$ to obtain an inexact gradient of $f_{\tau,\eta}$. The stopping criterion used here is set as $\|\pi_{t+1} - \pi(U_t)\|_1 \leq \hat{\epsilon}$ which implies that $\pi_{t+1}$ is $\epsilon$-approximate optimal transport plan for $U_t \in \mathrm{St}(d,k)$.

The remaining issue is to approximately solve an entropic regularized UOT efficiently. We leverage the approach in Pham et al. (2020) and obtain the desired output $\pi_{t+1}$ for $U_t \in \mathrm{St}(d,k)$ using the Sinkhorn iteration between unbalanced distributions. By adapting the proof presented by Pham *et.al.* (Pham et al. (2020), Theorem 2), we derive that Sinkhorn iteration achieves a finite-time guarantee.,

We first show that $f_{\tau,\eta}$ is continuously differentiable over $\mathbb{R}^{d\times k}$ and the classical gradient inequality holds true over $\mathrm{St}(d,k)$. The derivative is novel and uncovers the structure of the computation of entropic regularized PRW. Let $g : \mathbb{R}^{d\times k}\times\mathbb{R}_+^{n\times n}\to\mathbb{R}$ be defined by

$$g(U,\pi)=\sum_{i=1}^n\sum_{j=1}^n\pi_{i,j}\|U^Tx_i-U^Ty_j\|^2+\tau\mathbf{KL}(\pi\mathbf{1}_n\|\mathbf{r})+\tau\mathbf{KL}(\pi^T\mathbf{1}_n\|\mathbf{c})-\eta H(\pi).$$

It is obvious that $g(U,\bullet)$ is strongly convex with respect to $l_1$-norm.

**Lemma 7.3** *$f_{\tau,\eta}$ is differentiable over $\mathbb{R}^{d\times k}$ and $\|\nabla f_{\tau,\eta}(U)\|_F\le\left((1+\frac{\eta\log(n)}{\tau-\eta\log(n)})(\alpha+\beta)+\frac{1}{3\log(n)}\right)\cdot\|C\|_\infty$ for all $U\in\mathrm{St}(d,k)$.*

**Proof:** It is clear that we have $f_{\tau,\eta}(\bullet)=\min_{\pi\in\mathbb{R}_+^{n\times n}}g(\bullet,\pi)$. Furthermore, $\pi^*(\bullet)\arg\min_{\pi\in\mathbb{R}_+^{n\times n}}g(\bullet,\pi)$ is uniquely defined. Putting these pieces together, Danskin's theorem Rockafellar (1970) implies that $f_{\tau,\eta}$ is continuously differentiable and the gradient is

$$\nabla f_{\tau,\eta}(U)=2V_{\pi^*(U)}U \text{ for any } U\in\mathbb{R}^{d\times k},$$

By Corollary 2 Pham et al. (2020), we have $x^*(U)\le\left(\frac{1}{2}+\frac{\eta\log(n)}{2\tau-2\eta\log(n)}\right)(\alpha+\beta)+\frac{1}{6\log(n)}$. Denote $H=\left(\frac{1}{2}+\frac{\eta\log(n)}{2\tau-2\eta\log(n)}\right)(\alpha+\beta)+\frac{1}{6\log(n)}$, there is $\pi^*\in\Pi(H)$. Since $U\in\mathrm{St}(d,k)$, we have

$$\|\nabla f_{\tau,\eta}(U)\|_F=2\|V_{\pi^*(U)}U\|_F\le2\|V_{\pi^*(U)}\|_F\le2H\|C\|_\infty.$$

This completes the proof. ■

**Lemma 7.4** *For all $U_1,U_2\in\mathrm{St}(d,k)$, the following statement holds true,*

$$|f_{\tau,\eta}(U_1)-f_{\tau,\eta}(U_2)-\langle\nabla f_{\tau,\eta}(U_2),U_1-U_2\rangle|\le\left(H\|C\|_\infty+\frac{2\|C\|_\infty^2}{\eta+2\tau}\right)\|U_1-U_2\|_F^2.$$

**Proof:** It suffices to prove hat

$$\|\nabla f_{\tau,\eta}(\alpha U_1+(1-\alpha)U_2)-\nabla f_{\tau,\eta}(U_2)\|_F\le\left(2H\|C\|_\infty+\frac{4\|C\|_\infty^2}{\eta+2\tau}\right)\alpha\|U_1-U_2\|_F,$$

For any $U_1,U_2\in\mathrm{St}(d,k)$ and any $\alpha\in[0,1]$. Indeed, let $U_\alpha=\alpha U_1+(1-\alpha)U_2$, we have

$$\|\nabla f_{\tau,\eta}(U_\alpha)-\nabla f_{\tau,\eta}(U_2)\|_F\le2\|V_{\pi^*(U_\alpha)}\|_F\|U_\alpha-U_2\|_F+2\|V_{\pi^*(U_\alpha)}-V_{\pi^*(U_2)}\|_F.$$

Since $\pi^*(U_\alpha)\in\Pi(S)$, we have $\|V_{\pi^*(U_\alpha)}\|_F\le S\|C\|_\infty$. By the definition of $V_\pi$, we have

$$\|V_{\pi^*(U_\alpha)}-V_{\pi^*(U_2)}\|_F\le\sum_{i=1}^n\sum_{j=1}^n|\pi_{i,j}^*(U_\alpha)-\pi_{i,j}^*(U_2)|\|x_i-y_j\|^2\le\|C\|_\infty\|\pi^*(U_\alpha)-\pi^*(U_2)\|_1.$$

Putting these pieces together yields that

$$\|\nabla f_{\tau,\eta}(U_\alpha)-\nabla f_{\tau,\eta}(U_2)\|_F\le2H\|C\|_\infty\|U_\alpha-U_2\|_F+2\|C\|_\infty\|\pi^*(U_\alpha)-\pi^*(U_2)\|_1 \tag{12}$$

Using the property of he enropy regularization $H(\bullet)$, we have $g(U,\bullet)$ is strongly convex with respect to $l_1$-norm and the module is $\eta$. This implies that

$$g(U_\alpha,\pi^*(U_2))-g(U_\alpha,\pi^*(U_\alpha))-\langle\nabla_\pi g(U_\alpha,\pi^*(U_\alpha)),\pi^*(U_2)-\pi^*(U_\alpha)\rangle\ge\frac{\eta+2\tau}{2}\|\pi^*(U_\alpha)-\pi^*(U_2)\|_1^2,$$

$$g(U_\alpha,\pi^*(U_\alpha))-g(U_\alpha,\pi^*(U_2))-\langle\nabla_\pi g(U_\alpha,\pi^*(U_2)),\pi^*(U_\alpha)-\pi^*(U_2)\rangle\ge\frac{\eta+2\tau}{2}\|\pi^*(U_\alpha)-\pi^*(U_2)\|_1^2.$$

Summing up these inequalities yields

$$\langle\nabla_\pi g(U_\alpha,\pi^*(U_\alpha))-\nabla_\pi g(U_\alpha,\pi^*(U_2)),\pi^*(U_\alpha)-\pi^*(U_2)\rangle\ge(\eta+2\tau)\|\pi^*(U_\alpha)-\pi^*(U_2)\|_1^2. \tag{13}$$

Futhermore, by the first-order optimality condition of $\pi^*(U_1)$ ad $\pi^*(U_2)$, we have

$$\langle \nabla_\pi g(U_\alpha, \pi^*(U_\alpha)), \pi^*(U_2) - \pi^*(U_\alpha) \rangle \geq 0,$$
$$\langle \nabla_\pi g(U_2, \pi^*(U_2)), \pi^*(U_\alpha) - \pi^*(U_2) \rangle \geq 0.$$

Summing up these inequalities yields

$$\langle \nabla_\pi g(U_2, \pi^*(U_2)) - \nabla_\pi g(U_\alpha, \pi^*(U_\alpha)), \pi^*(U_\alpha) - \pi^*(U_2) \rangle \geq 0. \tag{14}$$

Summing up the Eq. 13 and Eq. 14 and further using Holder's inequality, we have

$$\|\pi^*(U_\alpha) - \pi^*(U_2)\|_1 \leq \frac{1}{\eta + 2\tau} \|\nabla_\pi g(U_2, \pi^*(U_2)) - \nabla_\pi g(U_\alpha, \pi^*(U_2))\|_\infty.$$

By the definition of function $g$, we have

$$\begin{aligned}
\|\nabla_\pi g(U_2, \pi^*(U_2)) - \nabla_\pi g(U_\alpha, \pi^*(U_2))\|_\infty &\leq \max_{1 \leq i,j \leq n} |(x_i - x_j)^T (U_2 U_2^T - U_\alpha U_\alpha^T)(x_i - x_j)| \\
&\leq \left( \max_{1 \leq i,j \leq n} \|x_i - y_j\|^2 \right) \|U_2 U_2^T - U_\alpha U_\alpha^T\|_F \\
&= \|C\|_\infty \|U_2 U_2^T - U_\alpha U_\alpha^T\|_F.
\end{aligned}$$

Since $U_1, U_2 \in \mathrm{St}(d, k)$, we have

$$\begin{aligned}
\|U_2 U_2^T - U_\alpha U_\alpha^T\|_F &\leq \|U_2 (U_2 - U_\alpha)^T\|_F + \|(U_2 - U_\alpha) U_\alpha^T\|_F \\
&\leq \|U_2 - U_\alpha\|_F + \|(U_2 - U_\alpha)(\alpha U_1 + (1-\alpha) U_2)^T\|_F \\
&\leq \|U_2 - U_\alpha\|_F + \alpha \|(U_2 - U_\alpha) U_1^T\|_F + (1-\alpha) \|(U_2 - U_\alpha) U_2^T\|_F \\
&\leq 2 \|(U_2 - U_\alpha\|_F.
\end{aligned}$$

Putting these pieces together yields that

$$\|\pi^*(U_\alpha) - \pi^*(U_2)\|_1 \leq \frac{2\|C\|_\infty}{\eta + 2\tau} \|U_\alpha - U_2\|_F. \tag{15}$$

Plugging Eq. 15 into Eq. 12 yields the desired result. ■

Based on the lemma above, we further extend the Lemma from the worS of Lin *et.al.* from optimal transport problem to unbalanced optimal transport problem Lin et al. (2020).

**Lemma 7.5** *Let $\{(U_t, \pi^t)\}_{t \geq 1}$ be the iterates generated by Algorithm RGAS-UOT. We have*

$$\frac{1}{T} \left( \sum_{t=0}^{T-1} \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 \right) \leq \frac{4\Delta_f}{\gamma T} + \frac{3\epsilon^2}{25}.$$

*where $\Delta_f = \max_{U \in St(d,k), UU^T = I} f_{\tau,\eta}(U) - f_{\tau,\eta}(U_0)$ is the initial objective gap.*

**Proof:** Using lemma 7.4 with $U_1 = U_{t+1}$ and $U_2 = U_t$, we have

$$f_{\tau,\eta}(U_{t+1}) - f_{\tau,\eta}(U_t) - \langle \nabla f_{\tau,\eta}(U_t), U_{t+1} - U_t \rangle \geq -\left( H\|C\|_\infty + \frac{2\|C\|_\infty^2}{\eta + 2\tau} \right) \|U_{t+1} - U_t\|_F^2. \tag{16}$$

By the definition of $U_{t+1}$, we have

$$\begin{aligned}
\langle \nabla f_{\tau,\eta}(U_t), U_{t+1} - U_t \rangle =& \langle \nabla f_{\tau,\eta}(U_t), \mathrm{Retr}_{U_t}(\gamma \xi_{t+1}) - U_t \rangle \\
&= \langle \nabla f_{\tau,\eta}(U_t), \gamma \xi_{t+1} \rangle + \langle \nabla f_{\tau,\eta}(U_t), \mathrm{Retr}_{U_t}(\gamma \xi_{t+1}) - (U_t + \gamma \xi_{t+1}) \rangle \\
&\geq \langle \nabla f_{\tau,\eta}(U_t), \gamma \xi_{t+1} \rangle - \|\nabla f_{\tau,\eta}(U_t)\|_F \|\mathrm{Retr}_{U_t}(\gamma \xi_{t+1}) - (U_t + \gamma \xi_{t+1})\|_F
\end{aligned} \tag{17}$$

By Lemma 7.3, we have $\|\nabla f_{\tau,\eta}(U)\| \leq 2H\|C\|_\infty$. Putting these pieces with Proposition 7 yields that

$$\langle \nabla f_{\tau,\eta}(U_t), U_{t+1} - U_t \rangle \geq \gamma \langle \nabla f_{\tau,\eta}(U_t), \xi_{t+1} \rangle - 2H\gamma^2 L_2 \|C\|_\infty \|\xi_{t+1}\|_F^2. \tag{18}$$

Using Proposition 7 again, we have

$$\|U_{t+1} - U_t\|_F^2 = |\mathrm{Retr}_{U_t}(\gamma\xi_{t+1}) - U_t\|_F^2 \leq \gamma^2 L_1^2 \|\xi_{t+1}\|_F^2. \tag{19}$$

Combining Eq. 16, Eq. 18 and Eq. 19 yields

$$f_{\tau,\eta}(U_{t+1}) - f_{\tau,\eta}(U_t) \geq \gamma \langle \nabla f_{\tau,\eta}(U_t), \xi_{t+1} \rangle - \gamma^2 \|\xi_{t+1}\|_F^2 \left( (L_1^2 H + 2L_2 H)\|C\|_\infty + 2(\eta + 2\tau)^{-1} L_1^2 \|C\|_\infty^2 \right). \tag{20}$$

Recall that $\mathrm{grad} f_\eta(U_t) = P_{\mathrm{T}_{U_t}\mathrm{St}}(\nabla f_\eta(U_t))$ and $\xi_{t+1} = P_{\mathrm{T}_{U_t}\mathrm{St}}(2V_{\pi_{t+1}}U_t)$, we have

$$\langle \nabla f_{\tau,\eta}(U_t), \xi_{t+1} \rangle = \langle \mathrm{grad} f_{\tau,\eta}(U_t), \xi_{t+1} \rangle = \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 + \langle \mathrm{grad} f_{\tau,\eta}(U_t), \xi_{t+1} - \mathrm{grad} f_{\tau,\eta}(U_t) \rangle.$$

Using Young's inequality, we have

$$\langle \nabla f_{\tau,\eta}(U_t), \xi_{t+1} \rangle \geq \frac{1}{2} \left( \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 - \|\xi_{t+1} - \mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 \right).$$

Furthermore, we have $\|\xi_{t+1}\|_F^2 \leq 2\left( \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 + \|\xi_{t+1} - \mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 \right)$. Putting these pieces together with Eq. 20 yields that

$$\begin{aligned} f_{\tau,\eta}(U_{t+1}) - f_{\tau,\eta}(U_t) \geq & \gamma \left( \frac{1}{2} - \gamma(2L_1^2 H\|C\|_\infty + 4L_2 H\|C\|_\infty + 4(\eta+2\tau)^{-1} L_1^2 \|C\|_\infty^2) \right) \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 \\ & - \gamma \left( \frac{1}{2} + \gamma(2L_1^2 H\|C\|_\infty + 4L_2 H\|C\|_\infty + 4(\eta+2\tau)^{-1} L_1^2 \|C\|_\infty^2) \right) \|\xi_{t+1} - \mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2. \end{aligned} \tag{21}$$

Since $\xi_{t+1} = P_{\mathrm{T}_{U_t}\mathrm{St}}(2V_{\pi_{t+1}}U_t)$ and $\mathrm{grad} f_{\tau,\eta}(U_t) = P_{\mathrm{T}_{U_t}\mathrm{St}}(\nabla\tau, \eta(U_t))$ where $\pi_t^*$ is the minimizer of the entropic regularized UOT problem, *i.e.* $\pi_t^* \in \arg\min_{\pi \in \Pi(H)} \left\{ \langle UU^T, V_\pi \rangle + \tau \mathbf{KL}(\pi \mathbf{1}_n \| \mathbf{r}) + \tau \mathbf{KL}(\pi^T \mathbf{1}_n \| \mathbf{c}) - \eta H(\pi) \right\}$, we have

$$\|\xi_{t+1} - \mathrm{grad} f_{\tau,\eta}(U_t)\|_F \leq 2\|(V_{\pi_{t+1}} - V_{\pi_t^*})U_t\|_F = 2\|(V_{\pi_{t+1}} - V_{\pi_t^*})\|_F.$$

By the definition of $V_\pi$ and using the stopping criterion: $\|\pi_{t+1} - \pi_t^*\|_1 \leq \hat{\epsilon} = \frac{\epsilon_2}{10\|C\|_\infty}$, we have

$$\|(V_{\pi_{t+1}} - V_{\pi_t^*})\|_F \leq \|C\|_\infty \|\pi_{t+1} - \pi_t^*\|_1 \leq \frac{\epsilon_2}{10}.$$

Putting these pieces together yields that

$$\|\xi_{t+1} - \mathrm{grad} f_{\tau,\eta}(U_t)\|_F \leq \frac{\epsilon_2}{5}. \tag{22}$$

Plugging Eq. 22 into Ept. 21 with the definition of $\gamma$ yields that

$$f_{\tau,\eta}(U_{t+1}) - f_{\tau,\eta}(U_t) \geq \frac{\gamma \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2}{4} - \frac{3\gamma {\epsilon_2}^2}{100}.$$

Summing and rearranging the resulting inequality yields that

$$\frac{1}{T} \left( \sum_{t=0}^{T-1} \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2 \right) \leq \frac{4(f_{\tau,\eta}(U_t) - f_{\tau,\eta}(U_0))}{\gamma T} + \frac{3{\epsilon_2}^2}{25}.$$

■

**Lemma 7.6** *Consider a sphere set* $\mathcal{S} = \{x \in \mathbb{R}^d | \}$

**Theorem 7.1** *Letting* $\{(U_T, \pi^T)\}_{T \geq 1}$ *be the iterates generated by Algorithm RGAS-UOT, the number of iterations required to reach dist*$(0, \mathrm{subdiff} f(U_T)) \leq \epsilon_1$ *satisfies that*

$$T \leq \frac{100\Delta_f}{(25{\epsilon_1}^2 - 3{\epsilon_2}^2)\gamma}$$

**Proof:**

Given that $\|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F > \epsilon_1$ for all $t = 0, 1, \ldots, T-1$ and combine the inequality with Lemma 7.5 , we have

$$\frac{1}{T}\left(\sum_{t=0}^{T-1} \|\mathrm{grad} f_{\tau,\eta}(U_t)\|_F^2\right) \leq \frac{4\Delta_f}{\gamma T} + \frac{3{\epsilon_2}^2}{25}$$

Also,

$$\begin{aligned}\frac{1}{\gamma} &= (8L_1^2 H + 16L_2 H)\|C\|_\infty + 16(\eta + 2\tau)^{-1} L_1^2 \|C\|_\infty^2 \\ &= H(8L_1^2 + 16L_2)\|C\|_\infty + 16L_1^2\|C\|_\infty^2 \cdot \left(\min\left\{\frac{\epsilon_2}{S+D}, \frac{1}{2}, \frac{\tau}{4\log(n)}, \frac{\tau}{4(\alpha+\beta)\log(n))}, \frac{\epsilon_1}{40\log(n)}\right\} + 2\tau\right)^{-1}.\end{aligned}$$

We conclude that the upper bound $T$ must satisfy

$$\begin{aligned}&25{\epsilon_1}^2 - 3{\epsilon_2}^2 \\ &\quad \leq \frac{100\Delta_f}{T}\left(H(8L_1^2 + 16L_2)\|C\|_\infty + 16L_1^2\|C\|_\infty^2 \cdot \left(\min\left\{\frac{\epsilon_2}{S+D}, \frac{1}{2}, \frac{\tau}{4\log(n)}, \frac{\tau}{4(\alpha+\beta)\log(n))}, \frac{\epsilon_1}{40\log(n)}\right\} + 2\tau\right)^{-1}\right).\end{aligned}$$

Using Lemma 7.4, we have

$$\begin{aligned}\Delta_f &\leq \left(H\|C\|_\infty + \frac{2\|C\|_\infty^2}{\eta + 2\tau}\right)\left(\max_{U\in \mathrm{St}(d,k)} \|U_1 - U_2\|_F^2\right) + 2H\|C\|_\infty\left(\max_{U\in \mathrm{St}(d,k)} \|U_1 - U_2\|_F^2\right) \\ &= k \cdot \left(6H\|C\|_\infty + \frac{4\|C\|_\infty^2}{\eta + 2\tau}\right) \\ &= k \cdot \left(6H\|C\|_\infty + 4\|C\|_\infty^2 \cdot \left(\min\left\{\frac{\epsilon_2}{S+D}, \frac{1}{2}, \frac{\tau}{4\log(n)}, \frac{\tau}{4(\alpha+\beta)\log(n))}, \frac{\epsilon_1}{40\log(n)}\right\} + 2\tau\right)^{-1}\right)\end{aligned}$$

Putting these pieces together implies the desired result.

### 7.1 Additional Experiments

| | D | G | I | KB1 | KB2 | TM | T |
|---|---|---|---|---|---|---|---|
| D | 0.000 | 0.120 | 0.131 | 0.152 | 0.161 | 0.133 | **0.101** |
| G | 0.120 | 0.000 | 0.096 | 0.145 | 0.156 | **0.090** | 0.126 |
| I | 0.131 | 0.096 | 0.000 | 0.143 | 0.153 | **0.092** | 0.129 |
| KB1 | 0.152 | 0.145 | 0.143 | 0.000 | **0.088** | 0.144 | 0.125 |
| KB2 | 0.161 | 0.156 | 0.153 | **0.088** | 0.000 | 0.147 | 0.131 |
| TM | 0.133 | **0.090** | 0.092 | 0.144 | 0.147 | 0.000 | 0.132 |
| T | **0.101** | 0.126 | 0.129 | 0.125 | 0.131 | 0.132 | 0.000 |

Table 3: Each entry is $\mathcal{S}_k^2/\mathcal{P}_k^2$ distance between different movie scripts. D = Dunkirk, G = Gravity, I = Interstellar, KB1 = Kill Bill Vol.1, KB2 = Kill Bill Vol.2, TM = The Martian, T = Titanic

■

**Ability to capture the dimension of sampled measures** Figure **??** presents the behavior of $\mathcal{P}_k^2(\hat{\mu}, \hat{\nu})$ as a function of $k^* \in \{2, 4, 7, 10\}$, where $\hat{m}u, \hat{n}u$ are empirical distributions corresponding to $\mu$ and $\nu$, respectively. The sequence is concave and increases slowly after $k = k^*$ for both algorithms, which is reasonable since the last $d - k^*$ dimensions only represent noise. We see that the solutions of both the RBCD-UOT and the RGAS-UOT algorithms achieve almost the same quality.

**Comparing RGAS-UOT and RBCD-UOT** We also compare the computational time of the two different PRUOT algorithms, RGAS-UOT and RBCD-UOT. We fix $k^{=k=2}$, and generate the Fragmented Hypercube with varying $n, d$. 5 - 8

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.00/ 0.00 | 0.79/ 0.97 | 0.60/ **0.80** | 0.90/ 1.20 | 0.90/ 1.23 | 0.71/ 1.03 | **0.60**/ 0.81 | 0.66/ 0.86 | 0.77/ 1.06 | 0.80/ 1.09 |
| 1 | 0.79/ 0.97 | 0.00/ 0.00 | 0.52/ 0.66 | 0.70/ 0.86 | 0.53/ 0.68 | 0.68/ 0.84 | 0.65/ 0.80 | **0.47/ 0.58** | 0.70/ 0.88 | 0.69/ 0.85 |
| 2 | 0.60/ 0.80 | 0.52/ **0.66** | 0.00/ 0.00 | 0.53/ 0.73 | 0.78/ 1.08 | 0.81/ 1.08 | 0.69/ 0.90 | **0.51**/ 0.70 | 0.51/ 0.68 | 0.79/ 1.07 |
| 3 | 0.90/ 1.20 | 0.70/ 0.86 | 0.53/ 0.73 | 0.00/ 0.00 | 0.89/ 1.20 | **0.43/ 0.58** | 0.93/ 1.23 | 0.55/ 0.72 | 0.64/ 0.88 | 0.64/ 0.83 |
| 4 | 0.90/ 1.23 | 0.53/ 0.68 | 0.78/ 1.08 | 0.89/ 1.20 | 0.00/ 0.00 | 0.76/ 1.00 | 0.61/ 0.85 | 0.62/ 0.79 | 0.75/ 1.09 | **0.38/ 0.49** |
| 5 | 0.71/ 1.03 | 0.68/ 0.84 | 0.81/ 1.08 | **0.43/ 0.58** | 0.76/ 1.00 | 0.00/ 0.00 | 0.52/ 0.72 | 0.70/ 0.91 | 0.54/ 0.72 | 0.58/ 0.78 |
| 6 | 0.60/ 0.81 | 0.65/ 0.80 | 0.69/ 0.90 | 0.93/ 1.23 | 0.61/ 0.85 | **0.52/ 0.72** | 0.00/ 0.00 | 0.83/ 1.11 | 0.66/ 0.92 | 0.81/ 1.22 |
| 7 | 0.66/ 0.86 | 0.47/ **0.58** | 0.51/ 0.70 | 0.55/ 0.72 | 0.62/ 0.79 | 0.70/ 0.91 | 0.83/ 1.11 | 0.00/ 0.00 | 0.71/ 1.07 | **0.46**/ 0.62 |
| 8 | 0.77/ 1.06 | 0.70/ 0.88 | **0.51/ 0.68** | 0.64/ 0.88 | 0.75/ 1.09 | 0.54/ 0.72 | 0.66/ 0.92 | 0.71/ 1.07 | 0.00/ 0.00 | 0.61/ 0.87 |
| 9 | 0.80/ 1.09 | 0.69/ 0.85 | 0.79/ 1.07 | 0.64/ 0.83 | **0.38/ 0.49** | 0.58/ 0.78 | 0.81/ 1.22 | 0.46/ 0.62 | 0.61/ 0.87 | 0.00/ 0.00 |

Table 4: Each entry is scaled $\mathcal{S}_k^2/\mathcal{P}_k^2$ distance between different hand-written digits.

show the comparison for different algorithms with different $(n, d)$ pairs. All the reported CPU times are in seconds. We run each $n, d$ pair for 50 times and take the average. From Tables 1 - 4, we see that our RBCD-UOT algorithm runs faster than the RGAS-UOT algorithm in all cases. Moreover, we found that the advantage of RBCD-UOT over RGAS-UOT is more significant when n is relatively larger than d.

Table 5: CPU time for calculating PRUOT of the fragmented hypercube problem. We set $n = 100$

| DIMENSION d | 20 | 50 | 100 | 250 | 500 |
|---|---|---|---|---|---|
| RBCD-UOT | 0.64 | 0.66 | 0.80 | 2.84 | 5.58 |
| RGAS-UOT | 1.58 | 2.02 | 2.31 | 5.64 | 7.01 |

Table 6: CPU time for calculating PRW of the fragmented hypercube problem. We set $d = 50$.

| n | 50 | 100 | 250 | 500 | 1000 |
|---|---|---|---|---|---|
| RBCD-UOT | 0.15 | 0.68 | 2.49 | 4.06 | 9.60 |
| RGAS-UOT | 0.38 | 2.16 | 8.61 | 8.71 | 14.28 |

Table 7: CPU time for calculating PRUOT of the fragmented hypercube problem. We set $n = d$.

| DIMENSION d | 10 | 20 | 50 | 100 | 250 |
|---|---|---|---|---|---|
| RBCD-UOT | 0.13 | 0.08 | 0.11 | 0.76 | 10.37 |
| RGAS-UOT | 0.29 | 0.30 | 0.39 | 2.39 | 27.11 |

Table 8: CPU time for calculating PRUOT of the fragmented hypercube problem. We set $n = 10d$.

| DIMENSION d | 10 | 20 | 50 | 100 | 250 |
|---|---|---|---|---|---|
| RBCD-UOT | 0.48 | 5.68 | 3.49 | 11.01 | 174.26 |
| RGAS-UOT | 1.30 | 7.34 | 8.10 | 16.64 | 253.29 |

## References


P.-A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. In *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

J.-D. Benamou. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

N. Boumal, P.-A. Absil, and C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

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

P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

R. M. Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

K. Fatras, T. Séjourné, R. Flamary, and N. Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR, 2021.

M. Huang, S. Ma, and L. Lai. A riemannian block coordinate descent method for computing the projection robust wasserstein distance. In *International Conference on Machine Learning*, pages 4446–4455. PMLR, 2021.

J. Lee, N. P. Bertrand, and C. J. Rozell. Parallel unbalanced optimal transport regularization for large scale imaging problems. *arXiv preprint arXiv:1909.00149*, 2019.

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

T. Lin, C. Fan, N. Ho, M. Cuturi, and M. Jordan. Projection robust wasserstein distance and riemannian optimization. *Advances in neural information processing systems*, 33:9383–9397, 2020.

T. Lin, Z. Zheng, E. Chen, M. Cuturi, and M. I. Jordan. On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics*, pages 262–270. PMLR, 2021.

T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1008`.

J. Niles-Weed and P. Rigollet. Estimation of wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

K. O'Connor, K. McGoff, and A. B. Nobel. Optimal transport for stationary markov chains via policy iteration. *J. Mach. Learn. Res.*, 23:45–1, 2022.

S. Ozair, C. Lynch, Y. Bengio, A. Van den Oord, S. Levine, and P. Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

F.-P. Paty and M. Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

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

J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

R. T. Rockafellar. *Convex analysis*, volume 36. Princeton university press, 1970.

T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8(2):231–259, 1983.

F. Zhan, Y. Yu, K. Cui, G. Zhang, S. Lu, J. Pan, C. Zhang, F. Ma, X. Xie, and C. Miao. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021.