Low-Rank Optimal Transport through Factor Relaxation with Latent Coupling

Peter Halmos[1,*], **Xinhao Liu**[1,*], **Julian Gold**[2,*], and **Benjamin J. Raphael**[1]

[1]Department of Computer Science, Princeton University
[2]Center for Statistics and Machine Learning, Princeton University

Abstract

Optimal transport (OT) is a general framework for finding a minimum-cost transport plan, or coupling, between probability distributions, and has many applications in machine learning. A key challenge in applying OT to massive datasets is the quadratic scaling of the coupling matrix with the size of the dataset. Forrow et al. (2019) introduced a factored coupling for the k-Wasserstein barycenter problem, which Scetbon et al. (2021) adapted to solve the primal low-rank OT problem. We derive an alternative parameterization of the low-rank problem based on the *latent coupling* (LC) factorization previously introduced by Lin et al. (2021) generalizing Forrow et al. (2019). The LC factorization has multiple advantages for low-rank OT including decoupling the problem into three OT problems and greater flexibility and interpretability. We leverage these advantages to derive a new algorithm *Factor Relaxation with Latent Coupling* (FRLC), which uses *coordinate* mirror descent to compute the LC factorization. FRLC handles multiple OT objectives (Wasserstein, Gromov-Wasserstein, Fused Gromov-Wasserstein), and marginal constraints (balanced, unbalanced, and semi-relaxed) with linear space complexity. We provide theoretical results on FRLC, and demonstrate superior performance on diverse applications – including graph clustering and spatial transcriptomics – while demonstrating its interpretability.

1 Introduction

Optimal transport (OT) is a powerful geometric framework for comparing probability distributions. OT problems seek a transport plan $\boldsymbol{P}$ efficiently transforming one distribution ($\boldsymbol{a}$) into another ($\boldsymbol{b}$), subject to a ground cost $\boldsymbol{C}$. The minimum cost yields a distance between $\boldsymbol{a}$ and $\boldsymbol{b}$, while the optimal transport plan reveals key structural similarities between the distributions. Owing to its versatility – different ground costs result in different ways to compare data – OT has found many applications in machine learning and beyond: from self-attention Tay et al. (2020); Sander et al. (2022); Geshkovski et al. (2023) and domain adaptation Courty et al. (2014); Solomon et al. (2015) to computational biology Schiebinger et al. (2019); Yang et al. (2020); Bunne et al. (2023); Liu et al. (2023).

This versatility is compounded by several variants using different forms of the objective function and/or constraints on the transport plan $\boldsymbol{P}$. Wasserstein (W) OT Kantorovich (1942) compares distributions over the same space through the expected work of $\boldsymbol{P}$, while Gromov-Wasserstein (GW) OT Mémoli (2011) compares distributions supported on distinct geometries through the expected metric distortion of $\boldsymbol{P}$. Fused Gromov-Wasserstein (FGW) Vayer et al. (2020) OT is suited to structured data, taking a convex combination of the former two objectives. Independently, one can relax constraints on the *marginals* of $\boldsymbol{P}$: in computational applications, $\boldsymbol{P}$ is a matrix whose row-sum $\boldsymbol{P}\mathbf{1}_m$ and column-sum $\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n$ are called its left and right marginals. *Balanced* OT requires $\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}$ and $\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{b}$. *Unbalanced* OT Frogner et al. (2015) replaces these constraints with penalties in the

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

transport cost, and is more robust to outliers. *Semi-relaxed* OT can be used to understand how one dataset embeds into another by imposing one hard constraint on either the left or right marginal, used for feature transfer Dong et al. (2023), and alignment of spatiotemporal data Halmos et al. (2024).

An important consideration in applying OT is the quadratic space of the transport plan. To address both the quadratic complexity and to provide robustness under sampling noise, Forrow et al. (2019) introduced another variant of OT, optimizing a k-Wasserstein Barycenter proxy for the rank-constrained Wasserstein objective. Their approach factors the transport plan through a small set of anchor points called hubs. Generalizing this approach, Scetbon et al. (2021) introduce the factorization $\boldsymbol{P} = \boldsymbol{Q}\,\mathrm{diag}\,(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}$ comprised of *sub-coupling* matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ sharing an *inner marginal* $\boldsymbol{g}$, meaning $\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{g}$. Building on this, Scetbon et al. (2021, 2022, 2023) derived algorithms to compute low-rank optimal transport plans for the primal OT problem with general costs, extending low-rank OT to GW and unbalanced problems using factored couplings.

Interestingly, a different factorization of $\boldsymbol{P}$ was proposed by Lin et al. (2021) in the context of k-Wasserstein barycenters. We call their factorization a *latent coupling* (LC) factorization, given by $\boldsymbol{P} = \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}$, with *two* inner marginals $\boldsymbol{g}_Q = \boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n$ and $\boldsymbol{g}_R = \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_n$ and a general coupling $\boldsymbol{T}$. Lin et al. (2021) constrain the transport between $\boldsymbol{a}$ and $\boldsymbol{b}$ through two sets of learned anchor points, where the factorization is defined by three transport plans computed from three cost matrices between the points and their anchors. This objective differs from that of Forrow et al. (2019); Scetbon et al. (2021), who seek a minimal rank coupling with respect to a single, fixed cost $\boldsymbol{C}$. We observe that factored couplings of Forrow et al. (2019) correspond to LC factorizations with diagonal $\boldsymbol{T}$, suggesting the LC factorization of Lin et al. (2021) may provide an alternative parameterization of transport plans for the low-rank OT problem considered in Forrow et al. (2019); Scetbon et al. (2021). To our knowledge, this idea has not yet been explored.

Contributions. We present a new algorithm, Factor Relaxation with Latent Coupling (FRLC, with the informal mnemonic "frolic"), to compute a minimum cost low-rank transport plan using the LC factorization. Parameterizing low-rank transport plans with the LC factorization has a number of advantages. First, optimization of the low-rank OT objective decouples into three OT sub-problems on the LC factors $\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}$, leading to a simpler optimization algorithm. Second, this decoupling provides straightfoward extensions of FRLC to low-rank unbalanced and semi-relaxed OT; similar extensions for factored couplings required additional work Scetbon et al. (2023) beyond the balanced case. Third, the latent coupling $\boldsymbol{T}$ in the LC factorization provides additional flexibility to model transport between datasets with different numbers of clusters, and to model mass-splitting between these clusters, providing a high-level and interpretable description of $\boldsymbol{P}$ that differs from the factored couplings of Forrow et al. (2019). FRLC computes the LC factorization using a novel *coordinate* mirror descent scheme, alternating descent steps on variables $(\boldsymbol{Q}, \boldsymbol{R})$ and $\boldsymbol{T}$, inspired by the mirror descent approach of Scetbon et al. (2021). We call the descent step on $(\boldsymbol{Q}, \boldsymbol{R})$ *factor relaxation*, as the factors $\boldsymbol{Q}$ and $\boldsymbol{R}$ have relaxed inner marginals, allowing FRLC to be solved by OT sub-problems. FRLC handles multiple OT objectives (Wasserstein, Gromov-Wasserstein, Fused Gromov-Wasserstein), and marginal constraints (balanced, unbalanced, and semi-relaxed). We show FRLC performs better than existing state-of-the-art low-rank methods on a range of synthetic and real datasets, retaining the interpretability of Lin et al. (2021), and inheriting the broad applicability of Scetbon et al. (2021); Scetbon & Cuturi (2022); Scetbon et al. (2022, 2023).

2 Background

Wasserstein OT. Let $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots y_m\}$ be datasets in a metric space $\mathcal{X}$, and let Δ_d be the probability simplex of size d. Through probability vectors $\boldsymbol{a} \in \Delta_n$ and $\boldsymbol{b} \in \Delta_m$, each dataset is encoded as a probability measure: $\mu = \sum_{i=1}^n \boldsymbol{a}_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m \boldsymbol{b}_j \delta_{y_j}$. Let

$$\Pi_{\boldsymbol{a},\cdot} := \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} : \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}\},\ \Pi_{\cdot,\boldsymbol{b}} := \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} : \boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{b}\},\ \Pi_{\boldsymbol{a},\boldsymbol{b}} := \Pi_{\boldsymbol{a},\cdot} \cap \Pi_{\cdot,\boldsymbol{b}}.$$

Thus, $\Pi_{\boldsymbol{a},\boldsymbol{b}}$ is the set of transport plans (probabilistic coupling matrices) with marginals $\boldsymbol{a}$ and $\boldsymbol{b}$. Given a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, define the cost matrix $\boldsymbol{C} \in \mathbb{R}_+^{n\times m}$ via $\boldsymbol{C}_{ij} = c(x_i, y_j)$. The Kantorovich formulation Kantorovich (1942) of discrete OT, also called the Wasserstein problem, seeks a transport plan $\boldsymbol{P}$ of minimal cost :

$$\mathrm{W}(\mu, \nu) := \min_{\boldsymbol{P}\in\Pi_{\boldsymbol{a},\boldsymbol{b}}} \langle \boldsymbol{C}, \boldsymbol{P}\rangle_F. \tag{1}$$

Gromov-Wasserstein OT. In many applications, one wishes to compare datasets $\{x_1, \dots, x_n\} \subset \mathcal{X}$ and $\{y_1, \dots, y_m\} \subset \mathcal{Y}$ across distinct metric spaces $\mathcal{X}$ and $\mathcal{Y}$. The Gromov-Wasserstein (GW) objective Mémoli (2007, 2011) addresses the absence of a common metric or coordinate system through intra-domain cost functions $c_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and $c_2 : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, leading to intra-domain cost matrices $\boldsymbol{A}_{ik} = c_1(x_i, x_k)$ and $\boldsymbol{B}_{jl} = c_2(y_j, y_l)$. The GW objective function $\mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P}) := \sum_{i,j,k,l}(\boldsymbol{A}_{ik} - \boldsymbol{B}_{jl})^2 \boldsymbol{P}_{ij}\boldsymbol{P}_{kl}$ quantifies the expected metric distortion under $\boldsymbol{P}$, leading to the optimization problem:

$$\mathrm{GW}(\mu, \nu) := \min_{\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P}). \tag{2}$$

The Fused Gromov-Wasserstein (FGW) objective function Vayer et al. (2020) is a convex combination of the W and GW objectives, given as $\alpha\langle \boldsymbol{C}, \boldsymbol{P}\rangle_F + (1-\alpha)\mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P})$, for hyperparameter $\alpha \in (0, 1)$.

Relaxed marginal constraints. *Balanced* OT (1) constrains $\boldsymbol{P}$ to lie in $\Pi_{\boldsymbol{a},\boldsymbol{b}}$. *Unbalanced* OT relaxes constraints $\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}$ and $\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{b}$, replacing them with penalties in the form of KL divergences (or other divergences, see Chizat et al. (2018)):

$$\text{U-W}(\mu, \nu) := \min_{\boldsymbol{P} \in \mathbb{R}_+^{n \times m}} \langle \boldsymbol{C}, \boldsymbol{P}\rangle_F + \tau_L \mathrm{KL}(\boldsymbol{P}\mathbf{1}_m \| \boldsymbol{a}) + \tau_R \mathrm{KL}(\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n \| \boldsymbol{b}), \tag{3}$$

where $\tau_L, \tau_R > 0$ control the strength of each penalty. *Semi-relaxed* optimal transport relaxes exactly one of the hard constraints $\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}$ and $\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{b}$ in the same manner. The semi-relaxed version of (1) obtained by relaxing only the "right" marginal constraint on $\boldsymbol{b}$ is:

$$\text{SR}^{\mathrm{R}}\text{-W}(\mu, \nu) := \min_{\boldsymbol{P} \in \Pi_{\boldsymbol{a},\cdot}} \langle \boldsymbol{C}, \boldsymbol{P}\rangle + \tau \mathrm{KL}(\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n \| \boldsymbol{b}), \tag{4}$$

while it's "left" marginal counterpart $\text{SR}^{\mathrm{L}}\text{-W}(\mu, \nu)$ is defined analogously over $\boldsymbol{P} \in \Pi_{\cdot,\boldsymbol{b}}$, using penalty $\tau \mathrm{KL}(\boldsymbol{P}\mathbf{1}_m \| \boldsymbol{a})$. Likewise, one can form semi-relaxed or unbalanced GW and FGW problems.

Entropy regularization. The seminal work Cuturi (2013b) introduced the Sinkhorn algorithm to solve an entropy regularized version of (1), $\mathrm{W}_\epsilon(\mu, \nu) := \min_{\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}} \langle \boldsymbol{C}, \boldsymbol{P}\rangle_F - \epsilon H(\boldsymbol{P})$, massively improving the $O(n^3 \log n)$ time complexity of classical techniques Orlin (1997); Tarjan (1997). Above, H is the entropy, $H(\boldsymbol{P}) = -\sum_{ij} \boldsymbol{P}_{ij}(\log \boldsymbol{P}_{ij} - 1)$, and $\epsilon > 0$ is the regularization strength.

Low-rank regularization. The nonnegative rank $\mathrm{rk}_+(\boldsymbol{M})$ of matrix $\boldsymbol{M}$ is the least number of nonnegative rank-one matrices summing to $\boldsymbol{M}$. For $r \geq 1$, define

$$\Pi_{\boldsymbol{a},\cdot}(r) = \{\boldsymbol{P} \in \Pi_{\boldsymbol{a},\cdot} : \mathrm{rk}_+(\boldsymbol{P}) \leq r\}, \quad \Pi_{\cdot,\boldsymbol{b}}(r) = \{\boldsymbol{P} \in \Pi_{\cdot,\boldsymbol{b}} : \mathrm{rk}_+(\boldsymbol{P}) \leq r\}, \tag{5}$$

and let $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r) = \Pi_{\boldsymbol{a},\cdot}(r) \cap \Pi_{\cdot,\boldsymbol{b}}(r)$. To estimate Wasserstein distances with greater stability and accuracy under sampling noise, Forrow et al. (2019) proposed a low-rank regularization on the coupling matrix, factoring the transport through a small set of anchor points. More explicitly, Scetbon et al. (2021) parameterized the set as $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ through the set $\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ of *factored couplings*,

$$\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}) \in \mathbb{R}_+^{n \times r} \times \mathbb{R}_+^{m \times r} \times (\mathbb{R}_+^*)^r : \boldsymbol{Q} \in \Pi_{\boldsymbol{a},\boldsymbol{g}}, \boldsymbol{R} \in \Pi_{\boldsymbol{b},\boldsymbol{g}}\}.$$

The set $\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ parameterizes $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ through $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}) \mapsto \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}$, as shown by Cohen & Rothblum (1993).

Scetbon et al. (2021) apply this factorization to solve the Wasserstein problem subject to $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ for general cost matrices:

$$\mathrm{W}_r(\mu, \nu) := \min_{\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)} \langle \boldsymbol{C}, \boldsymbol{P}\rangle_F \tag{6}$$

GW, unbalanced and semi-relaxed low-rank OT problems are defined as in (2), (3) and (4), replacing $\mathbb{R}_+^{n \times m}$, $\Pi_{\boldsymbol{a},\cdot}$, or $\Pi_{\cdot,\boldsymbol{b}}$ with rank-constrained counterparts (5). Scetbon & Cuturi (2022); Scetbon et al. (2022, 2023) developed a robust framework for solving all of these problems.

3 Factor Relaxation with Latent Coupling (FRLC) algorithm

3.1 Latent Coupling Factorization

We parameterize low-rank coupling matrices $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ using a factorization introduced in Lin et al. (2021), which we call the *latent coupling (LC) factorization* (Fig. 1). The key property of

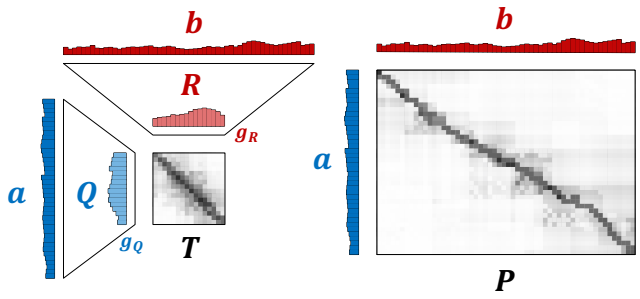


Figure 1: (Left) The LC factorization $\boldsymbol{P} = \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}$ of coupling matrix $\boldsymbol{P}$ with outer marginals $\boldsymbol{a}, \boldsymbol{b}$, inner marginals $\boldsymbol{g}_Q, \boldsymbol{g}_R$, factors $\boldsymbol{Q}, \boldsymbol{R}$, and latent coupling $\boldsymbol{T}$. (Right) Full-rank coupling matrix $\boldsymbol{P}$.

this factorization is the presence of a coupling matrix $\boldsymbol{T}$ linking two distinct inner marginals. For simplicity we describe this factorization using an r-dimensional latent space, but we also extend to non-square matrices linking two latent spaces of different dimensions, as demonstrated in the results.

Definition 3.1 (Inner marginals)**.** Given a factorization $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$ of a coupling matrix $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$, the *inner marginals* of $\boldsymbol{Q}$ and $\boldsymbol{R}$ are $\boldsymbol{g}_Q := \boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n$ and $\boldsymbol{g}_R := \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m$, respectively, where $\boldsymbol{g}_Q, \boldsymbol{g}_R \in \Delta_r$.

To distinguish the different marginals, we refer to $\boldsymbol{a}$ and $\boldsymbol{b}$ as *outer marginals*.

Definition 3.2 (LC factorization)**.** Given a coupling matrix $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$, a *latent coupling (LC) factorization of* $\boldsymbol{P}$ is $\boldsymbol{P} = \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}$, where $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ are the inner marginals of $\boldsymbol{Q}$ and $\boldsymbol{R}$, $\boldsymbol{Q} \in \Pi_{\boldsymbol{a},\cdot}$, $\boldsymbol{R} \in \Pi_{\boldsymbol{b},\cdot}$, and $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}$.

We call the factors $\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}$ in an LC factorization *sub-couplings*. Let $\mathcal{R}_+ := \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{r\times r}$. Given probability vectors $\boldsymbol{a} \in \Delta_n$, $\boldsymbol{b} \in \Delta_m$ and a positive integer rank r, let

$$\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{Q} \in \Pi_{\boldsymbol{a},\cdot}, \boldsymbol{R} \in \Pi_{\boldsymbol{b},\cdot}, \boldsymbol{T} \in \Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}\},$$

be the set of admissible sub-couplings for the LC factorization. Definition 3.2 gives the following map from $\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ to $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$:

$$(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \mapsto \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}} =: \boldsymbol{P}_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})}. \tag{7}$$

Since this map is surjective, the set $\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ parameterizes $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$. Surjectivity follows from the fact that $\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ maps injectively into $\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)$, through $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}) \mapsto (\boldsymbol{Q}, \boldsymbol{R}, \mathrm{diag}(\boldsymbol{g}))$, and $\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ maps surjectively onto $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ via $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}) \mapsto \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}$. Definition 3.1 and Definition 3.2 are readily extended in two directions: the case when the outer marginal constraints are relaxed such that $\boldsymbol{Q} \in \mathbb{R}_+^{n\times m}$ or $\boldsymbol{R} \in \mathbb{R}_+^{n\times m}$, while maintaining the constraint that $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}$; as well as the case of non-square $\boldsymbol{T}$.

3.2 The Balanced FRLC Algorithm

We introduce an algorithm Factor Relaxation with Latent Coupling (FRLC), to compute a LC factorization of minimum cost. We first describe the FRLC algorithm for the balanced Wasserstein problem. Extensions to other and marginal constraints are discussed later. The FRLC objective function, for low-rank, balanced Wasserstein OT, is

$$\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) := \langle \boldsymbol{C}, \boldsymbol{P}_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})}\rangle_F, \tag{8}$$

where $P_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})}$ is defined by (7). Since $\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)$ parameterizes $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$, problem (8) is equivalent to low rank problem (6). The FRLC algorithm is built from projections onto convex sets, described by constraints on the outer marginals alone for $(\boldsymbol{Q}, \boldsymbol{R})$ and by the inner marginals alone for $\boldsymbol{T}$. Given $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)$, sub-couplings $\boldsymbol{Q}$ and $\boldsymbol{R}$ are constrained by:

$$\mathcal{C}_1(\boldsymbol{a}) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{Q}\mathbf{1}_r = \boldsymbol{a}\}, \quad \mathcal{C}_1(\boldsymbol{b}) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}\}.$$

The convex sets constraining the latent coupling matrix $\boldsymbol{T}$ are

$$\mathcal{C}_2(\boldsymbol{g}_Q) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{T}\mathbf{1}_r = \boldsymbol{g}_Q\}, \quad \mathcal{C}_2(\boldsymbol{g}_R) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \boldsymbol{g}_R\},$$

where $\boldsymbol{g}_Q = \boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n$ and $\boldsymbol{g}_R = \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m$ as per Definition 3.1. Writing $\mathcal{C}_1 = \mathcal{C}_1(\boldsymbol{a}) \cap \mathcal{C}_1(\boldsymbol{b})$ and $\mathcal{C}_2 = \mathcal{C}_2(\boldsymbol{g}_Q) \cap \mathcal{C}_2(\boldsymbol{g}_R)$, one has $\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r) = \mathcal{C}_1 \cap \mathcal{C}_2$.

We use *coordinate* mirror descent to optimize (8), building on the mirror descent (MD) approach of Scetbon et al. (2021); Scetbon & Cuturi (2022); Scetbon et al. (2022, 2023) for the low-rank problem. First we take a descent step in the variables $(\boldsymbol{Q}, \boldsymbol{R})$ for a fixed $\boldsymbol{T}$, using KL penalties on their inner marginals. These "soft" constraints allow the joint optimization in $(\boldsymbol{Q}, \boldsymbol{R})$ to decouple into two semi-relaxed OT problems, one for each variable. We call this step *factor relaxation* as this allows $(\boldsymbol{Q}, \boldsymbol{R})$ to have relaxed inner marginals $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$. Next we take a descent step in the latent coupling variable $\boldsymbol{T}$, fixing the $\boldsymbol{Q}$ and $\boldsymbol{R}$, equivalent to solving a balanced OT problem. Thus, solving both coordinate descent steps corresponds to solving three OT problems.

We now provide further details on these coordinate descent steps, with the full algorithm given in Algorithm 1. Let $(\gamma_k)_{k=1}^N$ be a sequence of step sizes. As in Scetbon & Cuturi (2022), we choose ℓ^∞-normalization for the step-sizes. Our coordinate mirror descent in the factor relaxation step is:

$$\begin{aligned}(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}) \leftarrow \underset{(\boldsymbol{Q},\boldsymbol{R}) \,:\, (\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}_k)\in\mathcal{C}_1}{\arg\min} & \langle (\boldsymbol{Q}, \boldsymbol{R}), \nabla_{(\boldsymbol{Q},\boldsymbol{R})}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}((\boldsymbol{Q}, \boldsymbol{R})\|(\boldsymbol{Q}_k, \boldsymbol{R}_k)) \\ & + \tau\mathrm{KL}((\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m)\|(\boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m))\end{aligned}$$

The Sinkhorn kernels for the semi-relaxed OT problems arising from the factor relaxation step are:

$$\begin{aligned}\boldsymbol{K}_{\boldsymbol{Q}}^{(k)} &:= \boldsymbol{Q}_k \odot \exp(\,-\gamma_k(\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}} - \mathbf{1}_n\mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{Q_k}))^{\mathrm{T}})\,) \\ \boldsymbol{K}_{\boldsymbol{R}}^{(k)} &:= \boldsymbol{R}_k \odot \exp(\,-\gamma_k(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k - \mathbf{1}_m\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_{R_k})\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k)^{\mathrm{T}})\,),\end{aligned}$$

introducing the shorthand $\boldsymbol{X} = \mathrm{diag}\,(1/\boldsymbol{g}_Q)\,\boldsymbol{T}\,\mathrm{diag}\,(1/\boldsymbol{g}_R)$ and where $\mathrm{diag}^{-1}(\cdot) : \mathbb{R}^{r\times r} \to \mathbb{R}^r$ denotes the matrix-to-vector extraction of the diagonal. This τ-dependent regularization also allows us to show smoothness of the objective in Proposition E.5, from which the convergence guarantee Proposition 3.3 follows. We derive the semi-relaxed projection Algorithm 2 of the sub-couplings $\boldsymbol{Q}$ and $\boldsymbol{R}$ in Appendix G for completeness. We also show in Lemma A.1 that $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ induced by the semi-relaxed projection are both feasible and locally optimal, not requiring separate optimization.

As $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}) \in \mathcal{C}_2$ if and only if $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_{Q_{k+1}},\boldsymbol{g}_{R_{k+1}}}$, after the factor relaxation step, we next take a coordinate MD step on the latent coupling $\boldsymbol{T}$:

$$\boldsymbol{T}_{k+1} \leftarrow \underset{\boldsymbol{T}\,:\,(\boldsymbol{Q}_{k+1},\boldsymbol{R}_{k+1},\boldsymbol{T})\in\mathcal{C}_2}{\arg\min} \langle \boldsymbol{T}, \nabla_{\boldsymbol{T}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{T}\|\boldsymbol{T}_k). \tag{9}$$

This is equivalent to applying Sinkhorn (Algorithm 5) to $\boldsymbol{T}$ given $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ with the kernel:

$$\boldsymbol{K}_{\boldsymbol{T}}^{(k)} := \mathbf{T}_k \odot \exp(\,-\gamma_k\mathrm{diag}(1/\boldsymbol{g}_{Q_{k+1}})\boldsymbol{Q}_{k+1}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{R_{k+1}})\,).$$

After the final iteration of the coordinate-MD scheme, $\boldsymbol{X} = \mathrm{diag}\,(1/\boldsymbol{g}_Q)\,\boldsymbol{T}\,\mathrm{diag}\,(1/\boldsymbol{g}_R)$ satisfies $\boldsymbol{X}\boldsymbol{g}_R = \mathbf{1}_r$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{g}_Q = \mathbf{1}_r$ as $\boldsymbol{T}$ is a coupling between $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$. Thus $\boldsymbol{P}_r = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}$ and the iterates $(\boldsymbol{Q}_k, \boldsymbol{T}_k, \boldsymbol{R}_k)$ remain in the intersection of the constraint sets. Thus, in contrast to other approaches Scetbon et al. (2021); Lin et al. (2021); Forrow et al. (2019), we do not require Dykstra projections back into the intersection to maintain feasability. We note that our implementation of FRLC allows for a non-square latent coupling $\boldsymbol{T}$, providing greater interpretability in problem-specific applications. Above, we presented FRLC in the simplest case that $\boldsymbol{T}$ is square.

3.3 Initialization, convergence, and FRLC extensions

Full-rank random initializations of the sub-coupling matrices. We propose a new initialization of the sub-couplings $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ for the LC-factorization in Algorithm 6.This generates a full-rank initialization (Proposition F.1) in the set of rank-r couplings $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ and is accomplished by applying Sinkhorn to random matrices. Our approach differs from Scetbon et al. (2021); Scetbon & Cuturi (2022) who use initializations for the diagonal factorization of Forrow et al. (2019), and are not applicable to a latent coupling that is non-diagonal, non-square, or with two distinct inner marginals.

Algorithm 1 Balanced FRLC

Input $\boldsymbol{C}, r, \boldsymbol{a}, \boldsymbol{b}, \tau, \gamma, \delta, \varepsilon$
Initialize $\boldsymbol{g}_Q, \boldsymbol{g}_R = \frac{1}{r}\mathbf{1}_r$
$\boldsymbol{Q}_0, \boldsymbol{R}_0, \boldsymbol{T}_0 \leftarrow$ *Initialize-Couplings*$(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{g}_Q, \boldsymbol{g}_R)$ # Alg. 6
$\boldsymbol{X}_0 \leftarrow \mathrm{diag}(1/\boldsymbol{Q}_0^{\mathrm{T}}\mathbf{1}_n)\boldsymbol{T}_0\,\mathrm{diag}(1/\boldsymbol{R}_0^{\mathrm{T}}\mathbf{1}_m)$
while $\Delta((\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k), (\boldsymbol{Q}_{k-1}, \boldsymbol{R}_{k-1}, \boldsymbol{T}_{k-1})) > \varepsilon$ **do** # Δ as in (10)
  $\nabla_{\boldsymbol{Q}} \leftarrow \boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}} - \mathbf{1}_n\mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}}$
  $\nabla_{\boldsymbol{R}} \leftarrow \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k - \mathbf{1}_m\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k)^{\mathrm{T}}$
  $\gamma_k \leftarrow \gamma/\max\{\|\nabla_{\boldsymbol{Q}}\|_\infty, \|\nabla_{\boldsymbol{R}}\|_\infty\}$ # ℓ^∞-normalization of Scetbon & Cuturi (2022)
  $\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \boldsymbol{K}_{\boldsymbol{R}}^{(k)} \leftarrow \boldsymbol{Q}_k \odot \exp(-\gamma_k\nabla_{\boldsymbol{Q}}), \boldsymbol{R}_k \odot \exp(-\gamma_k\nabla_{\boldsymbol{R}})$
  $\boldsymbol{Q}_k \leftarrow$ SR^{R}*-projection*$(\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \gamma_k, \tau, \boldsymbol{a}, \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n, \delta)$ # Semi-relaxed OT, Alg. 2
  $\boldsymbol{R}_k \leftarrow \mathrm{SR}^{\mathrm{R}}$*-projection*$(\boldsymbol{K}_{\boldsymbol{R}}^{(k)}, \gamma_k, \tau, \boldsymbol{b}, \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m, \delta)$ # Semi-relaxed OT
  $\boldsymbol{g}_Q, \boldsymbol{g}_R \leftarrow \boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m$
  $\nabla_{\boldsymbol{T}} = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}_k\,\mathrm{diag}(1/\boldsymbol{g}_R)$
  $\gamma_{\boldsymbol{T}} = \gamma/\|\nabla_{\boldsymbol{T}}\|_\infty$ # ℓ^∞-normalization
  $\boldsymbol{K}_{\boldsymbol{T}}^{(k)} \leftarrow \boldsymbol{T}_k \odot \exp(-\gamma_{\boldsymbol{T}}\nabla_{\boldsymbol{T}})$
  $\boldsymbol{T}_k \leftarrow$ *Sinkhorn*$(\boldsymbol{K}_{\boldsymbol{T}}^{(k)}, \boldsymbol{g}_R, \boldsymbol{g}_Q, \delta)$ # Balanced OT, Alg. 5
  $\boldsymbol{X}_k \leftarrow \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)$
end while
Return $\boldsymbol{P}_r = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$

Convergence analysis of FRLC. As objective (8) is non-convex, it is important to have convergence guarantees. Our convergence criterion $\Delta(\cdot,\cdot)$ is defined in (10). To prove convergence we require a lower bound on the entries of $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$. Previous works introduce a lower-bound vector $\boldsymbol{\alpha} \leq \boldsymbol{g}$ enforced element-wise for stability and smoothness Scetbon et al. (2021). In FRLC the use of semi-relaxed projections naturally enforces a lower-bound. In Appendix E.5, we show that for any $\delta \in (0, \frac{1}{r})$, the FRLC algorithm's τ-weighted regularization on the inner marginals can guarantee a uniform lower-bound of δ on the entries: for sufficiently large τ and $\tilde{O}(m^2/\epsilon)$ iterations for the sub-coupling Pham et al. (2020), one guarantees a lower bound of δ on $\boldsymbol{g}_R$ and $\boldsymbol{g}_Q$. This allows us to show objective smoothness in Proposition E.5. Previous work on low-rank optimal transport Scetbon et al. (2021) use the non-asymptotic convergence criterion of Ghadimi et al. (2014). Following existing works Dang & Lan (2015) establishing convergence rates of coordinate mirror-descent for smooth objectives, we show in Proposition 3.3 this criterion may be extended to coordinate-MD by adapting the block-descent lemma of Beck & Tetruashvili (2013).

Proposition 3.3. *Suppose one has $f \in C^1(\mathcal{X}, \mathbb{R})$ with block-coordinate Lipschitz gradient and block smoothness constants $(L_i)_{i=1}^p$, and a function $h \in C(\mathcal{X}, \mathbb{R})$ which is α-strongly convex. For $\Phi = f + h$, suppose one performs coordinate mirror descent on Φ minimized over a product of closed convex sets $\mathcal{X} = \prod_{i=1}^p \mathcal{X}_i$. Let the sub-iterates with respect to the i-th block update be $\{\mathbf{x}_k^i\}_{i=0}^p$ where $\mathbf{x}_k := \mathbf{x}_k^0$ for $k \in [N]$ outer iterations. Then one has:*

$$\min_k \Delta(\mathbf{x}_k, \mathbf{x}_{k-1}) \leq \frac{D^2 L}{N(\alpha^2/2L)} = \frac{2D^2L^2}{N\alpha^2},$$

where D is (36)*, L is the global smoothness constant, stepsizes $\gamma_{k,i} := \alpha/L$, and convergence criterion $\Delta(\mathbf{x}_k, \mathbf{x}_{k-1})$ is given in* (35)*.*

Specialized to the LC-parametrization, the criterion $\Delta_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1})$ is:

$$\Delta_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) := \frac{1}{\gamma_k^2}\left[\|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k\|_F^2 + \|\boldsymbol{R}_{k+1} - \boldsymbol{R}_k\|_F^2 + \|\boldsymbol{T}_{k+1} - \boldsymbol{T}_k\|_F^2\right] \tag{10}$$

for $\boldsymbol{x}_k = (\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k)$. We show through Propositions 3.3, E.5 the following result:

Proposition 3.4. *The FRLC algorithm with step-sizes $\gamma_k = \alpha/L$ and iterates $\boldsymbol{x}_k = (\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k)$ has non-asymptotic stationary convergence in the criterion $\Delta(\cdot,\cdot)$ with:*

$$\min_{k \in 1,..,N-1} \Delta_k(\boldsymbol{x}_k, \boldsymbol{x}_{k+1}) \leq 2D^2L^2/N\alpha^2$$

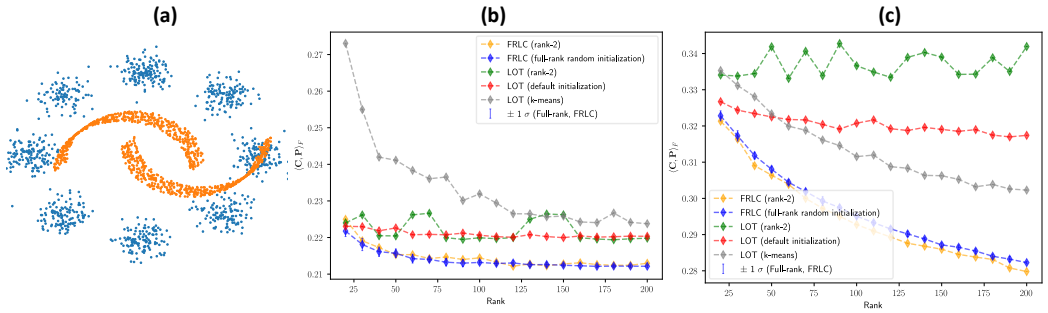


Figure 2: (a) Simulated dataset containing points from two moons (orange) and eight Gaussians (blue). (b) Transport cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ achieved by FRLC and LOT Scetbon et al. (2021) for the balanced Wasserstein problem on the dataset in (a) for different ranks and initializations. FLRC full rank (blue curve) is average over 10 random initializations. (c) Results on the 10D mixture of Gaussians dataset.

Where N is the number of iterations, D the optimality-gap as in (36)*, and $L = \max_{i \in \{1,..,3\}}(L_i)$ the global smoothness for $L_i = \text{poly}(\|\boldsymbol{C}\|_F, n, m, r, \delta)$ the block-wise smoothness constants.*

The proof of Proposition 3.4 follows directly from our extension of the non-asymptotic criterion with the block-descent result Proposition 3.3 and the proof that this lemma holds in FRLC Proposition E.5. We also mention two improvements to other low rank approximation results in literature. In Proposition F.2 we show that one can *analytically* solve for the block-optimal $\boldsymbol{g}$ for the factorization of Scetbon et al. (2021), and we improve the bound on the low-rank approximation error in Proposition E.7.

FRLC for other marginal constraints and objectives. The balanced FRLC algorithm can be extended simply to other marginal constraints owing to the decoupling of the coordinate MD scheme. In particular, by using either the semi-relaxed projections (Algorithm 2) or fully-relaxed (unbalanced) projections (Algorithm 3) on sub-couplings $\boldsymbol{Q}$ and $\boldsymbol{R}$, one can solve the balanced problem, the problem with the left or right marginal relaxed, or the unbalanced problem. As such, *all* variants of marginal constraints can be handled by a single algorithm, given in Algorithm 4.

We also extend the FRLC algorithm to the Gromov-Wasserstein problem. This consists of computing a GW-specific gradient with the appropriate marginal constraints applied to simplify their form, and re-computing Sinkhorn kernels as exponentiations of these gradients. The matrix form of the quadratic GW objective is $\mathbf{1}_m^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{A}^{\odot 2} \boldsymbol{P} \mathbf{1}_m + \mathbf{1}_n^{\mathrm{T}} \boldsymbol{P} \boldsymbol{B}^{\odot 2} \boldsymbol{P}^{\mathrm{T}} \mathbf{1}_n - 2\langle \boldsymbol{A}\boldsymbol{P}\boldsymbol{B}, \boldsymbol{P} \rangle$, where $\odot$ denotes the Hadamard (entrywise) product. Then the GW-specific Sinkhorn kernels are

$$
\begin{aligned}
\boldsymbol{K}_{\boldsymbol{Q}}^{(k)} &\leftarrow \exp\left(2\gamma_k(2\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}} - \boldsymbol{A}^{\odot 2}\boldsymbol{Q}\mathbf{1}_r\mathbf{1}_r^{\mathrm{T}})\right), \\
\boldsymbol{K}_{\boldsymbol{R}}^{(k)} &\leftarrow \exp\left(2\gamma_k(2\boldsymbol{B}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X} - \boldsymbol{B}^{\odot 2}\boldsymbol{R}\mathbf{1}_r\mathbf{1}_r^{\mathrm{T}})\right), \\
\boldsymbol{K}_{\boldsymbol{T}}^{(k)} &\leftarrow \exp(4\gamma_k \operatorname{diag}(\boldsymbol{g}_Q^{-1})\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}\operatorname{diag}(\boldsymbol{g}_R^{-1})).
\end{aligned}
$$

In Algorithm 4, one can solve the GW-problem by using the kernels above. Here, we present the kernels omitting a rank-1 perturbation, which is given in Appendix D. From the Wasserstein and GW gradients, the FGW gradient is easily taken as a convex combination of the two. In this work, we primarily focus on the LC-factorization for the rank r Wasserstein problem (6).

4 Experimental Results

We compare FRLC to existing low-rank and full-rank optimal transport algorithms on several datasets: simulated datasets previously used in Tong et al. (2023) and Scetbon et al. (2021); a massive spatial-transcriptomics dataset Chen et al. (2022); and a graph partitioning task Chowdhury & Needham (2021). Further details of each experiment (e.g. pre-processing, validation) are in Appendices K, L, and M. In the section below, LOT refers to the works of Scetbon et al. (2021, 2023, 2022) and Latent OT refers to Lin et al. (2021).

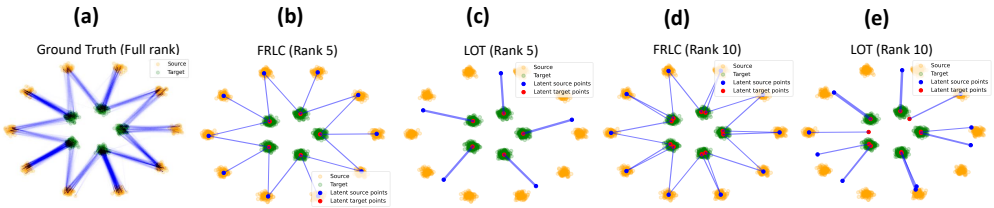


Figure 3: LC-projections of couplings of Gaussians centered on the 5^{th}-roots of unity (green) and 10^{th} roots of unity (yellow). (a) Ground-truth full-rank coupling. (b) Non-square rank-5 latent-coupling of FRLC (c) LC-projection barycenters aligned with rank-5 diagonal coupling of LOT Scetbon et al. (2021). (d) Square rank-10 latent coupling of FRLC. (e) Rank-10 diagonal coupling of LOT .

4.1 Evaluation of Low-rank Approximations for Balanced OT on Synthetic Data

We first compare the balanced OT version of FRLC with the the low-rank balanced OT algorithm LOT of Scetbon et al. (2021) on a synthetic dataset following Tong et al. (2023). The dataset consists of $m = 1000$ points from two moons and $n = 1000$ points sampled from eight 2D Gaussian densities (Fig. 2a). We solve the Wasserstein problem (1) with cost matrix $\boldsymbol{C}$ computed using the Euclidean distance. The full-rank coupling matrix $\boldsymbol{P}$ has rank 1000, and we compute both FRLC and LOT solutions with rank between 20 and 200. For each rank, we initialize FRLC adapting the deterministic rank-2 initialization proposed in Scetbon et al. (2021) and the random initialization of Alg. 6. We initialize LOT using the rank-2 initialization and two other options in `ott-jax` Cuturi et al. (2022).

We find that FRLC obtains lower transport cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ with increasing rank (Fig. 2b) and consistently achieves lower transport cost than LOT across all ranks and all initializations. Specifically, starting both methods at the same rank-2 initialization, FRLC consistently achieves a lower cost than LOT for all ranks. Additionally, we observe smooth convergence of FRLC for both rank-2 initialization and the full-rank random initialization of Alg. 6 (Fig. 5).

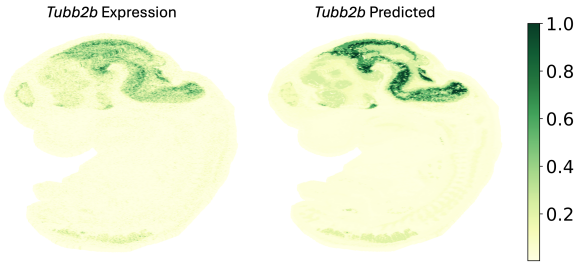
We also evaluate FRLC and LOT on two datasets of Gaussian mixtures, one in 2-dimensions and one in 10-dimensions, each with $n = m = 5,000$ points from two mixtures of Gaussians, following Scetbon et al. (2021), with further details in Appendix K. We observe the same trend as the previous simulation for both datasets (Fig. 2c, Fig. 7), with FRLC achieving lower transport costs than LOT across all ranks and all initializations. In addition FRLC has half the runtime of LOT (CPU) – including the setup time of FRLC but excluding the setup time of LOT in `ott-jax` – on datasets of $n = m = 1000$ points from all three datasets with rank $r = 100$ (Table 2). At the same time FRLC achieves lower primal cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ with tighter marginals $\|\boldsymbol{P}\mathbf{1}_n - \boldsymbol{a}\|_2$ and $\|\boldsymbol{P}^T\mathbf{1}_m - \boldsymbol{b}\|_2$. Lin et al. (2021) only solves a proxy for the rank-constrained Wasserstein problem, and thus is not the focus of our comparisons. Nevertheless, we verify that on all synthetic experiments that FRLC achieves significantly lower primal OT cost than Latent OT (Table 5).

4.2 Interpretation of the Latent Coupling and LC-Projection

We demonstrate the intepretability of the latent coupling $\boldsymbol{T}$ in the LC factorization. In both the LC factorization and factored couplings, the sub-couplings $\boldsymbol{Q}$ and $\boldsymbol{R}$ each have associated barycentric projection operators which coarse-grain input datasets $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}$. In particular, the LC projection is defined from the LC factorization as follows.

Definition 4.1 (LC-Projection)**.** Let $\boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}$ be an LC factorization of of a coupling matrix $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ computed from datasets $\mathbf{Z}^{(1)} \in \mathbb{R}^{n\times d}, \mathbf{Z}^{(2)} \in \mathbb{R}^{m\times d}$, with $\boldsymbol{T} \in \mathbb{R}_+^{r_1\times r_2}$. The *LC-projections* $\mathbf{Y}^{(1)}$ and $\mathbf{Y}^{(2)}$ of $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ are $\mathbf{Y}^{(1)} := \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{Q}^{\mathrm{T}}\mathbf{Z}^{(1)}$, and $\mathbf{Y}^{(2)} := \mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\mathbf{Z}^{(2)}$.

By interpreting any factored coupling $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g})$ as an LC factorization $(\boldsymbol{Q}, \boldsymbol{R}, \mathrm{diag}(\boldsymbol{g}))$, Definition 4.1 describes the barycentric projections for both factorizations. We compare the projections of the coupling computed by FRLC to those of LOT Scetbon et al. (2021) on a dataset containing 1000 samples from 2D-Gaussians centered at the 5^{th}-roots of unity and 1000 samples from 2D Gaussians



(a)

Objective	Algorithm	Spearman ρ	ARI	AMI
W	LOT-U	0.394	0.332	0.397
	FRLC-U	**0.465**	**0.466**	**0.484**
	LOT-SR	0.391	0.319	0.396
	FRLC-SR	**0.467**	**0.475**	**0.492**
GW	LOT-U	0.005	0.0	0.0
	FRLC-U	**0.266**	**0.299**	**0.364**
	LOT-SR	0.004	0.0	0.0
	FRLC-SR	**0.275**	**0.318**	**0.381**
FGW	LOT-U	**0.391**	0.330	0.400
	FRLC-U	0.368	**0.391**	**0.420**
	LOT-SR	0.396	0.337	0.401
	FRLC-SR	**0.465**	**0.469**	**0.499**

Figure 4: (a) Brain marker gene *Tubb2b* expression and FRLC prediction. (b) Comparison of the low-rank unbalanced (LOT-U) algorithm of Scetbon et al. (2023) and FRLC on aligning spatial transcriptomics data. Bold indicates top performing method for each metric on each objective.

centered at the 10^{th}-roots of unity (the latter scaled by a factor of two, Fig. 3a). In both cases, the latent coupling $\boldsymbol{T}$ or $\mathrm{diag}(\boldsymbol{g})$ is visualized as a transport between barycenters. We run FRLC and LOT with ranks $r = 5$ and $r = 10$ to match the number of target and source clusters. In the rank-5 case, FRLC uses a *non-square* latent coupling $\boldsymbol{T} \in \mathbb{R}_+^{10\times 5}$ which correctly captures the coupling between clusters (Fig. 3(b)), while the LOT rank-5 projection computes barycenters that are outside of the clusters (Fig. 3c). A similar result is observed for square rank-10 latent couplings computed by FRLC (Fig. 3d) and LOT (Fig. 3e) demonstrating that the LOT barycenters in Fig. 3b) are not an artifact of using the lowest rank. We observe similar results on other simulated datasets (Fig. 11).

4.3 Evaluation on Spatial Transcriptomics Alignment

We compare FRLC and the algorithm (LOT-U) of Scetbon et al. (2023) (which solves unbalanced low-rank Wasserstein, GW, and FGW problems) on the task of computing an alignment between cells from different time points during mouse embryonic development. Specifically, we compute an alignment between a spatial transcriptomics (ST) dataset of an E11.5 stage mouse embryo and an E12.5 stage mouse embryo Chen et al. (2022). Optimal transport is a popular approach to align single-cell Schiebinger et al. (2019) and spatial trancriptomics datasets Zeira et al. (2022); Liu et al. (2023); Klein et al. (2023). In single-cell transcriptomics, one measures a gene expression vector for each cell, and in spatial transcriptomics one additionally measures the 2D location of each cell. The cost matrix $\boldsymbol{C}$ describes the difference between gene expression vectors and intra-domain cost matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are derived from the 2D coordinates within each slice. Therefore, OT problems of W, GW, and FGW objectives can be solved and the coupling matrix represents the cell-cell alignment (Appendix M). However, computation of a full-rank OT solution is not feasible in our large-scale dataset: the E11.5 slice has about 30,000 cells while the E12.5 slice has about 50,000 cells.

We evaluate the alignments by assessing performance on two prediction tasks from Scetbon et al. (2023): (1) a *gene expression prediction* task where we predict the expression of a gene in E12.5 from expression of the gene in E11.5 using the alignment; (2) a *cell type prediction task* where we predict the cell types of E12.5 from the cell type clustering of E11.5 (Appendix M). We evaluate the accuracy of the gene expression prediction task through the Spearman correlation ρ between the predicted expression and the ground truth expression of 10 test marker genes. We evaluate the accuracy of the cell type prediction task by computing the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) between the predicted cell types and the cell types derived in the original publication Chen et al. (2022). Being a comparison between different objectives, this relies on downstream metrics. For completeness, we validate the efficacy of FRLC on directly minimizing the balanced Wasserstein cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ against Scetbon et al. (2021) in Figure 8.

For a direct comparison, we use FRLC to solve the same unbalanced problems (denoted FRLC-U). We perform an extensive grid search (Appendix M.3) to pick the best hyperparameters (including rank $\ll 30,000$) for all algorithms. Scetbon et al. (2023) previously showed that unbalanced FGW algorithm has the best performance on ST alignment. We find that unbalanced FRLC achieves comparable or better results than the previous state-of-the-art unbalanced low-rank method on all three objectives (Table 4). We also solve a semi-relaxed version of each problem motivated by the

Method	Factorization	Cost	Variables	Algorithm	Sub-routine for coupling
Factored Coupling Forrow et al. (2019)	Factored coupling	k-Wasserstein barycenter	Anchors & sub-couplings	Lloyd-type	Dykstra's
Latent OT Lin et al. (2021)	Latent coupling	Extension of k-Wasserstein barycenter	Anchors & sub-couplings	Lloyd-type	Dykstra's
LOT Scetbon et al. (2021)	Factored coupling	Primal OT cost	Sub-couplings & inner marginal	Mirror-descent	Dykstra's
FRLC (this work)	Latent coupling	Primal OT cost	Sub-couplings	Coordinate mirror-descent	OT

Table 1: Comparing aspects of low-rank OT methods. Factorization indicates the structure of the inner matrix.

observation that all cells from E12.5 have an ancestor, but not all cells from E11.5 have the same number of descendants due to cell growth and death. Thus the former marginal is tight, and the latter relaxed Halmos et al. (2024). We run both semi-relaxed FRLC (FRLC-SR) and a setting of LOT-U that recovers the semi-relaxed problem (LOT-SR). Semi-relaxed FRLC achieves the best results on all three metrics by a large margin (Table 4). As one example, the expression of *Tubb2b*, a mouse brain marker gene, agreeing with the expression predicted from the semi-relaxed alignment of FRLC (Fig. 4a).

4.4 Additional Experiments

We evaluate FRLC on an unsupervised graph partitioning problem Chowdhury & Needham (2021) on four real-world graph datasets Yang & Leskovec (2012); Yin et al. (2017); Banerjee et al. (2013). We benchmark the performance of the semi-relaxed and GW settings of FRLC against (1) GWL Xu et al. (2019), solving a balanced GW problem; (2) SpecGWL Chowdhury & Needham (2021) using the heat kernel on the graph Laplacian as the cost matrix. We find FRLC achieves the better clustering performance than GWL and SpecGWL on 9/12 and 11/12 of the datasets (Table 3 and Appendix L).

5 Discussion

We provide comparison of existing low-rank solvers in Table 1. The FRLC algorithm has a number of advantages, including (1) coarsening a full-rank plan $\boldsymbol{P}$ to non-diagonal latent coupling $\boldsymbol{T}$; (2) minimizing the primal OT problem for general cost $\boldsymbol{C}$ rather than a barycenteric problem; (3) optimizing only sub-couplings; and (4) using Sinkhorn alone as the sub-routine for low-rank OT. While we argue these are substantial advantages, FRLC has limitations which warrant follow-up work. In particular, three key limitations of our work, common to the existing low-rank OT algorithms, are: (1) selecting values of the latent coupling ranks; (2) strengthening the convergence criterion; (3) addressing sensitivity to the initialization from non-convexity of the objective. A limitation specific to our work is the selection of the τ hyperparameter controlling the smoothness of the trajectory. These and other limitations are discussed in Section N of the Appendix. Another direction for further investigation is to better understand what structure LC factorizations capture when the optimal plan is known to have full rank, e.g. when the Monge map exists, as has been explored by Liu et al. (2021).

6 Conclusion

We introduce FRLC, an algorithm to compute low-rank optimal transport plan from the latent coupling (LC) factorization. FRLC handles different OT objective costs and relaxations of the marginal constraints. Moreover, the LC factorization provides an interpretable coarse-graining of the full transport plan and its marginals through the mapping $(\boldsymbol{P}, \boldsymbol{a}, \boldsymbol{b}) \to (\boldsymbol{T}, \boldsymbol{g}_Q, \boldsymbol{g}_R)$. We demonstrate the superior performance of FRLC compared to state-of-the-art low-rank methods on real and synthetic datasets.

Acknowledgments and Disclosure of Funding

This work is supported by NCI grant U24CA248453 to B.J.R. J.G. gratefully acknowledges support from the Schmidt DataX Fund at Princeton University made possible through a major gift from the Schmidt Futures Foundation.

References

Bakshi, A. and Woodruff, D. Sublinear Time Low-Rank Approximation of Distance Matrices. *Advances in Neural Information Processing Systems*, 31, 2018.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., and Jackson, M. O. The Diffusion of Microfinance. *Science*, 341(6144):1236498, 2013.

Bauschke, H. H. and Lewis, A. S. Dykstras algorithm with Bregman projections: A convergence proof. *Optimization*, 48(4):409–427, January 2000. ISSN 1029-4945. doi: 10.1080/02331930008844513. URL http://dx.doi.org/10.1080/02331930008844513.

Beck, A. and Tetruashvili, L. On the Convergence of Block Coordinate Descent Type Methods. *SIAM J. Optim.*, 23:2037–2060, 2013. URL https://api.semanticscholar.org/CorpusID:6866704.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman Projections for Regularized Transportation Problems. *SIAM Journal on Scientific Computing*, 37 (2):A1111–A1138, January 2015. ISSN 1095-7197. doi: 10.1137/141000439. URL http://dx.doi.org/10.1137/141000439.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, September 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01969-x. URL http://dx.doi.org/10.1038/s41592-023-01969-x.

Charikar, M., Chen, B., Ré, C., and Waingarten, E. Fast Algorithms for a New Relaxation of Optimal Transport. In Neu, G. and Rosasco, L. (eds.), *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4831–4862. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/charikar23a.html.

Chen, A., Liao, S., Cheng, M., Ma, K., Wu, L., Lai, Y., Qiu, X., Yang, J., Xu, J., Hao, S., et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell*, 185(10):1777–1792, 2022.

Chen, X. and Price, E. Condition number-free query and active learning of linear families. 2017.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced Optimal Transport: Dynamic and Kantorovich Formulations. *Journal of Functional Analysis*, 274(11):3090–3123, June 2018. ISSN 0022-1236. doi: 10.1016/j.jfa.2018.03.008. URL http://dx.doi.org/10.1016/j.jfa.2018.03.008.

Chowdhury, S. and Needham, T. Generalized Spectral Clustering via Gromov-Wasserstein Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 712–720. PMLR, 2021.

Chung, F. Laplacians and the Cheeger Inequality for Directed Graphs. *Annals of Combinatorics*, 9: 1–19, 2005.

Cohen, J. E. and Rothblum, U. G. Nonnegative Ranks, Decompositions, and Factorizations of Nonnegative Matrices. *Linear Algebra and its Applications*, 190:149–168, 1993.

Courty, N., Flamary, R., and Tuia, D. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pp. 274–289. Springer, 2014.

Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Advances in neural information processing systems*, 26, 2013a.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013b. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Dang, C. D. and Lan, G. Stochastic Block Mirror Descent Methods for Nonsmooth and Stochastic Optimization. *SIAM J. Optim.*, 25(2):856–881, January 2015.

Dong, S., Pan, Z., Fu, Y., Xu, D., Shi, K., Yang, Q., Shi, Y., and Zhuo, C. Partial Unbalanced Feature Transport for Cross-Modality Cardiac Image Segmentation. *IEEE Transactions on Medical Imaging*, 2023.

Dykstra, R. L. An Algorithm for Restricted Least Squares Regression. *Journal of the American Statistical Association*, 78(384):837–842, December 1983. ISSN 1537-274X. doi: 10.1080/01621459.1983.10477029. URL http://dx.doi.org/10.1080/01621459.1983.10477029.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. Statistical Optimal Transport via Factored Couplings. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2454–2465. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/forrow19a.html.

Frieze, A., Kannan, R., and Vempala, S. Fast Monte-Carlo Algorithms for Finding Low-rank Approximations. *J. ACM*, 51(6):1025–1041, nov 2004. ISSN 0004-5411. doi: 10.1145/1039488.1039494. URL https://doi.org/10.1145/1039488.1039494.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a Wasserstein Loss. *Advances in neural information processing systems*, 28, 2015.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. A mathematical perspective on Transformers. *arXiv preprint arXiv:2312.10794*, 2023.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1–2):267–305, December 2014. ISSN 1436-4646. doi: 10.1007/s10107-014-0846-1. URL http://dx.doi.org/10.1007/s10107-014-0846-1.

Halmos, P., Liu, X., Gold, J., Chen, F., Ding, L., and Raphael, B. J. DeST-OT: Alignment of Spatiotemporal Transcriptomics Data. In *International Conference on Research in Computational Molecular Biology*, pp. 434–437. Springer, 2024.

Indyk, P., Vakilian, A., Wagner, T., and Woodruff, D. P. Sample-optimal low-rank approximation of distance matrices. In Beygelzimer, A. and Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1723–1751. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/indyk19a.html.

Kantorovich, L. On the Translocation of Masses: Doklady akademii nauk ussr. 1942.

Klein, D., Palla, G., Lange, M., Klein, M., Piran, Z., Gander, M., Meng-Papaxanthos, L., Sterr, M., Bastidas-Ponce, A., Tarquis-Medina, M., et al. Mapping cells through time and space with moscot. *bioRxiv*, pp. 2023–05, 2023.

Lin, C.-H., Azabou, M., and Dyer, E. L. Making transport more robust and interpretable by moving data through a small number of anchor points. *Proceedings of machine learning research*, 139: 6631, 2021.

Liu, W., Zhang, C., Zheng, N., and Qian, H. Approximating optimal transport via low-rank and sparse factorization. *CoRR*, abs/2111.06546, 2021. URL https://arxiv.org/abs/2111.06546.

Liu, X., Zeira, R., and Raphael, B. J. Partial alignment of multislice spatially resolved transcriptomics data. *Genome Research*, 33(7):1124–1132, 2023.

Mémoli, F. On the use of Gromov-Hausdorff Distances for Shape Comparison. 2007.

Mémoli, F. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of computational mathematics*, 11:417–487, 2011.

Nesterov, Y. Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Orlin, J. B. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, Aug 1997. ISSN 1436-4646. doi: 10.1007/BF02614365. URL https://link.springer.com/content/pdf/10.1007/BF02614365.pdf.

Pham, K., Le, K., Ho, N., Pham, T., and Bui, H. H. On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:211068892.

Sander, M. E., Ablin, P., Blondel, M., and Peyré, G. Sinkformers: Transformers with Doubly Stochastic Attention. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/sander22a.html.

Scetbon, M. and Cuturi, M. Low-rank Optimal Transport: Approximation, Statistics and Debiasing. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4btNeXKFAQ.

Scetbon, M., Cuturi, M., and Peyré, G. Low-Rank Sinkhorn Factorization. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:232147563.

Scetbon, M., Peyré, G., and Cuturi, M. Linear-time Gromov Wasserstein Distances using Low Rank Couplings and Costs. In *International Conference on Machine Learning*, pp. 19347–19365. PMLR, 2022.

Scetbon, M., Klein, M., Palla, G., and Cuturi, M. Unbalanced Low-rank Optimal Transport Solvers, 2023.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943, 2019.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

Tarjan, R. E. Dynamic trees as search trees via Euler tours, applied to the network simplex algorithm. *Mathematical Programming*, 78(2):169–177, Aug 1997. ISSN 1436-4646. doi: 10.1007/BF02614369. URL https://link.springer.com/content/pdf/10.1007/BF02614369.pdf.

Tay, Y., Bahri, D., Yang, L., Metzler, D., and Juan, D.-C. Sparse Sinkhorn Attention. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9438–9447. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/tay20a.html.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Vayer, T., Chapel, L., Flamary, R., Tavenard, R., and Courty, N. Fused Gromov-Wasserstein distance for structured objects. *Algorithms*, 13(9):212, August 2020. ISSN 1999-4893. doi: 10.3390/a13090212. URL http://dx.doi.org/10.3390/a13090212.

Vincent-Cuaz, C., Flamary, R., Corneli, M., Vayer, T., and Courty, N. Semi-relaxed Gromov-Wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RShaMexjc-x.

Wolf, F. A., Angerer, P., and Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Xu, H., Luo, D., Zha, H., and Duke, L. C. Gromov-Wasserstein Learning for Graph Matching and Node Embedding. In *International conference on machine learning*, pp. 6932–6941. PMLR, 2019.

Yang, J. and Leskovec, J. Defining and Evaluating Network Communities based on Ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, pp. 1–8, 2012.

Yang, K. D., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G., and Uhler, C. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

Yin, H., Benson, A. R., Leskovec, J., and Gleich, D. F. Local Higher-Order Graph Clustering. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 555–564, 2017.

Zeira, R., Land, M., Strzalkowski, A., and Raphael, B. J. Alignment and integration of spatial transcriptomics data. *Nature Methods*, 19(5):567–575, 2022.

A Low-rank optimal transport

A.1 Low-rank factorizations

The set of low-rank couplings. Given $\boldsymbol{M} \in \mathbb{R}_+^{n\times m}$, the *nonnegative rank* of $\boldsymbol{M}$ is the least number of nonnegative, rank-1 matrices that sum to $\boldsymbol{M}$:

$$\mathrm{rk}_+(\boldsymbol{M}) = \min_{r\geq 1}\left\{\boldsymbol{M} = \sum_{i=1}^{r}\boldsymbol{M}_i, \text{ such that } \mathrm{rk}(\boldsymbol{M}_i) = 1 \text{ and } \boldsymbol{M}_i \geq 0 \text{ for all } i\right\}.$$

Let $\boldsymbol{a} \in \Delta_n, \boldsymbol{b} \in \Delta_m$ be probability vectors, and let $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ denote the set of rank-r coupling matrices with marginals $\boldsymbol{a}$ and $\boldsymbol{b}$:

$$\Pi_{\boldsymbol{a},\boldsymbol{b}}(r) = \{\boldsymbol{P} \in \mathbb{R}_+^{n\times m} : \boldsymbol{P}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{a},\ \boldsymbol{P}\mathbf{1}_n = \boldsymbol{b},\ \mathrm{rk}_+(\boldsymbol{P}) \leq r\}.$$

To optimize any cost over $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$, one requires a parameterization of this set.

Factored couplings. The *factored coupling* parameterization of $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ introduced in Forrow et al. (2019), and used by Scetbon et al. (2021); Scetbon & Cuturi (2022); Scetbon et al. (2022, 2023) is

$$\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r) := \{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times (\mathbb{R}_+^*)^r : \boldsymbol{Q} \in \Pi_{\boldsymbol{a},\boldsymbol{g}}, \boldsymbol{R} \in \Pi_{\boldsymbol{b},\boldsymbol{g}}\}.$$

Cohen & Rothblum (1993) show that any $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ may be decomposed as $\boldsymbol{P} = \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}$ for some triple $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g}) \in \mathsf{FC}$. Thus, for cost matrix $\boldsymbol{C} \in \mathbb{R}^{n\times m}$, the general low-rank optimal transport problem is equivalent to an optimization over factored couplings:

$$\min_{\boldsymbol{P}\in\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)} \langle \boldsymbol{C}, \boldsymbol{P}\rangle_F = \min_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g})\in\mathsf{FC}_{\boldsymbol{a},\boldsymbol{b}}(r)} \langle \boldsymbol{C}, \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}\rangle_F. \tag{11}$$

Latent coupling factorization. The *latent coupling* parameterization of $\Pi_{\boldsymbol{a},\boldsymbol{b}}(r)$ introduced in Lin et al. (2021), and used in the present work is

$$\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r) := \{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{r\times r} : \boldsymbol{Q} \in \Pi_{\boldsymbol{a},\cdot}, \boldsymbol{R} \in \Pi_{\boldsymbol{b},\cdot}, \boldsymbol{T} \in \Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}\},$$

where $\boldsymbol{g}_Q, \boldsymbol{g}_R$ are the inner marginals of $\boldsymbol{Q}$ and $\boldsymbol{R}$.

Latent coupling diagonalization. The LC-factorization recovers the factorization of Forrow et al. (2019) as a sub-case. While the diagonal factorization of previous works cannot be directly converted to the LC-factorization, the LC-factorization can easily recover the diagonal factorization. In particular, taking $\boldsymbol{Q}' \leftarrow \boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}$ one can refactor

$$\boldsymbol{P}_r = \boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{Q}'\,\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}$$

or alternatively taking $\boldsymbol{R}' = \boldsymbol{R}\,\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{T}^{\mathrm{T}}$ may refactor as

$$\boldsymbol{P}_r = \boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}} = \boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{g}_Q)(\boldsymbol{R}')^{\mathrm{T}}$$

So that instead of returning $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})$ one may alternatively return $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) \to (\boldsymbol{Q}',\boldsymbol{R},\mathrm{diag}(\boldsymbol{g}_R))$ or $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) \to (\boldsymbol{Q},\boldsymbol{R}',\mathrm{diag}(\boldsymbol{g}_Q))$ to recover the Forrow et al. (2019) factorization. An example of this diagonal-conversion is offered in Figure 12.

A.2 Balanced low-rank optimal transport

The FRLC optimization problem Our optimization problem is over the variables $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})$ and defined as follows:

$$\min_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})\in\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)} \mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}), \tag{12}$$

where our objective function $\mathcal{L}_{\mathrm{LC}}$ is

$$\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) = \langle \boldsymbol{C}, \boldsymbol{Q}(\mathrm{diag}(1/\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n))\boldsymbol{T}(\mathrm{diag}(1/\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m))\boldsymbol{R}^{\mathrm{T}}\rangle \tag{13}$$

Given $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) \in \mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r)$, sub-couplings $\boldsymbol{Q}$ and $\boldsymbol{R}$ are constrained by:

$$\mathcal{C}_1(\boldsymbol{a}) := \{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{Q}\mathbf{1}_r = \boldsymbol{a}\}, \quad \mathcal{C}_1(\boldsymbol{b}) := \{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}\},$$

while the convex sets constraining the latent coupling matrix $\boldsymbol{T}$ are

$$\mathcal{C}_2(\boldsymbol{g}_Q) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{T}\mathbf{1}_r = \boldsymbol{g}_Q\}, \quad \mathcal{C}_2(\boldsymbol{g}_R) := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathcal{R}_+ : \boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \boldsymbol{g}_R\},$$

where $\boldsymbol{g}_Q = \boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n$ and $\boldsymbol{g}_R = \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m$ as per Definition 3.1. Under these definitions, $\mathsf{LC}_{\boldsymbol{a},\boldsymbol{b}}(r) = \mathcal{C}_1 \cap \mathcal{C}_2$, where

$$\mathcal{C}_1 = \mathcal{C}_1(\boldsymbol{a}) \cap \mathcal{C}_1(\boldsymbol{b}) \quad \text{and} \quad \mathcal{C}_2 = \mathcal{C}_2(\boldsymbol{g}_Q) \cap \mathcal{C}_2(\boldsymbol{g}_R) \tag{14}$$

To solve (12), we separate the variables into two "blocks" of variables, $(\boldsymbol{Q}, \boldsymbol{R})$ and $\boldsymbol{T}$, and perform two block updates per iteration, as follows. Let $(\gamma_k)_{k\geq 0}$ be a positive sequence of stepsizes. Suppose we have $(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k) \in \mathcal{C}$. We update the first variable block $(\boldsymbol{Q}, \boldsymbol{R})$ by taking a locally optimal (mirror descent) update step, while the second variable block $\boldsymbol{T}$ is held fixed:

$$(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}) \leftarrow \operatorname*{arg\,min}_{(\boldsymbol{Q},\boldsymbol{R}):(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}_k)\in\mathcal{C}_1} \langle (\boldsymbol{Q}, \boldsymbol{R}), \nabla_{(\boldsymbol{Q},\boldsymbol{R})}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}((\boldsymbol{Q}, \boldsymbol{R})\|(\boldsymbol{Q}_k, \boldsymbol{R}_k)) \tag{15}$$

Here, we slightly abused the notation by putting $(\boldsymbol{Q}, \boldsymbol{R})$ inside an inner product. The triple $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}_k)$ produced by this update lies in $\mathcal{C}_1$. Next, we update $\boldsymbol{T}$, the second variable block, by taking another locally optimal (mirror descent) step, while the first variable block is held fixed.

$$\boldsymbol{T}_{k+1} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{T}:(\boldsymbol{Q}_{k+1},\boldsymbol{R}_{k+1},\boldsymbol{T})\in\mathcal{C}_2} \langle \boldsymbol{T}, \nabla_{\boldsymbol{T}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{T}\|\boldsymbol{T}_k). \tag{16}$$

By construction, the triple $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}_{k+1}) \in \mathcal{C}_2$. However, because the set $\mathcal{C}_1$ only constrains the first variable block $\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}$, we have that $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}_{k+1}) \in \mathcal{C}_2$, and hence $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}_{k+1}) \in \mathcal{C}$. Thus, each iteration produces a feasible triple $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}_{k+1})$ through a pair of locally optimal block updates.

The LOT optimization problem Scetbon et al. (2021) For comparison, recall the optimization problem that is solved in the LOT framework:

$$\min_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g})\in\widetilde{\mathcal{C}}} \mathcal{L}_{\mathrm{LOT}}, \tag{17}$$

where the objective function $\mathcal{L}_{\mathrm{LOT}}$ is

$$\mathcal{L}_{\mathrm{LOT}} := \langle \boldsymbol{C}, \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}\rangle \tag{18}$$

and where $\widetilde{\mathcal{C}} = \widetilde{\mathcal{C}}_1 \cap \widetilde{\mathcal{C}}_2$ with:

$$\widetilde{\mathcal{C}}_1 := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times (\mathbb{R}_+^*)^r : \boldsymbol{Q}\mathbf{1}_r = \boldsymbol{a}, \boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}\}$$
$$\widetilde{\mathcal{C}}_2 := \{(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{r} : \boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{g} = \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m\}.$$

Here, there are also three optimization variables $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g})$, but they are updated *together* in each iteration of LOT. That is, given feasible $(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{g}_k) \in \mathcal{C}$, an iteration of LOT updates this triple via

$$(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{g}_{k+1}) := \operatorname*{arg\,min}_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g})\in\widetilde{\mathcal{C}}} \langle (\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g}), \nabla_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g})}\mathcal{L}_{\mathrm{LOT}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}((\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{g})\|(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{g}_k)),$$

where $(\gamma_k)_{k\geq 0}$ is again a positive sequence of stepsizes. Scetbon et al. (2021) then compute the unconstrained argmin across all variables to yield a set of unconstrained kernels $(\boldsymbol{K}_Q, \boldsymbol{K}_R, \boldsymbol{k}_g)$, using Dykstra to jointly project the unconstrained update onto the intersection $\widetilde{\mathcal{C}}$ of the constraint sets.

OT subroutine in FRLC To see why we do not need Dykstra in the FRLC scheme, observe that the update (15) of variables $(\boldsymbol{Q}, \boldsymbol{R})$ can be equivalently expressed as

$$(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}) \leftarrow \operatorname*{arg\,min}_{\boldsymbol{Q},\boldsymbol{R}\,:\,(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}_k)\in\mathcal{C}_1} \langle (\boldsymbol{Q}, \boldsymbol{R}), \nabla_{(\boldsymbol{Q},\boldsymbol{R})}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}((\boldsymbol{Q}, \boldsymbol{R})\|(\boldsymbol{Q}_k, \boldsymbol{R}_k)),$$

Thus, even though the pair $(\boldsymbol{Q}, \boldsymbol{R})$ is being updated at this step, solving for $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1})$ above is equivalent to updating each individually because $\boldsymbol{Q}$ and $\boldsymbol{R}$ do not share an inner marginal:

$$\boldsymbol{Q}_{k+1} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{Q}\,:\,\boldsymbol{Q}\mathbf{1}_r=\boldsymbol{a}} \langle \boldsymbol{Q}, \nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{Q}\|\boldsymbol{Q}_k) \tag{19}$$

$$\boldsymbol{R}_{k+1} \leftarrow \underset{\boldsymbol{R}\,:\,\boldsymbol{R}\mathbf{1}_r=\boldsymbol{b}}{\arg\min} \langle \boldsymbol{R}, \nabla_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{R}\|\boldsymbol{R}_k) \tag{20}$$

We choose to add the regularization $\tau\mathrm{KL}((\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m)\|(\boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m))$ to turn each update here into an entropy-regularized semi-relaxed optimal transport problem, and to ensure β-smoothness:

$$\boldsymbol{Q}_{k+1} \leftarrow \underset{\boldsymbol{Q}\,:\,\boldsymbol{Q}\mathbf{1}_r=\boldsymbol{a}}{\arg\min} \langle \boldsymbol{Q}, \nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{Q}\|\boldsymbol{Q}_k) + \tau\mathrm{KL}(\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n\|\boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n) \tag{21}$$

$$\boldsymbol{R}_{k+1} \leftarrow \underset{\boldsymbol{R}\,:\,\boldsymbol{R}\mathbf{1}_r=\boldsymbol{b}}{\arg\min} \langle \boldsymbol{R}, \nabla_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{R}\|\boldsymbol{R}_k) + \tau\mathrm{KL}(\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m\|\boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m) \tag{22}$$

After updating $\boldsymbol{Q}$ and $\boldsymbol{R}$, the update on $\boldsymbol{T}$ then follows a similar form:

$$\boldsymbol{T}_{k+1} \leftarrow \underset{\boldsymbol{T}\,:\,(\boldsymbol{Q}_{k+1},\boldsymbol{R}_{k+1},\boldsymbol{T})\in\mathcal{C}_2}{\arg\min} \langle \boldsymbol{T}, \nabla_{\boldsymbol{T}}\mathcal{L}_{\mathrm{LC}}\rangle + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{T}\|\boldsymbol{T}_k), \tag{23}$$

leading to balanced constraints on $\boldsymbol{T}$ of the form $\boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \boldsymbol{Q}_{k+1}^{\mathrm{T}}\mathbf{1}_n$ and $\boldsymbol{T}\mathbf{1}_r = \boldsymbol{R}_{k+1}^{\mathrm{T}}\mathbf{1}_m$, allowing the problem to be solved by Sinkhorn.

Importantly, there are no constraints in $\mathcal{C}_1$ involving both $\boldsymbol{Q}$ and $\boldsymbol{R}$, which is what allows the optimization to split in this way. If there were such constraints, we would have needed to use Dykstra to update the pair $(\boldsymbol{Q}, \boldsymbol{R})$. Because our update scheme is equivalent to the three updates of individual variables given in (21), (22), (23), we can solve for each update using optimal transport. As $\boldsymbol{Q}$ and $\boldsymbol{R}$ are not required to match the inner marginals exactly, the OT problems associated to $\boldsymbol{Q}$ and $\boldsymbol{R}$ are semi-relaxed by construction.

The separation of our block updates into a step where $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}_k) \in \mathcal{C}_1$ and $(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_{k+1}, \boldsymbol{T}) \in \mathcal{C}_2$ allows us to entirely remove the optimization over inner marginals $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$, as done in all previous works on low-rank optimal transport which optimize $\boldsymbol{g}$ explicitly as both a variable and a constraint of the optimization Scetbon et al. (2021, 2023); Scetbon & Cuturi (2022). If one were to introduce an extended loss in the style of previous works which adds $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ as variables in the form $\mathcal{H}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}, \boldsymbol{g}_Q, \boldsymbol{g}_R) = \langle \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C}\rangle_F$, one observes an equivalence to simply taking a semi-relaxed projection.

Lemma A.1. *Define the function* $\mathcal{H}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}, \boldsymbol{g}_Q, \boldsymbol{g}_R) = \langle \boldsymbol{C}, \boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\rangle_F$ *and let* $\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ *be as in* (13)*. One has the following equivalence:*

$$\min_{\boldsymbol{g}_R\in\Delta_r, \boldsymbol{g}_Q\in\Delta_r, \boldsymbol{Q}\in\Pi_{\boldsymbol{a},\boldsymbol{g}_Q}, \boldsymbol{R}\in\Pi_{\boldsymbol{b},\boldsymbol{g}_R}} \mathcal{H}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}_k, \boldsymbol{g}_Q, \boldsymbol{g}_R) = \min_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}_k)\in\mathcal{C}_1} \mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}_k) \tag{24}$$

Thus the semi-relaxed projections yield locally optimal inner marginals.

Proof. To see why this is true, notice that, so long as the outer marginals are tightly satisfied $\boldsymbol{Q}\mathbf{1}_r = \boldsymbol{a}$ and $\boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}$ for $\boldsymbol{Q} \geq \mathbf{0}_{n\times r}, \boldsymbol{R} \geq \mathbf{0}_{m\times r}$, one has

$$\sum_i \boldsymbol{g}_{R,i} = \sum_i \langle \boldsymbol{R}_{i,\cdot}^{\mathrm{T}}, \mathbf{1}_m\rangle = \sum_i \langle \boldsymbol{R}_{\cdot,i}, \mathbf{1}_m\rangle = \sum_\ell \sum_i \boldsymbol{R}_{\ell,i} = \sum_\ell \boldsymbol{b}_\ell = 1$$

and

$$\sum_i \boldsymbol{g}_{Q,i} = \sum_i \langle \boldsymbol{Q}_{i,\cdot}^{\mathrm{T}}, \mathbf{1}_n\rangle = \sum_i \langle \boldsymbol{Q}_{\cdot,i}, \mathbf{1}_m\rangle = \sum_\ell \sum_i \boldsymbol{Q}_{\ell,i} = \sum_\ell \boldsymbol{a}_\ell = 1$$

Therefore, the inner marginals $\boldsymbol{g}_Q^* = (\boldsymbol{Q}^*)^{\mathrm{T}}\mathbf{1}_n$ and $\boldsymbol{g}_R^* = (\boldsymbol{R}^*)^{\mathrm{T}}\mathbf{1}_m$ induced by the optimal $\boldsymbol{Q}^*, \boldsymbol{R}^*$ of the optimization problem on the right hand side satisfy the constraints $\boldsymbol{g}_R^* \in \Delta_r, \boldsymbol{g}_Q^* \in \Delta_r$ on the left hand side, so the two minimums coincide. □

This implies that an extra optimization for $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ is unnecessary in a coordinate update which alternates $(\boldsymbol{Q}, \boldsymbol{R})$ and $\boldsymbol{T}$. If we did not perform a block-update, in the form of standard MD in the objective described above, we would encounter some difficulty. In particular $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ would optimized with the constraint $\boldsymbol{g}_Q \in \Delta_r$ and $\boldsymbol{g}_R \in \Delta_r$, and would concurrently constrain all of the other variables as $\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{g}_Q$, $\boldsymbol{T}\mathbf{1}_r = \boldsymbol{g}_Q$, and $\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{g}_R$, $\boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \boldsymbol{g}_R$.

B Block-Coordinate steps for the OT sub-problems

We use a latent non-diagonal coupling instead of an inner diagonal coupling $\mathrm{diag}(\boldsymbol{g})$ of the form of Forrow et al. (2019). This allows us to loosen the constraint that the inner marginals have to be joined by a common coupling $\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{g}$. The fundamental advantage of this choice is that we can decouple the convex-optimization problem for $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ entirely. One can simply solve for the optimal $\boldsymbol{Q}$ and $\boldsymbol{R}$ *independently*, yield the associated inner marginals for each $\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{g}_{\boldsymbol{Q}}$ and $\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{g}_{\boldsymbol{R}}$, and then find the optimal $\boldsymbol{T}$ which links the two. This link is provided by the aforementioned form of the problem, where:

$$\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$$

For $\boldsymbol{Q}, \boldsymbol{R}$ in either the appropriate set of couplings or a relaxation thereof (which we will describe shortly). $\boldsymbol{X}$ is related to $\boldsymbol{T}$ by:

$$\boldsymbol{X} = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)$$

And, $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_{\boldsymbol{Q}}, \boldsymbol{g}_{\boldsymbol{R}}}$ consistently for all cases. As the semi-relaxed case is intermediate between fully-relaxed and balanced, it has ideas which generalize to both directly. As such, we use it as the leading example again. As in Scetbon et al. (2021), we take proximal-steps of the form:

$$\min_{\boldsymbol{\zeta}} \langle \nabla \mathcal{L}(\boldsymbol{\zeta})\ |_{\boldsymbol{\zeta}_k}, \boldsymbol{\zeta} \rangle_F + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{\zeta} \| \boldsymbol{K}^{(k)})$$

Where these steps are now in block-wise fashion on $(\boldsymbol{Q}, \boldsymbol{R})$ and $\boldsymbol{T}$, rather than joint. One may identify for each sub-factor in $(\boldsymbol{Q}, \boldsymbol{R})$ and $\boldsymbol{T}$ a linearized gradient as before, which yields a set of objectives which each solve an independent optimal-transport for the sub-factors. In particular, we have that:

$$\begin{aligned}
\langle \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C} \rangle_F &= \mathrm{Tr}\left[\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\right] = \langle \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Q} \rangle_F \\
\langle \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C} \rangle_F &= \mathrm{Tr}\left[\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\right] = \langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}, \boldsymbol{R} \rangle_F \\
\langle \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C} \rangle_F &= \mathrm{Tr}\left[\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}\right] = \langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}, \boldsymbol{X} \rangle_F
\end{aligned}$$

A linearization in the left-slot of the inner product as $\langle \boldsymbol{Q}, \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}(\boldsymbol{Q}_k)^{\mathrm{T}} \rangle := \langle \boldsymbol{Q}, \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}} \rangle$ or $\langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}(\boldsymbol{R}_k), \boldsymbol{R} \rangle_F := \langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}, \boldsymbol{R} \rangle_F$ is common practice for quadratic problems. In this case, the directional derivative of $\boldsymbol{Q}$ and $\boldsymbol{R}$ in the matrix-direction $\boldsymbol{V}$ are respectively:

$$\begin{aligned}
D\langle \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Q} \rangle_F \circ (\boldsymbol{V}) = \langle \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{V} \rangle_F \implies \nabla_{\boldsymbol{Q}}\mathcal{L} = \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}} \\
D\langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}, \boldsymbol{R} \rangle_F \circ (\boldsymbol{V}) = \langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}, \boldsymbol{V} \rangle_F \implies \nabla_{\boldsymbol{R}}\mathcal{L} = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}
\end{aligned}$$

Without this linearization assumption on $\boldsymbol{X}$, the full gradient may be evaluated as well. In particular, we note that for $\mathrm{diag}^{-1}(\cdot)$ the matrix-to-vector extraction of the diagonal, the directional derivative on $\boldsymbol{Q}$ is:

$$\begin{aligned}
D\langle \boldsymbol{C}, \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}} \rangle_F \circ \boldsymbol{V} &= \langle \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{V} \rangle_F + \langle \boldsymbol{C}, \boldsymbol{Q}D\boldsymbol{X}^{\mathrm{T}} \circ (\boldsymbol{V})\boldsymbol{R}^{\mathrm{T}} \rangle_F \\
&= \langle \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}} - \mathbf{1}_n \mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}}, \boldsymbol{V} \rangle_F
\end{aligned}$$

Thus, without the linearization assumption one may use product rule on $\boldsymbol{X}$ as an implicit function of $\boldsymbol{Q}$ (resp. $\boldsymbol{R}$) to take the total derivative:

$$\nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{FRLC}} = \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}} - \mathbf{1}_n \mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}}$$

Likewise for $\boldsymbol{R}$,

$$\begin{aligned}
D\langle \boldsymbol{C}, \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}} \rangle_F \circ \boldsymbol{V} &= \langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}, \boldsymbol{V} \rangle_F + \langle \boldsymbol{C}, \boldsymbol{Q}D\boldsymbol{X} \circ (\boldsymbol{V})\boldsymbol{R}^{\mathrm{T}} \rangle_F \\
&= \langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X} - \mathbf{1}_m \mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X})^{\mathrm{T}}, \boldsymbol{V} \rangle_F,
\end{aligned}$$

and so

$$\nabla_{\boldsymbol{R}}\mathcal{L}_{\mathrm{FRLC}} = \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X} - \mathbf{1}_m \mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X})^{\mathrm{T}}.$$

Both the W and GW problem have these rank-one perturbations of the gradient from the derivative with respect to $\boldsymbol{X}$.

Lastly, owing to the block-coordinate updates which fix $(\boldsymbol{Q}, \boldsymbol{R})$ preceding the update on $\boldsymbol{T}$, the derivative on $\boldsymbol{T}$ follows directly by chain rule on $\boldsymbol{X}$:

$$D\langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}, \boldsymbol{X}\rangle_F \circ (\boldsymbol{V}) = \langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}, \boldsymbol{V}\rangle_F \implies \nabla_{\boldsymbol{X}}\mathcal{L} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}$$
$$\nabla_{\boldsymbol{T}}\mathcal{L} = \mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}})\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}\,\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}})$$

Let (γ_k) be a sequence of step sizes and consider the first-order conditions required for the proximal step as before. From these we have the updated proximal-step updates:

$$\begin{aligned}
\boldsymbol{K}_{\boldsymbol{Q}}^{(k)} &\leftarrow \min_{\boldsymbol{Q}} \langle \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}, \boldsymbol{Q}\rangle_F + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{Q}_k \| \boldsymbol{K}_{\boldsymbol{Q}}^{(k)}) \\
\boldsymbol{K}_{\boldsymbol{R}}^{(k)} &\leftarrow \min_{\boldsymbol{R}} \langle \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X}, \boldsymbol{R}\rangle_F + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{R}_k \| \boldsymbol{K}_{\boldsymbol{R}}^{(k)}) \\
\boldsymbol{K}_{\boldsymbol{T}}^{(k)} &\leftarrow \min_{\boldsymbol{T}} \langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}, \boldsymbol{X}\rangle_F + \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{T}_k \| \boldsymbol{K}_{\boldsymbol{T}}^{(k)}),
\end{aligned}$$

with kernels $\boldsymbol{K}_{\boldsymbol{\zeta}_j}^{(k)}$, for $j = 1, 2, 3$ given by

$$\begin{aligned}
\boldsymbol{K}_{\boldsymbol{Q}}^{(k)} &:= \boldsymbol{Q}_k \odot \exp(\, -\gamma_k \boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}}\,) \\
\boldsymbol{K}_{\boldsymbol{R}}^{(k)} &:= \boldsymbol{R}_k \odot \exp(\, -\gamma_k \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k\,) \\
\boldsymbol{K}_{\boldsymbol{T}}^{(k)} &:= \mathbf{T}_k \odot \exp(\, -\gamma_k \mathrm{diag}(\boldsymbol{g}_Q^{-1})\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}\mathrm{diag}(\boldsymbol{g}_R^{-1})\,)
\end{aligned}$$

Or, dropping the linearization assumption on $(\boldsymbol{Q}, \boldsymbol{R})$ the updates are: $\boldsymbol{K}_{\boldsymbol{\zeta}_j}^{(k)}$, for $j = 1, 2$ given by

$$\begin{aligned}
\boldsymbol{K}_{\boldsymbol{Q}}^{(k)} &:= \boldsymbol{Q}_k \odot \exp(\, -\gamma_k(\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}} - \mathbf{1}_n \mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{Q_k}))^{\mathrm{T}})\,) \\
\boldsymbol{K}_{\boldsymbol{R}}^{(k)} &:= \boldsymbol{R}_k \odot \exp(\, -\gamma_k(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k - \mathbf{1}_m \mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k)^{\mathrm{T}})\,)
\end{aligned}$$

The first projection is onto the set that satisfies the marginal constraint $\boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}$. In particular, following the discussion above, one has the coordinate-MD step:

$$\begin{aligned}
\min_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})} \quad & \frac{1}{\gamma_k}\mathrm{KL}\left((\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \,\|\, (\boldsymbol{K}_{\boldsymbol{Q}}, \boldsymbol{K}_{\boldsymbol{R}}, \boldsymbol{K}_{\boldsymbol{T}})\right) + \tau\mathrm{KL}(\boldsymbol{Q}\mathbf{1}_r \| \boldsymbol{a}) \\
s.t. \quad & \boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}
\end{aligned} \tag{25}$$

As before, there is no difference from the previous case, where one takes the unconstrained projection $\boldsymbol{R} = \mathrm{diag}(\boldsymbol{b}/\boldsymbol{K}_{\boldsymbol{R}}\mathbf{1}_r)\boldsymbol{K}_{\boldsymbol{R}}$. To generalize this, we also consider adding a soft-constraint on the inner marginal of $\boldsymbol{R}$ to be near that of the previous iteration. In particular, we consider the problem:

$$\begin{aligned}
\min_{(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})} \quad & \frac{1}{\gamma_k}\mathrm{KL}((\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \| (\boldsymbol{K}_{\boldsymbol{Q}}, \boldsymbol{K}_{\boldsymbol{R}}, \boldsymbol{K}_{\boldsymbol{T}})) \\
& \qquad + \tau\mathrm{KL}(\boldsymbol{Q}\mathbf{1}_r \| \boldsymbol{a}) + \tau\mathrm{KL}(\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m \| \boldsymbol{g}_R^{(k-1)} \equiv \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m) \\
s.t. \quad & \boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}
\end{aligned} \tag{26}$$

Which yields the relaxed solution of $\boldsymbol{R} = \mathrm{SR}^{\mathrm{R}}$-*projection*$(\boldsymbol{K}_{\boldsymbol{R}}, \gamma_k, \tau, \boldsymbol{b}, \boldsymbol{g}_R^{(k-1)})$ which generalizes the original projection and is equivalent to it for $\tau = 0$. This regularization is essential, as it ensures β-smoothness of the objective. For $\boldsymbol{Q}$, as the constraint on $\boldsymbol{g}$ is fully relaxed, the Lagrange multiplier $\boldsymbol{\lambda}_1 = \mathbf{0}$ entirely, such that the problem 40 now becomes fully-unconstrained:

$$\inf_{\boldsymbol{Q}} \left(\frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{Q} \| \boldsymbol{K}_{\boldsymbol{Q}}) + \tau\mathrm{KL}(\boldsymbol{Q}\mathbf{1}_r \| \boldsymbol{a}) \right) \tag{27}$$

To generalize this solution, we again consider adding a soft-regularization on the inner marginal of $\boldsymbol{Q}$, where we consider the alternate problem:

$$\inf_{\boldsymbol{Q}} \left(\frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{Q} \| \boldsymbol{K}_{\boldsymbol{Q}}) + \tau\mathrm{KL}(\boldsymbol{Q}\mathbf{1}_r \| \boldsymbol{a}) + \tau\mathrm{KL}(\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n \| \boldsymbol{g}_{\boldsymbol{Q}}^{(k-1)} \equiv \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n) \right) \tag{28}$$

Which trivially recovers the original for $\tau = 0$. This form has a solution given by an unbalanced optimal transport with kernel $\boldsymbol{K}_{\boldsymbol{Q}}$. We see these two, convex problems in sequence give *independent* optimal solutions for $\boldsymbol{Q}$ and $\boldsymbol{R}$ as the two matrices are not required to share an inner marginal. The last step is to link them via $\boldsymbol{T}$, corresponding to the projection of $(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T})$ onto the set of valid rank-r couplings $\min_{(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}) \in C} \mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}) := \min_{\boldsymbol{T} \in \Pi(\boldsymbol{g}_Q, \boldsymbol{g}_R)} \mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T})$. We verify in C that if $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_{\boldsymbol{Q}} = \boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n, \boldsymbol{g}_{\boldsymbol{R}} = \boldsymbol{R}^{\mathrm{T}} \mathbf{1}_m}$ then $\boldsymbol{P} \in \Pi_{\boldsymbol{a}, \boldsymbol{b}}$. As such, one does not require any projection onto the intersection of convex sets, as done in Scetbon et al. (2021, 2023) via the Dykstra projection algorithm Dykstra (1983). Alternating a coordinate-MD step in $(\boldsymbol{Q}, \boldsymbol{R})$ and a step on $\boldsymbol{T}$, one not only minimizes the objective in an alternating fashion but remains in the feasible set without the need for projection algorithms beyond Sinkhorn. As such, the final linking step is done via:

$$\begin{aligned} \min_{\boldsymbol{T}} \quad & \frac{1}{\gamma_k} \mathrm{KL}(\boldsymbol{T} \| \boldsymbol{K}_{\boldsymbol{T}}) \\ s.t. \quad & \boldsymbol{T} \mathbf{1}_r = \boldsymbol{g}_{\boldsymbol{Q}}, \boldsymbol{T}^{\mathrm{T}} \mathbf{1}_r = \boldsymbol{g}_{\boldsymbol{R}} \end{aligned} \tag{29}$$

This formulation amounts to solving a *balanced* Sinkhorn problem on $\boldsymbol{T}$ with respect to the proximal step kernel matrix.

Algorithm 2 SR^{R}-projection *(semi-relaxed OT, right marginal relaxed)*

Input $\boldsymbol{K}, \gamma, \tau, \boldsymbol{a}, \boldsymbol{b}, \delta$
$\boldsymbol{u} \leftarrow \mathbf{1}_n$
$\boldsymbol{v} \leftarrow \mathbf{1}_r$
repeat
  $\tilde{\boldsymbol{u}} \leftarrow \boldsymbol{u}$
  $\tilde{\boldsymbol{v}} \leftarrow \boldsymbol{v}$
  $\boldsymbol{u} \leftarrow (\boldsymbol{a} / \boldsymbol{K}\boldsymbol{v})$
  $\boldsymbol{v} \leftarrow \left(\boldsymbol{b} / \boldsymbol{K}^{\mathrm{T}} \boldsymbol{u}\right)^{\tau / (\tau + \gamma^{-1})}$
until $\gamma^{-1} \max\{\|\log \tilde{\boldsymbol{u}} / \boldsymbol{u}\|_\infty, \|\log \tilde{\boldsymbol{v}} / \boldsymbol{v}\|_\infty\} < \delta$
return $\mathrm{diag}(\boldsymbol{u}) \boldsymbol{K} \mathrm{diag}(\boldsymbol{v})$

Algorithm 3 U-projection *(unbalanced OT)*

Input $\boldsymbol{K}, \gamma, \tau, \boldsymbol{a}, \boldsymbol{b}, \delta$
$\boldsymbol{u} \leftarrow \mathbf{1}_n$
$\boldsymbol{v} \leftarrow \mathbf{1}_r$
repeat
  $\tilde{\boldsymbol{u}} \leftarrow \boldsymbol{u}$
  $\tilde{\boldsymbol{v}} \leftarrow \boldsymbol{v}$
  $\boldsymbol{u} \leftarrow (\boldsymbol{a} / \boldsymbol{K}\boldsymbol{v})^{\tau / (\tau + \gamma^{-1})}$
  $\boldsymbol{v} \leftarrow \left(\boldsymbol{b} / \boldsymbol{K}^{\mathrm{T}} \boldsymbol{u}\right)^{\tau / (\tau + \gamma^{-1})}$
until $\gamma^{-1} \max\{\|\log \tilde{\boldsymbol{u}} / \boldsymbol{u}\|_\infty, \|\log \tilde{\boldsymbol{v}} / \boldsymbol{v}\|_\infty\} < \delta$
return $\mathrm{diag}(\boldsymbol{u}) \boldsymbol{K} \mathrm{diag}(\boldsymbol{v})$

C The Latent Coupling Matrix

To solve the balanced form (and generalize the principle to the relaxed problems), we consider an alternative parametrization of the inner matrix. In particular, previous works Scetbon et al. (2021, 2023) consider $\mathrm{diag}(1/\boldsymbol{g})$ to be the inner matrix, with marginals $\boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n = \boldsymbol{R}^{\mathrm{T}} \mathbf{1}_m = \boldsymbol{g}$ to ensure that the outer conditions $\boldsymbol{Q} \mathbf{1}_r = \boldsymbol{a}$ and $\boldsymbol{R} \mathbf{1}_r = \boldsymbol{b}$ hold. We instead relax the constraint that $\boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n$ and $\boldsymbol{R}^{\mathrm{T}} \mathbf{1}_m$ are equal, by allowing $\boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n = \boldsymbol{g}_Q$ and $\boldsymbol{R}^{\mathrm{T}} \mathbf{1}_m = \boldsymbol{g}_R$ to vary arbitrarily, and considering a non-diagonal inner matrix $\boldsymbol{X} \in \mathbb{R}^{r \times r}$ in the place of $\mathrm{diag}(1/\boldsymbol{g})$ where we have the conditions:

$$\boldsymbol{X} \boldsymbol{g}_R = \boldsymbol{X} \boldsymbol{R}^{\mathrm{T}} \mathbf{1}_m = \mathbf{1}_r$$

And

$$\boldsymbol{X}^{\mathrm{T}} \boldsymbol{g}_Q = \boldsymbol{X} \boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n = \mathbf{1}_r$$

Thus, if one considers the coupling formed as $\boldsymbol{P}_r = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$, one maintains that $\boldsymbol{P}_r \in \Pi_{\boldsymbol{a},\boldsymbol{b}}$ from the condition of $\mathcal{C}_1$, defined in the balanced case as in (14). We clearly have that:

$$\boldsymbol{P}_r \mathbf{1}_m = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{Q}\mathbf{1}_r = \boldsymbol{a}$$

And:

$$\boldsymbol{P}_r^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}\mathbf{1}_n = \boldsymbol{R}\mathbf{1}_r = \boldsymbol{b}$$

We first consider two approaches to optimizing for such a $\boldsymbol{X}$, given that it is not a coupling. First, we consider the appropriate proximal step for $\boldsymbol{X}$

$$\min_{\boldsymbol{\zeta}} \langle \nabla \mathcal{L}(\boldsymbol{\zeta}) \mid_{\boldsymbol{\xi}_k}, \boldsymbol{\zeta} \rangle_F + \frac{1}{\gamma_k} \mathrm{KL}(\boldsymbol{\zeta} \| \boldsymbol{K}^{(k)})$$

We first note that the gradient of our loss with respect to $\boldsymbol{X}$ is given as $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}$. If one supposes that $\boldsymbol{X}$ is invertible with $\boldsymbol{X}^{-1} = \boldsymbol{T}$, we have that for any such $\boldsymbol{T}$:

$$\boldsymbol{X}^{-1}\mathbf{1}_r = \boldsymbol{T}\mathbf{1}_r = \boldsymbol{g}_R$$

And

$$\boldsymbol{X}^{-T}\mathbf{1}_r = \boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \boldsymbol{g}_Q$$

This implies that this inverse matrix $\boldsymbol{T}$ is a coupling such that $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_R,\boldsymbol{g}_Q}$, which also suggests one might be able to update it using Sinkhorn. In fact, being a density in $\mathbb{R}_+^{r\times r}$ it represents a transition matrix between the latent r-dimensional variables. In particular, writing the proximal step in full, we have:

$$\min_{\boldsymbol{T}} \langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}, \boldsymbol{T}^{-1} \rangle_F + \frac{1}{\gamma_k} \mathrm{KL}(\boldsymbol{T}_k \| \boldsymbol{K}_{\boldsymbol{T}}^{(k)})$$

Noting the derivative $D(\boldsymbol{X}^{-1}) \circ \boldsymbol{V} = -\boldsymbol{X}^{-1}\boldsymbol{V}\boldsymbol{X}^{-1}$, we have from the first-order condition that:

$$-\boldsymbol{T}^{-T}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}\boldsymbol{T}^{-T} + \frac{1}{\gamma_k} \log \left[\frac{\boldsymbol{T}_k}{\boldsymbol{K}_{\boldsymbol{T}}^{(k)}} \right] = \mathbf{0}$$

This implies the kernel matrix update:

$$\boldsymbol{K}_{\boldsymbol{T}}^{(k)} = \boldsymbol{T}_k \odot \exp\{+\gamma_k \boldsymbol{T}^{-T}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}\boldsymbol{T}^{-T}\}$$

Where one then takes the Sinkhorn projection J onto the set $\Pi_{\boldsymbol{g}_R,\boldsymbol{g}_Q}$ as $\mathcal{P}_{\Pi_{\boldsymbol{g}_R,\boldsymbol{g}_Q}}(\boldsymbol{K}_{\boldsymbol{T}}^{(k)})$ using the Sinkhorn algorithm Cuturi (2013b). However, a more stable and inversion-free update exists which ensures $\boldsymbol{X}$ remains positive by a diagonal re-scaling in the form introduced by Lin et al. (2021). In particular, if one takes $\boldsymbol{X} = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)$, then

$$\boldsymbol{X}\boldsymbol{g}_R = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{g}_R = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\mathbf{1}_r = \mathbf{1}_r$$

and likewise

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{g}_Q = \mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{T}^{\mathrm{T}}\,\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{g}_Q = \mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \mathbf{1}_r$$

so that $\boldsymbol{X}$ necessarily satisfies $\boldsymbol{X}\boldsymbol{g}_R = \mathbf{1}_r$ and $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{g}_Q = \mathbf{1}_r$. Thus $\boldsymbol{T}\mathbf{1}_r = \boldsymbol{g}_Q$ and $\boldsymbol{T}^{\mathrm{T}}\mathbf{1}_r = \boldsymbol{g}_R$ and $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}$. With analogous reasoning to before, one has a step for the coupling $\boldsymbol{T}$ in the form:

$$\min_{\boldsymbol{T}} \langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}, \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R) \rangle_F + \frac{1}{\gamma_k} \mathrm{KL}(\boldsymbol{T}_k \| \boldsymbol{K}_{\boldsymbol{T}}^{(k)})$$

Which yields the kernel matrix:

$$\boldsymbol{K}_{\boldsymbol{T}}^{(k)} = \boldsymbol{T}_k \odot \exp\{-\gamma_k\,\mathrm{diag}(\boldsymbol{g}_Q)^{-1}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}\,\mathrm{diag}(\boldsymbol{g}_R)^{-1}\}$$

Which is likewise projected onto $\Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}$ using the Sinkhorn algorithm. From this $\boldsymbol{T} \in \Pi_{\boldsymbol{g}_Q,\boldsymbol{g}_R}$, one takes $\boldsymbol{X} = \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)$ as the inner matrix that corresponds the unequal marginals $\boldsymbol{g}_Q$ and $\boldsymbol{g}_R$ which ensuring $\boldsymbol{P}_r \in \Pi_{\boldsymbol{a},\boldsymbol{b}}$.

Algorithm 4 FRLC *(General marginal constraint low-rank optimal transport)*

Input $\boldsymbol{C}, r, r_2, \boldsymbol{a}, \boldsymbol{b}, \tau, \tau_2, \gamma, \delta, \varepsilon$
Initialize $\boldsymbol{g}_Q, \boldsymbol{g}_R = \frac{1}{r}\mathbf{1}_r, \frac{1}{r_2}\mathbf{1}_{r_2}$
$\boldsymbol{Q}_0, \boldsymbol{R}_0, \boldsymbol{T}_0 \leftarrow$ *Initialize-Couplings*$(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{g}_Q, \boldsymbol{g}_R)$
if $r = r_2$ **then**
  $\boldsymbol{X}_0 \leftarrow \boldsymbol{T}_0^{-1}$ # Invertible case
else
  $\boldsymbol{X}_0 \leftarrow \mathrm{diag}(1/\boldsymbol{Q}_0^{\mathrm{T}}\mathbf{1}_n)\boldsymbol{T}_0\,\mathrm{diag}(1/\boldsymbol{R}_0^{\mathrm{T}}\mathbf{1}_m)$ # General case
end if
while $\Delta((\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k), (\boldsymbol{Q}_{k-1}, \boldsymbol{R}_{k-1}, \boldsymbol{T}_{k-1})) > \varepsilon$ **do**
  $\nabla_{\boldsymbol{Q}} \leftarrow \boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}} - \mathbf{1}_n\mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}_k\boldsymbol{X}_k^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}}$
  $\nabla_{\boldsymbol{R}} \leftarrow \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k - \mathbf{1}_m\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}_k^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_k\boldsymbol{X}_k)^{\mathrm{T}}$
  $\gamma_k \leftarrow \gamma/\max\{\|\nabla_{\boldsymbol{Q}}\|_\infty, \|\nabla_{\boldsymbol{R}}\|_\infty\}$ # ℓ^∞-normalization of Scetbon & Cuturi (2022)
  $\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \boldsymbol{K}_{\boldsymbol{R}}^{(k)} \leftarrow \boldsymbol{Q}_k \odot \exp(-\gamma_k\nabla_{\boldsymbol{Q}}), \boldsymbol{R}_k \odot \exp(-\gamma_k\nabla_{\boldsymbol{R}})$
  if Balanced **then**
    $\boldsymbol{Q}_k \leftarrow$ SR$^{\mathrm{R}}$-*projection*$(\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \gamma_k, \tau, \boldsymbol{a}, \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n, \delta)$ # Semi-relaxed OT
    $\boldsymbol{R}_k \leftarrow$ SR$^{\mathrm{R}}$-*projection*$(\boldsymbol{K}_{\boldsymbol{R}}^{(k)}, \gamma_k, \tau, \boldsymbol{b}, \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m, \delta)$ # Semi-relaxed OT
  else if Unbalanced **then**
    $\boldsymbol{Q}_k \leftarrow$ U-*projection*$(\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \gamma_k, \tau, \boldsymbol{a}, \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n, \delta)$ # Unbalanced OT
    $\boldsymbol{R}_k \leftarrow$ U-*projection*$(\boldsymbol{K}_{\boldsymbol{R}}^{(k)}, \gamma_k, \tau_2, \boldsymbol{b}, \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m, \delta)$ # Unbalanced OT
  else if Semi-Relaxed Left **then**
    $\boldsymbol{Q}_k \leftarrow$ U-*projection*$(\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \gamma_k, \tau, \boldsymbol{a}, \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n, \delta)$ # Unbalanced OT
    $\boldsymbol{R}_k \leftarrow$ SR$^{\mathrm{R}}$-*projection*$(\boldsymbol{K}_{\boldsymbol{R}}^{(k)}, \gamma_k, \tau, \boldsymbol{b}, \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m, \delta)$ # Semi-relaxed OT
  else if Semi-Relaxed Right **then**
    $\boldsymbol{Q}_k \leftarrow$ SR$^{\mathrm{R}}$-*projection*$(\boldsymbol{K}_{\boldsymbol{Q}}^{(k)}, \gamma_k, \tau, \boldsymbol{a}, \boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n, \delta)$ # Semi-relaxed OT
    $\boldsymbol{R}_k \leftarrow$ U-*projection*$(\boldsymbol{K}_{\boldsymbol{R}}^{(k)}, \gamma_k, \tau, \boldsymbol{b}, \boldsymbol{R}_{k-1}^{\mathrm{T}}\mathbf{1}_m, \delta)$ # Unbalanced OT
  end if
  $\boldsymbol{g}_Q, \boldsymbol{g}_R \leftarrow \boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n, \boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m$
  $\nabla_{\boldsymbol{T}} = \mathrm{diag}(\boldsymbol{g}_Q)^{-1}\boldsymbol{Q}_k^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}_k\,\mathrm{diag}(\boldsymbol{g}_R)^{-1}$
  $\gamma_{\boldsymbol{T}} = \gamma/\|\nabla_{\boldsymbol{T}}\|_\infty$ # ℓ^∞-normalization
  $\boldsymbol{K}_{\boldsymbol{T}}^{(k)} \leftarrow \boldsymbol{T}_k \odot \exp\{-\gamma_{\boldsymbol{T}}\nabla_{\boldsymbol{T}}\}$
  $\boldsymbol{T}_k \leftarrow$ *Sinkhorn*$(\boldsymbol{K}_{\boldsymbol{T}}^{(k)}, \boldsymbol{g}_R, \boldsymbol{g}_Q, \delta)$ # Balanced OT
  $\boldsymbol{X}_k \leftarrow \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{g}_R)$
end while
Return $\boldsymbol{P}_r = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$

Algorithm 5 Sinkhorn Algorithm *(Cuturi (2013b), balanced OT)*

Input $\boldsymbol{K}, \boldsymbol{a}, \boldsymbol{b}, \delta$
$\boldsymbol{u} \leftarrow \mathbf{1}_n$
$\boldsymbol{v} \leftarrow \mathbf{1}_m$
while $\|\mathrm{diag}(\boldsymbol{u})\boldsymbol{K}\boldsymbol{v} - \boldsymbol{a}\|_1 + \|\mathrm{diag}(\boldsymbol{v})\boldsymbol{K}^{\mathrm{T}}\boldsymbol{u} - \boldsymbol{b}\|_1 > \delta$ **do**
  $\boldsymbol{u}^{(l+1)} \leftarrow \boldsymbol{a}/\boldsymbol{K}\boldsymbol{v}^{(l)}$
  $\boldsymbol{v}^{(l+1)} \leftarrow \boldsymbol{b}/\boldsymbol{K}^{\mathbf{T}}\boldsymbol{u}^{(l+1)}$
end while
Return $\mathrm{diag}(\boldsymbol{u})\boldsymbol{K}\,\mathrm{diag}(\boldsymbol{v})$

D Gromov-Wasserstein (GW)

As defined in, the general Gromov-Wasserstein problem concerns a minimization of the energy:

$$\mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P}) = \sum_{i,j,k,l} (\boldsymbol{A}_{ik} - \boldsymbol{B}_{jl})^2 \boldsymbol{P}_{ij}\boldsymbol{P}_{kl} \tag{30}$$

Where the minimization is over the set of all couplings $\mathbf{\Pi}_{\boldsymbol{a},\boldsymbol{b}}$:

$$\mathrm{GW}(\mu,\nu) := \min_{\boldsymbol{P}\in\Pi_{\boldsymbol{a},\boldsymbol{b}}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P}) \tag{31}$$

We consider extending the semi-relaxed framework to the GW problem under the low-rank restriction on $\boldsymbol{P}$. This extension has thus far been considered for the balanced and unbalanced case in two previous works Scetbon et al. (2023, 2022). In both of these works, the algorithms for the Wasserstein case extend trivially to the Gromov-Wasserstein (GW) problem. In particular, each kernel with variable $\boldsymbol{\zeta}$ has an update of the form $\boldsymbol{K} \leftarrow \boldsymbol{\zeta} \odot \exp\left(-\gamma_k \nabla_{\boldsymbol{\zeta}} \mathcal{L}(\boldsymbol{\zeta})\right)$ for $\mathcal{L}(\boldsymbol{\zeta})$ heretofore taken to be the Wasserstein loss $\langle \boldsymbol{P}(\boldsymbol{\zeta}), \boldsymbol{C}\rangle_F$ where the coupling $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}$ is interpreted as a function of the low-rank sub-factor variables $\boldsymbol{\zeta} \in \{\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{X}\}$. Taking $\mathcal{L} := \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P}(\boldsymbol{\zeta}))$ one can simply extend the gradient through the GW-loss and directly use it in place of the Wasserstein gradient in the update $\boldsymbol{K} \leftarrow \boldsymbol{\zeta} \odot \exp\left(-\gamma_k \nabla_{\boldsymbol{\zeta}} \mathcal{L}(\boldsymbol{\zeta})\right)$ for $\mathcal{L}(\boldsymbol{\zeta})$ of each algorithm. The matrix-form of the GW-cost is expressed as:

$$\mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{P}) = \mathbf{1}_m^{\mathrm{T}} \boldsymbol{P}^{\mathrm{T}} \boldsymbol{A}^{\odot 2} \boldsymbol{P} \mathbf{1}_m + \mathbf{1}_n^{\mathrm{T}} \boldsymbol{P} \boldsymbol{B}^{\odot 2} \boldsymbol{P}^{\mathrm{T}} \mathbf{1}_n - 2\langle \boldsymbol{A}\boldsymbol{P}\boldsymbol{B}, \boldsymbol{P}\rangle$$

Which, using the constraints of $\mathcal{C}_2$ reduces the cost as a function of $\boldsymbol{Q}$, $\boldsymbol{R}$, $\boldsymbol{X}$ to:

$$\mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{X}) = \mathbf{1}_r^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} \boldsymbol{A}^{\odot 2} \boldsymbol{Q} \mathbf{1}_r + \mathbf{1}_r^{\mathrm{T}} \boldsymbol{R}^{\mathrm{T}} \boldsymbol{B}^{\odot 2} \boldsymbol{R} \mathbf{1}_r - 2\langle \boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\rangle_F$$

$$\nabla_{\boldsymbol{Q}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{X}) = 2\boldsymbol{A}^{\odot 2}\boldsymbol{Q}\mathbf{1}_r\mathbf{1}_r^{\mathrm{T}} - 4\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}$$

Which is proportional to $\nabla_{\boldsymbol{Q}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{X}) \propto -4\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}$ for the balanced and right-marginal semi-relaxed case. And:

$$\nabla_{\boldsymbol{R}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{X}) = 2\boldsymbol{B}^{\odot 2}\boldsymbol{R}\mathbf{1}_r\mathbf{1}_r^{\mathrm{T}} - 4\boldsymbol{B}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}$$

Which likewise can be reduced in proportionality to $\nabla_{\boldsymbol{R}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{g}) \propto -4\boldsymbol{B}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}$ in the balanced and left-marginal semi-relaxed case. The gradients, as presented above, assume a linearization in $\boldsymbol{X} \leftarrow \boldsymbol{X}_k$. If one does not make this assumption and takes $\boldsymbol{X}(\boldsymbol{Q},\boldsymbol{R}) = \mathrm{diag}(1/\boldsymbol{Q}^T\mathbf{1}_n)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{R}^T\mathbf{1}_m)$, a rank-one perturbation must be added to the $\boldsymbol{Q}$ and $\boldsymbol{R}$ gradient of the form:

$$\nabla_{\boldsymbol{Q}}^{(2)} = 4\mathbf{1}_n\,\mathrm{diag}^{-1}(\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}(\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}}))^{\mathrm{T}}$$
$$\nabla_{\boldsymbol{R}}^{(2)} = 4\mathbf{1}_m\,\mathrm{diag}^{-1}(\boldsymbol{X}^T\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}\,\mathrm{diag}(1/\boldsymbol{g}_R))^{\mathrm{T}}$$

Analogously, for the gradient on $\boldsymbol{T}$ one can simply take the gradient with respect to $\boldsymbol{X}$, and subsequently $\boldsymbol{T}$ as $\boldsymbol{X}(\boldsymbol{T}) = \mathrm{diag}(\boldsymbol{g}_Q)^{-1}\boldsymbol{T}\,\mathrm{diag}(\boldsymbol{g}_R)^{-1}$. The gradient with respect to $\boldsymbol{X}$ is given as:

$$\nabla_{\boldsymbol{X}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{X}) = -4\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}$$

And thus, the directional derivative with respect to $\boldsymbol{X}$ in the direction $\boldsymbol{V}_{\boldsymbol{X}} = \mathrm{diag}(\boldsymbol{g}_Q)^{-1}\boldsymbol{V}_{\boldsymbol{T}}\,\mathrm{diag}(\boldsymbol{g}_R)^{-1}$ and thus by the chain rule $\boldsymbol{V}_{\boldsymbol{T}}$ is:

$$\begin{aligned} D\mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{X}) \circ (\boldsymbol{V}_{\boldsymbol{X}}) &= -4\langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}, \boldsymbol{V}_{\boldsymbol{X}}\rangle_F \\ &= -4\langle \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}, \mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{V}_{\boldsymbol{T}}\,\mathrm{diag}(1/\boldsymbol{g}_R)\rangle_F \end{aligned}$$

So that the gradient with respect to the coupling matrix $\boldsymbol{T}$ is given as:

$$\nabla_{\boldsymbol{T}} \mathcal{Q}_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T}) = -4\,\mathrm{diag}(1/\boldsymbol{g}_Q)\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}\boldsymbol{X}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{R}\,\mathrm{diag}(1/\boldsymbol{g}_R)$$

E Convergence Analysis and Other Proofs

E.1 Convergence and Smoothness of the Objective

We show in Proposition 3.3 that directly applying the block-descent lemma of Beck & Tetruashvili (2013) to the template of Ghadimi et al.'s proof Ghadimi et al. (2014) is sufficient to show the non-asymptotic stationary convergence of a coordinate mirror descent procedure in Ghadimi's criterion. The non-asymptotic guarantee of Ghadimi et al. (2014) follows directly in the case of coordinate mirror descent using Lemma E.1 and Lemma E.2 below. For completeness, we define all notation used, describe the coordinate mirror descent algorithm in general, and discuss a few relevant preliminaries.

Suppose that the vector of n variables $\mathbf{x} \in \mathbb{R}^n$ is partitioned into p blocks, $\mathbf{x} = (\mathbf{x}(1), \ldots, \mathbf{x}(p))$, where $\mathbf{x}(i) \in \mathbb{R}^{n_i}$. Here, $n_1, \ldots, n_p$ are positive integers summing to n. Following the notation of Beck & Tetruashvili (2013); Nesterov (2012), we define matrices $\mathbf{U}_i \in \mathbb{R}^{n \times n_i}$ such that $\mathbf{x}(i) = \mathbf{U}_i^{\mathrm{T}} \mathbf{x}$ for all $i = 1, \ldots, p$. This also implies $\mathbf{x} = \sum_{i=1}^{p} \mathbf{U}_i \mathbf{x}(i)$. This allows us to define the vector of partial derivatives corresponding to each block of variables $\mathbf{x}(i)$:

$$\nabla_i f(\mathbf{x}) := \mathbf{U}_i^{\mathrm{T}} \nabla f(\mathbf{x}).$$

In Beck & Tetruashvili (2013), the gradient of f is assumed to be block-coordinate-wise Lipschitz, with L_i the smoothness constant associated to the i-th block of variables: for all $\mathbf{h}_i \in \mathbb{R}^{n_i}$, one has

$$\|\nabla_i f(\mathbf{x} + \mathbf{U}_i \mathbf{h}_i) - \nabla_i f(\mathbf{x})\| \leq L_i \|\mathbf{h}_i\|, \tag{32}$$

and for such functions we denote by $L := \max_i L_i$ the (global) smoothness constant of ∇f. To be clear, a *smoothness constant* associated to f is a Lipschitz constant of its gradient.

Lemma E.1 (Block descent lemma, Beck & Tetruashvili (2013), Lemma 3.2.)**.** *Suppose $f \in C^1(\mathbb{R}^n, \mathbb{R})$ is a continuously differentiable function over $\mathbb{R}^n$ whose gradient is block-coordinatewise Lipschitz* (32) *for L_i the smoothness constant associated to the i-th block of variables $\mathbf{x}(i)$. Let $\mathbf{u}, \mathbf{v}$ be two vectors differing only in the i-th block: there exists $\mathbf{h}_i \in \mathbb{R}^{n_i}$ such that $\mathbf{v} - \mathbf{u} = \mathbf{U}_i \mathbf{h}_i$. Then,*

$$f(\mathbf{v}) \leq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle + \frac{L_i}{2} \|\mathbf{u} - \mathbf{v}\|^2. \tag{33}$$

Lemma E.1 is central to adapting the proof of Theorem 1 of Ghadimi et al. (2014) to our case. Their Theorem 1 concerns the non-asymptotic convergence of mirror descent for objectives of the form

$$\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + h(\mathbf{x}),$$

where $\mathcal{X}$ is a closed, convex subset of $\mathbb{R}^n$, $f \in C^1(\mathcal{X}, \mathbb{R})$ is a possibly non-convex objective, and where h is an α-strongly convex function. Using notation similar to Ghadimi et al. (2014), we write $\Phi(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$. Additionally, Ghadimi et al. (2014) assume ∇f is L-Lipschitz for some $L > 0$.

To unify our assumptions on f, we suppose that $\nabla f \in C^1(\mathcal{X}, \mathbb{R})$ is block-coordinate Lipschitz (32) with block Lipschitz constants $(L_i)_{i=1}^p$, and that $\mathcal{X}$ itself decomposes as a product $\mathcal{X} = \prod_{i=1}^p \mathcal{X}_i$, where each $\mathcal{X}_i$ is a closed convex set constraining the block variables $\mathbf{x}(i)$.

The proof of Ghadimi relies on β-smoothness of the objective in all variables. We show that component-wise smoothness in each block is sufficient to achieve an analogous convergence result for a coordinate mirror descent. To provide context for Proposition 3.3, we now describe in general (1) mirror descent, and (2) block-coordinate mirror descent.

We again follow the notation of Ghadimi et al. (2014). A function $\omega : \mathcal{X} \to \mathbb{R}$ is a *distance generating function* with modulus $\alpha > 0$, with respect to the Euclidean norm $\|\cdot\|$, if ω is continuously differentiable and strongly convex, so that

$$\langle \mathbf{x} - \mathbf{z}, \nabla \omega(\mathbf{x}) - \nabla \omega(\mathbf{z}) \rangle \geq \alpha \|\mathbf{x} - \mathbf{z}\|^2, \quad \text{for all } \mathbf{x}, \mathbf{z} \in \mathcal{X}.$$

The *prox-function* (or Bregman divergence) associated with ω is then

$$V(\mathbf{x}, \mathbf{z}) = \omega(\mathbf{x}) - \omega(\mathbf{z}) - \langle \nabla \omega(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle,$$

and from this prox-function, $\gamma > 0$, and some $\mathbf{g} \in \mathbb{R}^n$, we define the *generalized projection*

$$\mathbf{x}^+ := \arg\min_{\mathbf{u} \in \mathcal{X}} \left(\langle \mathbf{g}, \mathbf{u} \rangle + \frac{1}{\gamma} V(\mathbf{u}, \mathbf{x}) + h(\mathbf{u}) \right). \tag{34}$$

To describe the projected gradient descent algorithm (which coincides with mirror descent in our case of interest), we first define the *generalized projected gradient of Φ at $\mathbf{x}$*:

$$P_{\mathcal{X}}(\mathbf{x}, \mathbf{g}, \gamma) := \frac{1}{\gamma}(\mathbf{x} - \mathbf{x}^+).$$

The *mirror descent (MD) algorithm* is as follows: given initial point $\mathbf{x}_0 \in \mathcal{X}$, a total number of iterations N, and positive stepsizes $(\gamma_k)_{k=1}^N$, at step k, the $(k+1)$-st iterate is computed via

$$\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{u} \in \mathcal{X}} \left(\langle \nabla f(\mathbf{x}_k), \mathbf{u} \rangle + \frac{1}{\gamma_k} V(\mathbf{u}, \mathbf{x}_k) + h(\mathbf{u}) \right).$$

Among all iterates $\mathbf{x}_k$, the MD algorithm outputs the one at which the generalized projected gradient is of least norm. Concretely, $\mathbf{x}_R$ is the output of the MD algorithm, where

$$R := \operatorname*{arg\,min}_{k=0,\ldots,N} \|\mathbf{g}_{\mathcal{X},k}\|^2,$$

and where $\mathbf{g}_{\mathcal{X},k}$ is

$$\mathbf{g}_{\mathcal{X},k} := P_{\mathcal{X}}(\mathbf{x}_k, \nabla f(\mathbf{x}_k), \gamma_k).$$

Having described the MD algorithm, let us consider a block-coordinate variant; we assume $\mathbf{x}$ admits the block-coordinate structure described above. To simplify the presentation, we suppose that in a given iteration k, the block variables are updated sequentially from $i = 1, \ldots, p$. This leads to doubly-indexed iterates $(\mathbf{x}_k^i)$ with $k = 1, \ldots, N$ indexing each full iteration through all variables, and $i = 0, 1, \ldots, p$ indexing the sub-iterations which update one block of variables at a time.

The *coordinate mirror descent (CMD) algorithm* takes as input an initial point $\mathbf{x}_0 \in \mathcal{X}$, a number of iterations N, and a sequence of positive stepsizes $(\gamma_{k,i})_{k=1,i=1}^{N,p}$. We set $\mathbf{x}_0^0 = \mathbf{x}_0$, and for $k = 0, \ldots, N-1$ and $i = 1, \ldots, p$, we compute $\mathbf{x}_k^i$ from $\mathbf{x}_k^{i-1}$ as follows:

$$\begin{aligned}
\mathbf{x}_k^i(i) &\leftarrow \operatorname*{arg\,min}_{\mathbf{u}_i \in \mathcal{X}_i} \left(\langle \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{u}_i \rangle + \frac{1}{\gamma_{k,i}} V_i(\mathbf{u}_i, \mathbf{U}_i^{\mathrm{T}} \mathbf{x}_k^{i-1}) + h_i(\mathbf{u}_i) \right) \\
\mathbf{x}_k^i(j) &\leftarrow \mathbf{x}_k^{i-1}(j) \quad \text{for } j \neq i
\end{aligned}$$

Here, we have assumed that V can be written as a composite function of the block variables,

$$V(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{p} V_i(\mathbf{x}(i), \mathbf{z}(i)),$$

as is the case for the KL divergence. We also have assumed that h has this composite structure (as with entropy):

$$h(\mathbf{x}) = \sum_{i=1}^{p} h_i(\mathbf{x}(i)).$$

Lastly, for $k = 0, \ldots, N-1$, we set $\mathbf{x}_{k+1}^0 := \mathbf{x}_k^p$. We define $\mathbf{g}_{\mathcal{X},k} := (\mathbf{g}_{\mathcal{X},k,1}, \ldots, \mathbf{g}_{\mathcal{X},k,p})$ to be the collection of block-wise differences, where by definition

$$\mathbf{g}_{\mathcal{X},k,i} = P_{\mathcal{X}_i}(\mathbf{x}_k^{i-1}, \nabla_i f(\mathbf{x}_k^{i-1}), \gamma_{k,i}) = \frac{1}{\gamma_{k,i}} \left(\mathbf{x}_k^{i-1} - \mathbf{x}_k^i \right) = \mathbf{U}_i^{\mathrm{T}} \mathbf{g}_{\mathcal{X},k}.$$

The convergence criterion Δ we use in Proposition 3.3 is the one used by Ghadimi et al. (2014) summed across blocks. In particular, the CMD algorithm returns iterate $\mathbf{x}_R$, where

$$\begin{aligned}
R &:= \operatorname*{arg\,min}_{k=0,\ldots,N} \Delta(\mathbf{x}_k, \mathbf{x}_{k-1}), \\
\Delta(\mathbf{x}_k, \mathbf{x}_{k-1}) &:= \|\mathbf{g}_{\mathcal{X},k}\|^2 = \sum_{i=1}^{p} \|\mathbf{g}_{\mathcal{X},k,i}\|^2.
\end{aligned} \tag{35}$$

For $\Phi = f + h$ as above (f has global smoothness constant L), let $\Phi^* = \arg\min_{\mathbf{x} \in \mathcal{X}} \Phi(\mathbf{x})$, and define

$$D := \left(\frac{\Phi(x_0) - \Phi^*}{L} \right)^{1/2}. \tag{36}$$

The other lemma used in the proof of Proposition 3.3 is as follows.

Lemma E.2 (Ghadimi et al. (2014), Lemma 1)**.** *Let* $\mathbf{x}^+ = \arg\min_{\mathbf{u} \in \mathcal{X}} \{ \langle \mathbf{g}, \mathbf{u} \rangle + \frac{1}{\gamma} V(\mathbf{u}, \mathbf{x}) + h(\mathbf{u}) \}$ *and* $P_{\mathcal{X}}(\mathbf{x}, \mathbf{g}, \gamma) = \frac{1}{\gamma}(\mathbf{x} - \mathbf{x}^+)$*. Then for all* $\mathbf{x} \in \mathbb{R}^n$*, all* $\mathbf{g} \in \mathbb{R}^n$*, and* $\gamma > 0$*, one has:*

$$\langle \mathbf{g}, P_{\mathcal{X}}(\mathbf{x}, \mathbf{g}, \gamma) \rangle \geq \alpha \|P_{\mathcal{X}}(\mathbf{x}, \mathbf{g}, \gamma)\|^2 + \frac{1}{\gamma}(h(\mathbf{x}^+) - h(\mathbf{x})).$$

Proposition E.3 (Proposition 3.3)**.** *Suppose one has $f \in C^1(\mathcal{X}, \mathbb{R})$ whose gradient is block-coordinate Lipschitz, with block smoothness constants $(L_i)_{i=1}^p$, and a function $h \in C(\mathcal{X}, \mathbb{R})$ which is α-strongly convex. For $\Phi = f + h$, suppose one performs a coordinate mirror descent on Φ minimized over a product of closed convex sets $\mathcal{X} = \prod_{i=1}^p \mathcal{X}_i$. Let the sub-iterates with respect to the i-th block update be $\{\mathbf{x}_k^i\}_{i=0}^p$ where $\mathbf{x}_k := \mathbf{x}_k^0$ for $k \in [N]$ outer iterations. Then one has:*

$$\min_k \Delta(\mathbf{x}_k, \mathbf{x}_{k-1}) \leq \frac{D^2 L}{N(\alpha^2/2L)} = \frac{2D^2L^2}{N\alpha^2},$$

where D is (36)*, L is the global smoothness constant of f, and convergence criterion $\Delta(\mathbf{x}_k, \mathbf{x}_{k-1})$ is given in* (35)*. Above, the stepsizes $\gamma_{k,i}$ in the coordinate mirror descent are $\gamma_{k,i} := \alpha/L$.*

Proof. As f satisfies the hypotheses of the block descent lemma, Lemma E.1, we apply (33) to obtain:

$$f(\mathbf{x}_k^i) \leq f(\mathbf{x}_k^{i-1}) + \langle \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{x}_k^i - \mathbf{x}_k^{i-1} \rangle + \frac{L_i}{2} \|\mathbf{x}_k^i - \mathbf{x}_k^{i-1}\|^2.$$

Noting the definition $\mathbf{g}_{\mathcal{X},k,i} = \frac{1}{\gamma_{k,i}} \left(\mathbf{x}_k^{i-1} - \mathbf{x}_k^i \right)$, one has

$$f(\mathbf{x}_k^i) \leq f(\mathbf{x}_k^{i-1}) - \gamma_{k,i} \langle \nabla_i f(\mathbf{x}_k^{i-1}), \mathbf{g}_{\mathcal{X},k,i} \rangle + \frac{L_i}{2} \gamma_{k,i}^2 \|\mathbf{g}_{\mathcal{X},k,i}\|^2.$$

Lemma 1 of Ghadimi et al. (2014) (stated as Lemma E.2 above) applies identically through block-wise optimality on $\mathcal{X}_i$ because $\mathbf{g}_{\mathcal{X},k,i} = P_{\mathcal{X}_i}(\mathbf{x}_k^{i-1}, \nabla_i f(\mathbf{x}_k^{i-1}), \gamma_{k,i})$. Thus for any value $\nabla_i f(\mathbf{x}_k^{i-1})$ takes,

$$f(\mathbf{x}_k^i) \leq f(\mathbf{x}_k^{i-1}) - \left[\alpha \gamma_{k,i} \|\mathbf{g}_{\mathcal{X},k,i}\|^2 + h(\mathbf{x}_k^i) - h(\mathbf{x}_k^{i-1}) \right] + \frac{L_i}{2} \gamma_{k,i}^2 \|\mathbf{g}_{\mathcal{X},k,i}\|^2,$$

and thus,

$$f(\mathbf{x}_k^i) + h(\mathbf{x}_k^i) \leq f(\mathbf{x}_k^{i-1}) + h(\mathbf{x}_k^{i-1}) - \left[\alpha \gamma_{k,i} - \frac{L_i}{2} \gamma_{k,i}^2 \right] \|g_{\mathcal{X},k,i}\|^2.$$

The right-hand side above only becomes larger, taking $L_i = L$ to be the global smoothness constant. Introducing a sum over sub-iterates and total iterates, one has

$$\sum_{k,i}^{N,p} f(\mathbf{x}_k^i) + h(\mathbf{x}_k^i) \leq \sum_{k,i}^{N,p} f(\mathbf{x}_k^{i-1}) + h(\mathbf{x}_k^{i-1}) - \sum_{k,i}^{N,p} \left[\alpha \gamma_{k,i} - \frac{L}{2} \gamma_{k,i}^2 \right] \|\mathbf{g}_{\mathcal{X},k,i}\|^2,$$

$$\sum_{k,i}^{N,p} \Phi(\mathbf{x}_k^i) \leq \sum_{k,i}^{N,p} \Phi(\mathbf{x}_k^{i-1}) - \sum_{k,i}^{N,p} \left[\alpha \gamma_{k,i} - \frac{L}{2} \gamma_{k,i}^2 \right] \|\mathbf{g}_{\mathcal{X},k,i}\|^2.$$

Noting the end-point condition $\Phi(\mathbf{x}_k^p) = \Phi(\mathbf{x}_{k+1}^0)$, one may cancel all intermediate terms:

$$\Phi^* \leq \Phi(\mathbf{x}_N) \leq \Phi(\mathbf{x}_0) - \sum_{k,i}^{N,p} \left[\alpha \gamma_{k,i} - \frac{L}{2} \gamma_{k,i}^2 \right] \|\mathbf{g}_{\mathcal{X},k,i}\|^2$$

Thus one finds the upper bound in terms of $\Phi(\mathbf{x}_0)$ and the minimum value $\Phi^* = \min_{\mathbf{x} \in \mathcal{X}} \Phi(\mathbf{x})$ of Φ:

$$\sum_k^N \sum_i^p \left[\alpha \gamma_{k,i} - \frac{L}{2} \gamma_{k,i}^2 \right] \|\mathbf{g}_{\mathcal{X},k,i}\|^2 \leq \Phi(\mathbf{x}_0) - \Phi^*. \tag{37}$$

Taking $\gamma_{k,i} = \alpha/L$ as in Ghadimi et al. (2014), the bracketed term directly above becomes $\alpha^2/2L$, and one has:

$$\begin{aligned}
\sum_k^N \left[\frac{\alpha^2}{2L} \right] \left(\min_k \Delta(\mathbf{x}_k, \mathbf{x}_{k-1}) \right) &= \sum_k^N \left[\frac{\alpha^2}{2L} \right] \left(\min_k \sum_i^p \|\mathbf{g}_{\mathcal{X},k,i}\|^2 \right) \\
&\leq \sum_k^N \left[\frac{\alpha^2}{2L} \right] \sum_i^p \|\mathbf{g}_{\mathcal{X},k,i}\|^2 \\
&\leq \Phi(\mathbf{x}_0) - \Phi^* = D^2 L,
\end{aligned}$$

where D is as in (36), and where we used (37) to obtain the last line. Thus,

$$\min_k \Delta(\mathbf{x}_k, \mathbf{x}_{k-1}) \leq \frac{D^2 L}{N(\alpha^2/2L)} = \frac{2D^2L^2}{N\alpha^2},$$

completing the proof. □

Definition E.4 (Relative smoothness)**.** Let $\beta > 0$ and let $g \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ be continuously differentiable. Additionally, let ω be a distance generating function with associated prox-function V. The function g is β*-smooth relative to* ω if the following holds:

$$g(\boldsymbol{y}) \leq g(\boldsymbol{x}) + \langle \nabla\omega(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{y}\rangle + \beta V(\boldsymbol{y}, \boldsymbol{x})$$

Proposition E.5. *Let $\epsilon > 0$ be a predefined error tolerance, $r > 0$ a small rank parameter, $\delta \in (0, \frac{1}{r})$ a lower-bound parameter, and N the number of inner iterations required for the semi-relaxed projection to converge for each iteration k as $\|\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n - \boldsymbol{g}_{\boldsymbol{Q}}^{(k-1)}\|_2 < \epsilon = \frac{1}{N}\left(\frac{1}{r} - \delta\right)$ for $\boldsymbol{g}_{\boldsymbol{Q}}^{(0)} = \frac{1}{r}\mathbf{1}_r$. Then, the FRLC objective*

$$\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) = \langle \boldsymbol{Q}\,\mathrm{diag}(1/\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n)\boldsymbol{T}\,\mathrm{diag}(1/\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m)\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C}\rangle_F$$

is component-wise smooth with respect to the variables $\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}$ with smoothness constants $\{\beta_i\}_{i=1}^3$ where $\beta_i = \mathrm{poly}(\|\boldsymbol{C}\|_F, n, m, r, \delta)$.

Proof. The addition of a pair of regularizations on the inner marginals $\tau\mathrm{KL}(\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n\|\boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n)$ and $\tau\mathrm{KL}(\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m\|\boldsymbol{R}_k^{\mathrm{T}}\mathbf{1}_m)$ in 21 and 22 ensures that we may use standard results on relaxed optimal-transport which bound how far the marginal $\boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n$ deviates across iterations.

In particular, for all $\epsilon > 0$ there exists τ and N (number of iterations) sufficiently large, so that $\|\boldsymbol{R}^{\mathrm{T}}\mathbf{1}_m - \boldsymbol{g}_{\boldsymbol{R}}\|_2^2 < \epsilon$ and $\|\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_n - \boldsymbol{g}_{\boldsymbol{Q}}\|_2^2 < \epsilon$. In particular, Pham et al. (2020) shows that one can attain convergence to any ϵ for the unbalanced problem in $N = \tilde{O}(m^2/\epsilon)$ iterations (hiding logarithmic and τ-factors) for each sub-problem solving for $\boldsymbol{R}_k$ in Algorithm 4 and analogously $N = \tilde{O}(n^2/\epsilon)$ for $\boldsymbol{Q}$. Under the uniform initialization of $\boldsymbol{g}_{\boldsymbol{R}}$, and after N iterations, we have:

$$\|\boldsymbol{g}_{\boldsymbol{R}}^{(0)} - \boldsymbol{g}_{\boldsymbol{R}}^{(N)}\|_2 = \left\|\frac{1}{r}\mathbf{1}_r - \boldsymbol{g}_{\boldsymbol{R}}^{(N)}\right\|_2 \leq \sum_{k=1}^{N}\|\boldsymbol{g}_{\boldsymbol{R}}^{(k)} - \boldsymbol{g}_{\boldsymbol{R}}^{(k-1)}\|_2.$$

With sufficiently large τ and $N = \tilde{O}(m^2/\epsilon)$ sub-iterations, one can guarantee that:

$$\|\boldsymbol{g}_{\boldsymbol{R}}^{(k)} - \boldsymbol{g}_{\boldsymbol{R}}^{(k-1)}\|_2 < \epsilon = \frac{1}{N}\left(\frac{1}{r} - \delta\right).$$

This implies, for all iterations of the algorithm and all indices i, that

$$(\boldsymbol{g}_{R_k})_i > \delta, \quad \text{and analogously,} \quad (\boldsymbol{g}_{Q_k})_i > \delta. \tag{38}$$

Thus, by adding the regularization on the inner marginal, one may guarantee a lower-bound on the entries of $\boldsymbol{g}_R$ and $\boldsymbol{g}_Q$. This is essential for demonstrating smoothness of the objective.

First, we consider smoothness in $\boldsymbol{Q}$. We note that the gradient in $\boldsymbol{Q}$ splits into two terms:

$$\begin{aligned}
\nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) &= \nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}} + \nabla_{\boldsymbol{Q}}^{(B)}\mathcal{L}_{\mathrm{LC}} \\
&= \boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}} - \mathbf{1}_n\mathrm{diag}^{-1}((\boldsymbol{C}\boldsymbol{R}\boldsymbol{X}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}} \\
&= \nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}} - \mathbf{1}_n\mathrm{diag}^{-1}((\nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}})^{\mathrm{T}}\boldsymbol{Q}\mathrm{diag}(1/\boldsymbol{g}_Q))^{\mathrm{T}}
\end{aligned}$$

Where

$$\begin{aligned}
&\|\nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_k, \boldsymbol{T}_k) - \nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k)\|_F \\
&\quad\leq \|\nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_k, \boldsymbol{T}_k) - \nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k)\|_F \\
&\qquad + \|\nabla_{\boldsymbol{Q}}^{(B)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_k, \boldsymbol{T}_k) - \nabla_{\boldsymbol{Q}}^{(B)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k)\|_F.
\end{aligned} \tag{39}$$

Starting with the first term on the right side of (39), one has:

$$\begin{aligned}
&\|\nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1},\boldsymbol{R}_k,\boldsymbol{T}_k)-\nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k,\boldsymbol{R}_k,\boldsymbol{T}_k)\|_F\\
&\quad=\|\boldsymbol{C}\boldsymbol{R}_k(\mathrm{diag}(1/\boldsymbol{g}_{Q_k})\boldsymbol{T}_k\mathrm{diag}(1/\boldsymbol{g}_{R_k}))^{\mathrm{T}}-\boldsymbol{C}\boldsymbol{R}_k(\mathrm{diag}(1/\boldsymbol{g}_{Q_{k-1}})\boldsymbol{T}_k\mathrm{diag}(1/\boldsymbol{g}_{R_k}))^{\mathrm{T}}\|_F\\
&\quad\leq\|\mathrm{diag}(1/\boldsymbol{g}_{R_k})\|_F\|\boldsymbol{C}\|_F\|\boldsymbol{R}_k\|_F\|\boldsymbol{T}_k\|_F\|\mathrm{diag}(1/\boldsymbol{g}_{Q_k})-\mathrm{diag}(1/\boldsymbol{g}_{Q_{k-1}})\|_F.
\end{aligned}$$

Note that $\|\boldsymbol{R}_k\|_F^2=\sum_{i,j}(\boldsymbol{R}_k)_{i,j}^2<\sum_{i,j}(\boldsymbol{R}_k)_{i,j}=1$, as $\boldsymbol{R}_k$ has marginals which sum to one. The same bound holds for $\|\boldsymbol{T}_k\|_F^2$, which is also a coupling. Invoking the lower-bound (38) of δ on the entries of the inner marginals, and continuing from the above display,

$$\begin{aligned}
&\|\nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1},\boldsymbol{R}_k,\boldsymbol{T}_k)-\nabla_{\boldsymbol{Q}}^{(A)}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k,\boldsymbol{R}_k,\boldsymbol{T}_k)\|_F\\
&\quad\leq\frac{\|\boldsymbol{C}\|_F}{\delta}\|\mathrm{diag}(1/\boldsymbol{g}_{Q_k})-\mathrm{diag}(1/\boldsymbol{g}_{Q_{k-1}})\|_F\\
&\quad=\frac{\|\boldsymbol{C}\|_F}{\delta}\|\mathrm{diag}(1/\boldsymbol{g}_{Q_{k-1}})\mathrm{diag}(1/\boldsymbol{g}_{Q_k})(\mathrm{diag}(\boldsymbol{g}_{Q_{k-1}})-\mathrm{diag}(\boldsymbol{g}_{Q_k}))\|_F\\
&\quad\leq\frac{\|\boldsymbol{C}\|_F}{\delta^3}\|\mathrm{diag}(\boldsymbol{g}_{Q_{k-1}})-\mathrm{diag}(\boldsymbol{g}_{Q_k})\|_F.
\end{aligned}$$

To further bound the right-hand side above, consider:

$$\|\mathrm{diag}(\boldsymbol{g}_{\boldsymbol{Q}_k})-\mathrm{diag}(\boldsymbol{g}_{\boldsymbol{Q}_{k-1}})\|_F^2=\|\boldsymbol{Q}_k^{\mathrm{T}}\mathbf{1}_n-\boldsymbol{Q}_{k-1}^{\mathrm{T}}\mathbf{1}_n\|_2^2=\sum_{i=1}^{r}\left(\sum_{j=1}^{r}(\boldsymbol{Q}_k)_{i,j}-(\boldsymbol{Q}_{k-1})_{i,j}\right)^2$$

While can easily be upper-bounded by an application of Jensen's inequality as

$$\begin{aligned}
&=\sum_{i=1}^{r}r^2\left(\sum_{j=1}^{r}\frac{1}{r}\left((\boldsymbol{Q}_k)_{i,j}-(\boldsymbol{Q}_{k-1})_{i,j}\right)\right)^2\\
&\leq\sum_{i=1}^{r}r^2\left(\sum_{j=1}^{r}\frac{1}{r}\left((\boldsymbol{Q}_k)_{i,j}-(\boldsymbol{Q}_{k-1})_{i,j}\right)^2\right)\\
&=r\sum_{i=1}^{r}\sum_{j=1}^{r}\left((\boldsymbol{Q}_k)_{i,j}-(\boldsymbol{Q}_{k-1})_{i,j}\right)^2\\
&=r\|\boldsymbol{Q}_k-\boldsymbol{Q}_{k-1}\|_F^2.
\end{aligned}$$

Likewise, we have that $\|\mathrm{diag}(\boldsymbol{g}_{\boldsymbol{R}_k})-\mathrm{diag}(\boldsymbol{g}_{\boldsymbol{R}_{k-1}})\|_F^2\leq r\|\boldsymbol{R}_k-\boldsymbol{R}_{k-1}\|_F^2$. Thus, it holds that

$$\frac{\|\boldsymbol{C}\|_F}{\delta^3}\|\boldsymbol{g}_{\boldsymbol{Q}_k}-\boldsymbol{g}_{\boldsymbol{Q}_{k-1}}\|_2\leq\frac{\|\boldsymbol{C}\|_F\sqrt{r}}{\delta^3}\|\boldsymbol{Q}_k-\boldsymbol{Q}_{k-1}\|_F.$$

Next, we focus on the $\nabla_{\boldsymbol{Q}}^{(B)}$ term. Observe that

$$\begin{aligned}
\|\mathbf{1}_n\mathrm{diag}^{-1}\mathbf{X}\|_F^2&=\mathrm{Tr}(\mathbf{1}_n\mathrm{diag}^{-1}\mathbf{X})^{\mathrm{T}}(\mathbf{1}_n\mathrm{diag}^{-1}\mathbf{X})\\
&=n\|\mathrm{diag}^{-1}\mathbf{X}\|_2^2\leq n\|\mathbf{X}\|_F^2.
\end{aligned}$$

Thus:

$$\|\nabla_{\boldsymbol{Q}_{k+1}}^{(B)}-\nabla_{\boldsymbol{Q}_k}^{(B)}\|_F\leq\sqrt{n}\|(\nabla_{\boldsymbol{Q}_{k+1}}^{(A)})^{\mathrm{T}}\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})-(\nabla_{\boldsymbol{Q}_k}^{(A)})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F$$

Adding and subtracting terms in the norm and applying triangle inequality:

$$\begin{aligned}
&\sqrt{n}\|(\nabla_{\boldsymbol{Q}_{k+1}}^{(A)})^{\mathrm{T}}\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})-(\nabla_{\boldsymbol{Q}_k}^{(A)})^{\mathrm{T}}\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})\\
&\quad+(\nabla_{\boldsymbol{Q}_k}^{(A)})^{\mathrm{T}}\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})-(\nabla_{\boldsymbol{Q}_k}^{(A)})^{\mathrm{T}}\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F\\
&\quad\leq\sqrt{n}\|\nabla_{\boldsymbol{Q}_{k+1}}^{(A)}-\nabla_{\boldsymbol{Q}_k}^{(A)}\|_F\|\boldsymbol{Q}_{k+1}\|_F\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})\|_F\\
&\qquad+\sqrt{n}\|\nabla_{\boldsymbol{Q}_k}^{(A)}\|_F\|(\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})-\boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F
\end{aligned}$$

Invoking the lower-bound on the marginal, continuing from the above display,

$$\leq \frac{\sqrt{n}}{\delta}\|\nabla^{(A)}_{\boldsymbol{Q}_{k+1}} - \nabla^{(A)}_{\boldsymbol{Q}_k}\|_F + \sqrt{n}\|\nabla^{(A)}_{\boldsymbol{Q}_k}\|_F\|(\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) - \boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F$$

Let us consider $\|\nabla^{(A)}_{\boldsymbol{Q}_k}\|_F \leq \|\boldsymbol{X}_k\|_F\|\boldsymbol{C}\|_F\|\boldsymbol{R}_k\|_F \leq \|\boldsymbol{X}_k\|_F\|\boldsymbol{C}\|_F$. We have that:

$$\|\boldsymbol{X}_k\|_F^2 = \sum_{i,j}\frac{1}{\boldsymbol{g}^2_{(\boldsymbol{Q}_k)_i}}\boldsymbol{T}^2_{ij}\frac{1}{\boldsymbol{g}^2_{(\boldsymbol{R}_k)_j}}$$

Where we always have that $\boldsymbol{T}_{ij} \leq \boldsymbol{g}_{\boldsymbol{Q}i}$ and $\boldsymbol{T}_{ij} \leq \boldsymbol{g}_{\boldsymbol{R}j}$ by definition of $\boldsymbol{T}$ as a coupling. As such:

$$\leq \sum_{ij}\frac{1}{\boldsymbol{g}^2_{\boldsymbol{Q}i}}\left(\boldsymbol{g}_{\boldsymbol{Q}i}\boldsymbol{g}_{\boldsymbol{R}j}\right)\frac{1}{\boldsymbol{g}^2_{\boldsymbol{R}j}} = \sum_{ij}\frac{1}{\boldsymbol{g}_{\boldsymbol{Q}i}}\frac{1}{\boldsymbol{g}_{\boldsymbol{R}j}} = \left\langle \boldsymbol{g}_{\boldsymbol{Q}}^{-1}\boldsymbol{g}_{\boldsymbol{R}}^{-T}, \mathbf{1}_m\mathbf{1}_r^{\mathrm{T}}\right\rangle_F \leq \frac{mr}{\delta^2}$$

Thus $\|\nabla^{(A)}_{\boldsymbol{Q}_k}\|_F \leq \frac{\sqrt{mr}}{\delta}\|\boldsymbol{C}\|_F$, and the bound above reduces to

$$\leq \frac{\sqrt{n}}{\delta}\|\nabla^{(A)}_{\boldsymbol{Q}_{k+1}} - \nabla^{(A)}_{\boldsymbol{Q}_k}\|_F + \frac{\sqrt{nmr}}{\delta}\|\boldsymbol{C}\|_F\|(\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) - \boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F$$

Further bounding the last term in the norm

$$\begin{aligned}
&\frac{\sqrt{nmr}}{\delta}\|\boldsymbol{C}\|_F\|(\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) - \boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) \\
&\qquad\qquad + \boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) - \boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F \\
&\leq \frac{\sqrt{nmr}}{\delta}\|\boldsymbol{C}\|_F(\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})\|_F\|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k\|_F \\
&\qquad\qquad + \|\boldsymbol{Q}_k\|_F\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) - \mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F) \\
&\leq \frac{\sqrt{nmr}}{\delta}\|\boldsymbol{C}\|_F\left(\frac{1}{\delta} + \frac{\sqrt{r}}{\delta^2}\right)\|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k\|_F
\end{aligned}$$

Thus, the final bound on the $\nabla^{(B)}_{\boldsymbol{Q}}$ term is:

$$\begin{aligned}
&\leq \frac{\sqrt{n}}{\delta}\|\nabla^{(A)}_{\boldsymbol{Q}_{k+1}} - \nabla^{(A)}_{\boldsymbol{Q}_k}\|_F + \frac{\sqrt{nmr}}{\delta}\|\boldsymbol{C}\|_F\|(\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}}) - \boldsymbol{Q}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_k})\|_F \\
&\leq \left(\frac{\|\boldsymbol{C}\|_F\sqrt{nr}}{\delta^4} + \frac{\sqrt{nmr}}{\delta^2}\|\boldsymbol{C}\|_F\left(1 + \frac{\sqrt{r}}{\delta}\right)\right)\|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k\|_F
\end{aligned}$$

The total component-wise smoothness bound on $\boldsymbol{Q}$ is then

$$\begin{aligned}
&\|\nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1}, \boldsymbol{R}_k, \boldsymbol{T}_k) - \nabla_{\boldsymbol{Q}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_k, \boldsymbol{R}_k, \boldsymbol{T}_k)\|_F \\
&\leq \frac{\|\boldsymbol{C}\|_F}{\delta^2}\left(\left(\frac{\sqrt{nr}}{\delta^2} + \sqrt{nmr}\left(1 + \frac{\sqrt{r}}{\delta}\right)\right) + \frac{\sqrt{r}}{\delta}\right)\|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k\|_F
\end{aligned}$$

Identical reasoning applies for $\nabla_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}$, where we similarly have the gradient split into two terms:

$$\begin{aligned}
\nabla_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) &= \nabla^{(A)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}\nabla^{(B)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}} \\
&= \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X} - \mathbf{1}_m\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X})^{\mathrm{T}} \\
&= \nabla^{(A)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}} - \mathbf{1}_m\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\nabla^{(A)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}})^{\mathrm{T}}
\end{aligned}$$

As before, we may first show smoothness in $\nabla^{(A)}_{\boldsymbol{R}}$ using the same steps as for $\boldsymbol{Q}$

$$\begin{aligned}
&\|\nabla^{(A)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1},\boldsymbol{R}_{k+1},\boldsymbol{T}_k) - \nabla^{(A)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1},\boldsymbol{R}_k,\boldsymbol{T}_k)\|_F \\
&\quad= \|\boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})\boldsymbol{T}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_{k+1}}) - \boldsymbol{C}^{\mathrm{T}}\boldsymbol{Q}_{k+1}\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})\boldsymbol{T}_k\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_k})\|_F \\
&\quad\leq \|\boldsymbol{C}\|_F\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{Q}_{k+1}})\|_F\|\boldsymbol{Q}_{k+1}\|_F\|\boldsymbol{T}_k\|_F\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_{k+1}}) - \mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_k})\|_F \\
&\quad\leq \frac{\|\boldsymbol{C}\|_F}{\delta}\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_{k+1}}) - \mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_k})\|_F \\
&\quad\leq \frac{\|\boldsymbol{C}\|_F}{\delta^3}\|\boldsymbol{g}_{\boldsymbol{R}_{k+1}} - \boldsymbol{g}_{\boldsymbol{R}_k}\|_2 \\
&\quad\leq \frac{\|\boldsymbol{C}\|_F\sqrt{r}}{\delta^3}\|\boldsymbol{R}_{k+1} - \boldsymbol{R}_k\|_F.
\end{aligned}$$

For $\nabla^{(B)}_{\boldsymbol{R}}$, one may use the same reasoning as before to find:

$$\begin{aligned}
&\|\nabla^{(B)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1},\boldsymbol{R}_{k+1},\boldsymbol{T}_k) - \nabla^{(B)}_{\boldsymbol{R}}\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q}_{k+1},\boldsymbol{R}_k,\boldsymbol{T}_k)\|_F \\
\leq \|\mathbf{1}_m\left[\mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_{k+1}})\boldsymbol{R}^{\mathrm{T}}_{k+1}\nabla^{(A)}_{\boldsymbol{R}_{k+1}}\mathcal{L}_{\mathrm{LC}}) - \mathrm{diag}^{-1}(\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_k})\boldsymbol{R}^{\mathrm{T}}_k\nabla^{(A)}_{\boldsymbol{R}_k}\mathcal{L}_{\mathrm{LC}})\right]^{\mathrm{T}}\|_F \\
&\leq \sqrt{m}\|\mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_{k+1}})\boldsymbol{R}^{\mathrm{T}}_{k+1}\nabla^{(A)}_{\boldsymbol{R}_{k+1}}\mathcal{L}_{\mathrm{LC}} - \mathrm{diag}(1/\boldsymbol{g}_{\boldsymbol{R}_k})\boldsymbol{R}^{\mathrm{T}}_k\nabla^{(A)}_{\boldsymbol{R}_k}\mathcal{L}_{\mathrm{LC}}\|_F
\end{aligned}$$

As before, one may apply three rounds of triangle inequality inside the norm to bound this directly in terms of $\|\nabla^{(A)}_{\boldsymbol{R}_{k+1}} - \nabla^{(A)}_{\boldsymbol{R}_k}\|_F$, $\|\mathrm{diag}(1/\boldsymbol{R}_{k+1}) - \mathrm{diag}(1/\boldsymbol{R}_k)\|_F$, and $\|\boldsymbol{R}_{k+1} - \boldsymbol{R}_k\|_F$. Each of these terms is smooth in $\boldsymbol{R}$ by the lower-bound argument, so that smoothness in $\boldsymbol{R}$ holds analogously to $\boldsymbol{Q}$. The remainder of the proof for smoothness in $\boldsymbol{R}$ thus follows identically to that of $\boldsymbol{Q}$ above.

For $\boldsymbol{T}$, the component-wise bound of

$$\|\nabla_{\boldsymbol{T}_{k+1}}\mathcal{L}_{\mathrm{LC}} - \nabla_{\boldsymbol{T}_k}\mathcal{L}_{\mathrm{LC}}\|_F = \|\boldsymbol{Q}_{k+1}\boldsymbol{C}\boldsymbol{R}^{\mathrm{T}}_{k+1} - \boldsymbol{Q}_{k+1}\boldsymbol{C}\boldsymbol{R}^{\mathrm{T}}_{k+1}\|^2_F \leq L_T\|\boldsymbol{T}_{k+1} - \boldsymbol{T}_k\|_F$$

holds trivially for any $L_T > 0$ as the gradient is uniquely determined by $\boldsymbol{Q}$ and $\boldsymbol{R}$ alone. Thus there exist $L_{\boldsymbol{Q}}, L_{\boldsymbol{R}}, L_{\boldsymbol{T}} > 0$ as component-wise smoothness constants for $\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})$. □

We next prove Proposition 3.4, restated just below for convenience.

Proposition E.6 (Proposition 3.4)**.** *Consider the FRLC objective* (8)*. The FRLC algorithm, Algorithm 4, yields β-smooth iterates for $\beta = \mathrm{poly}(\|\boldsymbol{C}\|_F, m, r, \delta)$, where δ denotes the lower-bound on the entries of $\boldsymbol{g}_{\boldsymbol{Q}}, \boldsymbol{g}_{\boldsymbol{R}}$. Consider the convergence metric of 3.3 adapted from Ghadimi et al. (2014), given as:*

$$\Delta_k(\boldsymbol{x}_k,\boldsymbol{x}_{k+1}) = \sum_{i=1}^{p}\|\mathbf{g}_{\mathcal{X},k,i}\|^2 = \frac{1}{\gamma_k^2}\left[\|\boldsymbol{Q}_k - \boldsymbol{Q}_{k-1}\|^2_F + \|\boldsymbol{R}_k - \boldsymbol{R}_{k-1}\|^2_F + \|\boldsymbol{T}_k - \boldsymbol{T}_{k-1}\|^2_F\right]$$

for $\boldsymbol{x}_k = (\boldsymbol{Q}_k, \boldsymbol{T}_k, \boldsymbol{R}_k)$. Define the gap to the optimal solution D as in (36)*, and let $L = \sup_i(L_i)$ to be the global smoothness constant across all components. Then for $\gamma_k = \alpha/L$ as defined in 3.3 the FRLC algorithm has the non-asymptotic stationary convergence guarantee that:*

$$\min_{k\in 1,..,N-1}\Delta_k \leq \frac{2D^2L^2}{N\alpha^2}$$

Proof. The proof of the non-asymptotic stationary convergence of mirror descent of Ghadimi et al. (2014), adapted for coordinate mirror descent using the block-descent lemma in 3.3, only requires component-wise smoothness in $(\boldsymbol{Q},\boldsymbol{R},\boldsymbol{T})$. The proof of this for FRLC is given in **??**, and the guarantee follows directly for this value of $L = \max(L_{\boldsymbol{Q}}, L_{\boldsymbol{R}}, L_{\boldsymbol{T}}) = \mathrm{poly}(\|\boldsymbol{C}\|_F, m, r, \delta)$. □

Proposition E.7. ***Low-rank Approximation Error.*** *Let $\text{SR-W}^\star_r$ denote the optimal rank-r approximation for the semi-relaxed low-rank optimal transport problem, and let $\text{SR-W}^\star$ denote the optimal solution for the full-rank semi-relaxed optimal transport problem.*

Additionally, suppose $c_{\boldsymbol{b}} = \sum_{j=1}^m \boldsymbol{b}_j$ denotes the sum of the entries of the second marginal ($c_{\boldsymbol{b}} = 1$ if a probability measure). Then we have the following upper-bound on the objective error:

$$|\text{SR-W}^\star_r(\mu_{\boldsymbol{b}}) - \text{SR-W}^\star(\mu_{\boldsymbol{b}})| \leq c_{\boldsymbol{b}}\left(\max_{p,q}\{\boldsymbol{C}_{pq}\} - \min_{p,q}\{\boldsymbol{C}_{pq}\}\right)\ln\left(\min\{n,m\}/(r-1)\right)$$

We note that this bound also applies for the standard balanced optimal transport case, giving:

$$|\mathrm{W}_r^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}}) - \mathrm{W}^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}})| \leq \left(\max_{p,q}\{\boldsymbol{C}_{pq}\} - \min_{p,q}\{\boldsymbol{C}_{pq}\}\right) \ln\left(\min\{n, m\}/(r-1)\right)$$

and improves the previous bound of $|\mathrm{W}_r^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}}) - \mathrm{W}^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}})| \leq \|\boldsymbol{C}\|_\infty \ln\left(\min\{n, m\}/(r-1)\right)$ *as the distance matrix* $\boldsymbol{C}$ *contains only non-negative entries.*

Proof. We adapt the proof from Scetbon & Cuturi (2022) for the balanced case, which previously gave the bound:

$$|\mathrm{W}_r^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}}) - \mathrm{W}^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}})| \leq \|\boldsymbol{C}\|_\infty \ln\left(\min\{n, m\}/(r-1)\right)$$

In particular, for $z = \min\{m, n\}$, there exists an optimal $\mathrm{rank}_+(\boldsymbol{P}^*) \leq z$ where one may express the optimal solution for the non-negative coupling matrix $\boldsymbol{P}$ as a sum of z rank-one, non-negative outer products $\tilde{\boldsymbol{q}}_k \tilde{\boldsymbol{r}}_k^{\mathrm{T}} \succcurlyeq 0$:

$$\boldsymbol{P}^* = \sum_{k=1}^{z} \tilde{\boldsymbol{q}}_k \tilde{\boldsymbol{r}}_k^{\mathrm{T}} = \sum_{k=1}^{z} \lambda_k \boldsymbol{q}_k \boldsymbol{r}_k^{\mathrm{T}}$$

Where we write this sum in terms of normalized vectors $\boldsymbol{q}_k = \tilde{\boldsymbol{q}}_k / \|\tilde{\boldsymbol{q}}_k\|_1$, $\boldsymbol{r}_k = \tilde{\boldsymbol{r}}_k / \|\tilde{\boldsymbol{r}}_k\|_1$, and $\lambda_k = \|\tilde{\boldsymbol{r}}_k\|_1 \|\tilde{\boldsymbol{q}}_k\|_1$. Without any loss of generality, $(\lambda_k)_{k=1}^z$ is ordered in terms of decreasing value such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_z$. Note that we have a fixed constraint for the sum of the entries of $\boldsymbol{P}$ for the semi-relaxed case, assuming $\boldsymbol{b}$ is a general positive measure, as $\mathbf{1}_n^{\mathrm{T}} \boldsymbol{P} \mathbf{1}_m = \left(\boldsymbol{P}^{\mathrm{T}} \mathbf{1}_n\right)^{\mathrm{T}} \mathbf{1}_m = \boldsymbol{b}^{\mathrm{T}} \mathbf{1}_m = \sum_{j=1}^m \boldsymbol{b}_j := c_{\boldsymbol{b}}$ (where $c_{\boldsymbol{b}} = 1$ if $\boldsymbol{b}$ is chosen to be a probability measure, i.e. in the balanced case). Moreover, it is simple to observe for these ordered values that $\lambda_k \leq (c_{\boldsymbol{b}}/k)$. As in Scetbon & Cuturi (2022), define the weighted average of the bottom $z - r + 1$ vectors of the decomposition to be:

$$\boldsymbol{\alpha}_r = \frac{\sum_{i=r}^z \lambda_i \boldsymbol{q}_i}{\sum_{i=r}^z \lambda_i}$$
$$\boldsymbol{\beta}_r = \frac{\sum_{i=r}^z \lambda_i \boldsymbol{r}_i}{\sum_{i=r}^z \lambda_i}$$

And take the rank-r approximation using the optimal $r-1$ vectors of OPT and this weighted average of the bottom to be:

$$\tilde{\boldsymbol{P}}_r = \sum_{i=1}^{r-1} \lambda_i \boldsymbol{q}_i \boldsymbol{r}_i^{\mathrm{T}} + \left(\sum_{i=r}^{z} \lambda_i\right) \boldsymbol{\alpha}_r \boldsymbol{\beta}_r^{\mathrm{T}}$$

Where, by the assumption that $\boldsymbol{P}^* \in \Pi_{\boldsymbol{b}}$ is feasible:

$$\tilde{\boldsymbol{P}}_r^{\mathrm{T}} \mathbf{1}_n = \sum_{i=1}^{r-1} \lambda_i \boldsymbol{r}_i \boldsymbol{q}_i^{\mathrm{T}} \mathbf{1}_n + \left(\sum_{i=r}^{z} \lambda_i\right) \boldsymbol{\beta}_r \boldsymbol{\alpha}_r^{\mathrm{T}} \mathbf{1}_n = \sum_{i=1}^{r-1} \lambda_i \boldsymbol{r}_i + \left(\sum_{i=r}^{z} \lambda_i\right) \boldsymbol{\beta}_r = \sum_{i=1}^{z} \lambda_i \boldsymbol{r}_i = \boldsymbol{b}$$

Thus $\tilde{\boldsymbol{P}}_r \in \Pi_{\boldsymbol{b},r}$ is a feasible rank-r solution by feasibility of $\boldsymbol{P}^*$. One can verify that if $\boldsymbol{P}^* \in \Pi_{\boldsymbol{a},\boldsymbol{b}}$, then $\tilde{\boldsymbol{P}}_r \mathbf{1}_m = \boldsymbol{a}$ and the solution is again feasible. From this, we observe that the difference between this solution and $\boldsymbol{P}^*$ is an upper-bound to the difference between $\boldsymbol{P}^*$ and the optimal rank-r solution:

$$\begin{aligned}
|\text{SR-W}_r^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}}) - \text{SR-W}^\star(\mu_{\boldsymbol{a}}, \mu_{\boldsymbol{b}})| &= |\langle \boldsymbol{P}_r^*, \boldsymbol{C} \rangle_F - \langle \boldsymbol{P}^*, \boldsymbol{C} \rangle_F| \leq \langle \tilde{\boldsymbol{P}}_r, \boldsymbol{C} \rangle_F - \langle \boldsymbol{P}^*, \boldsymbol{C} \rangle_F \\
&= \left\langle \sum_{i=1}^{r-1} \lambda_i \boldsymbol{q}_i \boldsymbol{r}_i^{\mathrm{T}} + \left(\sum_{i=r}^{z} \lambda_i\right) \boldsymbol{\alpha}_r \boldsymbol{\beta}_r^{\mathrm{T}} - \sum_{i=1}^{z} \lambda_i \boldsymbol{q}_i \boldsymbol{r}_i^{\mathrm{T}}, \boldsymbol{C} \right\rangle_F \\
&= \left\langle \left(\sum_{i=r}^{z} \lambda_i\right) \boldsymbol{\alpha}_r \boldsymbol{\beta}_r^{\mathrm{T}} - \sum_{i=r}^{z} \lambda_i \boldsymbol{q}_i \boldsymbol{r}_i^{\mathrm{T}}, \boldsymbol{C} \right\rangle_F
\end{aligned}$$

Noting that $\boldsymbol{\alpha}_r, \boldsymbol{\beta}_r$ and $\boldsymbol{q}_i, \boldsymbol{r}_i$ are unit normalized positive vectors, the sum of the entries of the outer product $\mathbf{1}_n^{\mathrm{T}} \boldsymbol{q}_i \boldsymbol{r}_i^{\mathrm{T}} \mathbf{1}_m = 1$, and likewise for $\boldsymbol{\alpha}_r \boldsymbol{\beta}_r^{\mathrm{T}}$. Thus, continuing from the above display:

$$
\begin{aligned}
&= (\sum_{i=r}^{z} \lambda_i) \langle \boldsymbol{\alpha}_r \boldsymbol{\beta}_r^{\mathrm{T}}, \boldsymbol{C} \rangle_F - \sum_{i=r}^{z} \lambda_i \langle \boldsymbol{q}_i \boldsymbol{r}_i^{\mathrm{T}}, \boldsymbol{C} \rangle_F \\
&\leq (\sum_{i=r}^{z} \lambda_i) \langle \boldsymbol{\alpha}_r \boldsymbol{\beta}_r^{\mathrm{T}}, \boldsymbol{C} \rangle_F - (\sum_{i=r}^{z} \lambda_i) \min_{p,q} \{\boldsymbol{C}_{pq}\} \\
&\leq \left(\max_{p,q} \{\boldsymbol{C}_{pq}\} - \min_{p,q} \{\boldsymbol{C}_{pq}\} \right) \sum_{i=r}^{z} \lambda_i \\
&\leq c_{\boldsymbol{b}} \left(\max_{p,q} \{\boldsymbol{C}_{pq}\} - \min_{p,q} \{\boldsymbol{C}_{pq}\} \right) \sum_{i=r}^{z} \frac{1}{i} \\
&\leq c_{\boldsymbol{b}} \left(\max_{p,q} \{\boldsymbol{C}_{pq}\} - \min_{p,q} \{\boldsymbol{C}_{pq}\} \right) \ln\left(z/(r-1)\right)
\end{aligned}
$$

Concluding the proof. As discussed, this directly applies to the balanced case (for $c_{\boldsymbol{b}} = 1$). □

F Initialization

We propose a new initialization of the sub-couplings $\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}$ for the LC-factorization. Algorithm 6 generates a random full-rank initial condition in the set of couplings $\Pi_{\boldsymbol{a},\boldsymbol{b}}$ which still satisfies the marginal constraints. It accomplishes this by sampling random matrices which are full-rank and applying the Sinkhorn algorithm to each of them. Scetbon et al. (2021) proposed an initialization which represents an improvement over the rank-1 product measure which is rank-2. Follow-up work proposed initialization using k-means Scetbon & Cuturi (2022). However, this assumes the previous diagonal factorization and is thus not application for generating a latent coupling which may be non-diagonal, non-square, and with two distinct inner marginals. Our initialization is tailored to the LC-factorization, is effective, and has a full-rank guarantee. In particular, higher-rank initializations may exhibit better convergence properties by allowing the gradient to explore a larger set of directions immediately in the optimization. This initialization is given in Algorithm 6.

Algorithm 6 Initialize-Couplings

Input $\boldsymbol{a} \in \Delta_n, \boldsymbol{b} \in \Delta_m, \boldsymbol{g}_Q \in \Delta_r, \boldsymbol{g}_R \in \Delta_r$
$\boldsymbol{C_Q} \sim [0,1]^{n \times r}, \boldsymbol{C_R} \sim [0,1]^{m \times r}, \boldsymbol{C_T} \sim [0,1]^{r \times r}$
$\boldsymbol{K_Q} \leftarrow e^{\boldsymbol{C_Q}}, \boldsymbol{K_R} \leftarrow e^{\boldsymbol{C_R}}, \boldsymbol{K_T} \leftarrow e^{\boldsymbol{C_T}}$
$\boldsymbol{Q} \leftarrow$ *Sinkhorn*$(\boldsymbol{K_Q}, \boldsymbol{a}, \boldsymbol{g}_Q)$
$\boldsymbol{R} \leftarrow$ *Sinkhorn*$(\boldsymbol{K_R}, \boldsymbol{b}, \boldsymbol{g}_R)$
$\boldsymbol{T} \leftarrow$ *Sinkhorn*$(\boldsymbol{K_T}, \boldsymbol{g}_Q = \boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n, \boldsymbol{g}_R = \boldsymbol{R}^{\mathrm{T}} \mathbf{1}_m)$
Return $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$

Proposition F.1. *Suppose one samples an initial condition on the optimal transport coupling using Algorithm 6, where we assume* $\boldsymbol{C}_{ij} \sim \mathrm{Unif}(0,1)$ *such that* $\mathbb{P}(\mathrm{rank}(\boldsymbol{C}) < \min\{n,m\}) = 0$. *Additionally, suppose that* $\boldsymbol{a}, \boldsymbol{b} > \mathbf{0}$ *holds elementwise for both marginals* $\boldsymbol{a} \in \Delta_n, \boldsymbol{b} \in \Delta_r$. *Then the elementwise exponential* $\exp\{\boldsymbol{C}\}$ *(or* $\exp\{-\boldsymbol{C}\}$*) has full-rank and the return Sinkhorn*$(e^{-\boldsymbol{C}}, \boldsymbol{a}, \boldsymbol{b})$ *has full-rank.*

Proof. It is established that a random matrix $\boldsymbol{C} \sim [0,1]^{n \times m}$ has full-rank with probability one. For $\boldsymbol{K} = \exp\{\boldsymbol{C}\}$, it holds that the matrix must be entry-wise positive with $\boldsymbol{K}_{ij} \geq 0$. If columns $\boldsymbol{C}_{\cdot,i} \neq \boldsymbol{C}_{\cdot,j}$ then clearly $\boldsymbol{C}_{\cdot,i}^{\odot k} \neq \boldsymbol{C}_{\cdot,j}^{\odot k}$, and if $\boldsymbol{C}_{\cdot,i}, \boldsymbol{C}_{\cdot,j} \succcurlyeq \mathbf{0}$ and are independent remain so under element-wise powers. One may easily show this by contrapositive. Suppose there exist constants c_1, c_2 such that:

$$
c_1 \boldsymbol{C}_{\cdot,i}^{\odot k} + c_2 \boldsymbol{C}_{\cdot,j}^{\odot k} = 0
$$

As $\boldsymbol{C}_{.,i}, \boldsymbol{C}_{.,j} \succcurlyeq \mathbf{0}$, without loss of generality one may assume $c_1 > 0$ and $c_2 < 0$. Then:

$$c_1 \boldsymbol{C}_{.,i}^{\odot k} = c_1 \mathbf{1} \odot \boldsymbol{C}_{.,i}^{\odot k} = -c_2 \mathbf{1} \odot \boldsymbol{C}_{.,j}^{\odot k} = -c_2 \boldsymbol{C}_{.,j}^{\odot k} \implies \left(-\frac{c_1}{c_2}\right)^{1/k} \mathbf{1} = c\mathbf{1} = \frac{\boldsymbol{C}_{.,j}}{\boldsymbol{C}_{.,i}}$$

So clearly one has that $\boldsymbol{C}_{.,j} - c\boldsymbol{C}_{.,i} = \mathbf{0}$ for $c > 0$. This implies the columns $\boldsymbol{C}_{.,j}$ and $\boldsymbol{C}_{.,i}$ are dependent. Thus it is clear that elementwise powers of entrywise positive independent vectors preserve independence. The same principle extends trivially to exponentiation of the columns, where if one assumes by contradiction that $c_1 e^{\boldsymbol{C}_{.,i}} + c_2 e^{\boldsymbol{C}_{.,j}} = \mathbf{0}$ for $c_1 > 0$, $c_2 < 0$, one finds $\log\left(-\frac{c_1}{c_2}\right)\mathbf{1} = \boldsymbol{C}_{.,j} - \boldsymbol{C}_{.,i}$. Without loss of generality, assume $0 < -c_2 \leq c_1$ and $\boldsymbol{C}_{.,j} > \boldsymbol{C}_{.,i} > \mathbf{0}$, so that $\delta = \log\left(-\frac{c_1}{c_2}\right) \geq 0$. Then, considering constants q_1, q_2:

$$q_2 \boldsymbol{C}_{.,j} + q_1 \boldsymbol{C}_{.,i} = (q_1 + q_2)\boldsymbol{C}_{.,i} + \delta q_2 \mathbf{1} = \mathbf{0}$$

Assuming $\boldsymbol{C}_{.,i}$ has greater than one unique entry, which we assume as the entries are sampled densely in $\mathbb{R}$, the two vectors are dependent if and only if $q_1 = -q_2$ and $\delta = 0$, implying $\boldsymbol{C}_{.,i} = \boldsymbol{C}_{.,j}$. Thus, for the set of independent column vectors of $\boldsymbol{C}$, given as $\{\boldsymbol{C}_{.,i}\}_{i=1}^m$, the set $\{e^{\boldsymbol{C}_{.,i}}\}_{i=1}^m$, is also linearly independent. This holds analogously for the row vectors. As $\boldsymbol{C}$ is full-rank and $\operatorname{span}(\{\boldsymbol{C}_{.,i}\}) = \mathbb{R}^{\min\{m,n\}}$, we have that $\operatorname{span}(\boldsymbol{K}) = \mathbb{R}^{\min\{m,n\}}$ as $\boldsymbol{K} = e^{\boldsymbol{C}} = \sum_{k=0}^{\infty} \frac{\boldsymbol{C}^{\odot k}}{k!}$ (analogously $e^{-\boldsymbol{C}}$) and remains full-rank.

Sinkhorn expresses each variable as

$$\boldsymbol{X} = \operatorname{diag}(\boldsymbol{u})\boldsymbol{K}\operatorname{diag}(\boldsymbol{v})$$

where $\operatorname{rank}(\boldsymbol{X}) = \operatorname{rank}(\operatorname{diag}(\boldsymbol{u})\boldsymbol{K}\operatorname{diag}(\boldsymbol{v}))$. As Cuturi (2013b) updates the vectors $\boldsymbol{u} \leftarrow \boldsymbol{a}/\boldsymbol{K}\boldsymbol{v}$ and $\boldsymbol{v} \leftarrow \boldsymbol{b}/\boldsymbol{K}^{\mathrm{T}}\boldsymbol{u}$ from $\boldsymbol{u}_0 = \mathbf{1}_n$ and $\boldsymbol{v}_0 = \mathbf{1}_m$, if $\boldsymbol{a}, \boldsymbol{b} > \mathbf{0}$ holds element-wise, one has that $\boldsymbol{u}, \boldsymbol{v} > \mathbf{0}$ elementwise as well. Then, one has that $\operatorname{null}\operatorname{diag}(\boldsymbol{v}) = \{\mathbf{0}\}$ and $\operatorname{null}\operatorname{diag}(\boldsymbol{u}) = \{\mathbf{0}\}$, implying that $\operatorname{rank}(\operatorname{diag}(\boldsymbol{u})\boldsymbol{K}\operatorname{diag}(\boldsymbol{v})) = \operatorname{rank}(\boldsymbol{K}) = \min\{n, m\}$.

Thus, our initialization returns a random coupling matrix $\boldsymbol{X} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}$ of full-rank. □

In the next proposition, we show that one can *analytically* solve for the block-optimal weights $\boldsymbol{g}$ for the factorization of the coupling matrix $\boldsymbol{P}$ as $\boldsymbol{P} = \boldsymbol{Q}\operatorname{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}$ Forrow et al. (2019); Scetbon et al. (2021).

Proposition F.2. *For the minimization problem expressed as*

$$\min_{\boldsymbol{g} \in \Delta_r} \langle \boldsymbol{Q}\operatorname{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C}\rangle_F$$

One has the closed-form minimizer of $\boldsymbol{g}^*$ defined entrywise as:

$$\boldsymbol{g}_i^* = \frac{\sqrt{\boldsymbol{\omega_i}}}{\sum_{j=1}^r \sqrt{\boldsymbol{\omega_j}}}$$

For $\boldsymbol{\omega} = \operatorname{diag}^{-1}(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R})$, when $\boldsymbol{\omega} \geq \mathbf{0}$ holds entrywise.

Proof. As $\boldsymbol{\omega} \geq \mathbf{0}$, we consider the simplex condition $\sum_{j=1}^r \boldsymbol{g}_j = 1$ alone. Writing out the Lagrangian associated to our objective, with $\lambda \in \mathbb{R}$ our equality-condition dual variable, we have:

$$\mathcal{L}(\boldsymbol{g}, \lambda) = \langle \boldsymbol{Q}\operatorname{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C}\rangle_F + \lambda(1 - \sum_j \boldsymbol{g}_j)$$

Let us consider a rewriting of the inner product term:

$$\begin{aligned}
\langle \boldsymbol{Q}\operatorname{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}, \boldsymbol{C}\rangle_F &= \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{C}_{ij} \sum_{k=1}^r \boldsymbol{Q}_{ik} \left(\frac{1}{\boldsymbol{g}_k}\right) \boldsymbol{R}_{kj}^{\mathrm{T}} \\
&= \sum_{k=1}^r \left(\frac{1}{\boldsymbol{g}_k}\right) \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{Q}_{ki}^{\mathrm{T}} \boldsymbol{C}_{ij} \boldsymbol{R}_{jk} \\
&= \sum_{k=1}^r \left(\frac{1}{\boldsymbol{g}_k}\right) \left(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{R}\right)_{k,k} \\
&= \sum_{k=1}^r \frac{\boldsymbol{\omega}_k}{\boldsymbol{g}_k} = \boldsymbol{\omega}^{\mathrm{T}} (1/\boldsymbol{g})
\end{aligned}$$

Where division is interpreted element-wise in the last line. Thus, one can interpret the problem as minimizing the weighted sum of reciprocals of a density. As a result, we can simplify our Lagrangian's Froebenius inner product to a vector dot-product as:

$$\mathcal{L}(\boldsymbol{g}, \lambda) = \boldsymbol{\omega}^{\mathrm{T}} (1/\boldsymbol{g}) - \lambda(1 - \sum_{j=1}^{r} \boldsymbol{g}_j)$$

Thus, the first order condition tells us that the value of the coupling weight $\boldsymbol{g}_j$ is related to λ as:

$$\partial_{\boldsymbol{g}_j} \mathcal{L}(\boldsymbol{g}, \lambda) = -\frac{\boldsymbol{\omega}_j}{\boldsymbol{g}_j^2} + \lambda = 0 \implies \boldsymbol{g}_j = \sqrt{\frac{\boldsymbol{\omega}_j}{\lambda}}$$

And by relying on the summation condition on the probability density $\boldsymbol{g}$, yields the Langrange multiplier as

$$\sum_{j=1}^{r} \boldsymbol{g}_j = 1 = \sum_{j=1}^{r} \sqrt{\frac{\boldsymbol{\omega}_j}{\lambda}}$$

so that one finds

$$1 = \frac{1}{\sqrt{\lambda}} \sum_{j=1}^{r} \sqrt{\boldsymbol{\omega}_j} \implies \lambda = \left(\sum_{j=1}^{r} \sqrt{\boldsymbol{\omega}_j} \right)^2$$

Plugging our Lagrange-multiplier into the above expression yields:

$$\boldsymbol{g}_j = \sqrt{\frac{\boldsymbol{\omega}_j}{\lambda}} = \sqrt{\frac{\boldsymbol{\omega}_j}{\left(\sum_{i=1}^{r} \sqrt{\boldsymbol{\omega}_i} \right)^2}} = \frac{\sqrt{\boldsymbol{\omega}_j}}{\sum_{i=1}^{r} \sqrt{\boldsymbol{\omega}_i}}$$

As the Hessian $\nabla^2_{\boldsymbol{g}} \mathcal{L} = \mathrm{diag}(\frac{\boldsymbol{\omega}}{\boldsymbol{g}^3}) \succcurlyeq \mathbf{0}$, we conclude that this value of $\boldsymbol{g}$ indeed minimizes the loss over Δ_r. □

G Alternating updates on the dual variables

For the problem:

$$\inf_{(\boldsymbol{Q}, \boldsymbol{R}_k, \boldsymbol{g}_k) \in \mathcal{C}_1 \cap \mathcal{C}_2} \left(\frac{1}{\gamma_k} \mathrm{KL}(\boldsymbol{Q} \| \boldsymbol{K}_{\boldsymbol{Q}}) + \tau \mathrm{KL}(\boldsymbol{Q} \mathbf{1}_r \| \boldsymbol{a}) - \boldsymbol{\lambda}_1^{\mathrm{T}} \boldsymbol{Q}^{\mathrm{T}} \mathbf{1}_n \right) \tag{40}$$

one can find a simple set of semi-relaxed updates for the coupling matrix. We note the primal-dual relationship of Sinkhorn, $\boldsymbol{Q} = \mathrm{diag}(e^{\gamma_k \boldsymbol{f}_1}) \boldsymbol{K}_{\boldsymbol{Q}} \mathrm{diag}(e^{\gamma_k \boldsymbol{h}_1})$, and consider the entry-wise first-order condition required for the sub-coupling $\boldsymbol{Q}$:

$$\begin{aligned} 0 &= \gamma_k^{-1} \log \left(\frac{\boldsymbol{Q}_{ij}}{(\boldsymbol{K}_{\boldsymbol{Q}})_{ij}} \right) + \tau \log \left(\frac{\langle \boldsymbol{Q}_{i,.}, \mathbf{1}_r \rangle}{\boldsymbol{a}_i} \right) - \boldsymbol{\lambda}_{1,i} \\ \implies \log \boldsymbol{Q}_{ij} &= \tau \gamma_k \log \left(\frac{\boldsymbol{a}_i}{\langle \boldsymbol{Q}_{i,.}, \mathbf{1}_r \rangle} \right) + \log (\boldsymbol{K}_{\boldsymbol{Q}})_{ij} - \boldsymbol{\lambda}_{1j} \gamma_k \end{aligned} \tag{41}$$

Thus:

$$\boldsymbol{Q}_{ij} = \left(\frac{\boldsymbol{a}_i}{\boldsymbol{Q}_{i,.}^{\mathrm{T}} \mathbf{1}_r} \right)^{\tau \gamma_k} (\boldsymbol{K}_{\boldsymbol{Q}})_{ij} \, e^{-\boldsymbol{\lambda}_{1j} \gamma_k}$$

And in matrix-form, this yields:

$$\begin{aligned} \boldsymbol{Q} &= \mathrm{diag} \left(\frac{\boldsymbol{a}}{\boldsymbol{Q} \mathbf{1}_r} \right)^{\tau \gamma_k} \boldsymbol{K}_{\boldsymbol{Q}} \, \mathrm{diag}(e^{-\gamma_k \boldsymbol{\lambda}_1}) \\ &= \mathrm{diag}(e^{\gamma_k \boldsymbol{f}_1}) \boldsymbol{K}_{\boldsymbol{Q}} \, \mathrm{diag}(e^{\gamma_k \boldsymbol{h}_1}) \end{aligned}$$

And expanding the $\boldsymbol{Q}\mathbf{1}_r$ term explicitly, noting that $\boldsymbol{X}\,\mathrm{diag}(\boldsymbol{v})\mathbf{1} = \boldsymbol{X}\boldsymbol{v}$, we have:

$$\begin{aligned}
&\mathrm{diag}\left(\frac{\boldsymbol{a}}{\boldsymbol{Q}\mathbf{1}_r}\right)^{\tau\gamma_k} \boldsymbol{K}_{\boldsymbol{Q}}\,\mathrm{diag}(e^{-\gamma_k\boldsymbol{\lambda}_1}) \\
&\quad= \mathrm{diag}\left(\frac{\boldsymbol{a}}{\mathrm{diag}(e^{\gamma_k\boldsymbol{f}_1})\boldsymbol{K}_{\boldsymbol{Q}}\,\mathrm{diag}(e^{\gamma_k\boldsymbol{h}_1})\mathbf{1}_r}\right)^{\tau\gamma_k} \boldsymbol{K}_{\boldsymbol{Q}}\,\mathrm{diag}(e^{-\gamma_k\boldsymbol{\lambda}_1}) \\
&\quad= \mathrm{diag}\left(\frac{\boldsymbol{a}}{e^{\gamma_k\boldsymbol{f}_1}\odot \boldsymbol{K}_{\boldsymbol{Q}}e^{\gamma_k\boldsymbol{h}_1}}\right)^{\tau\gamma_k} \boldsymbol{K}_{\boldsymbol{Q}}\,\mathrm{diag}(e^{-\gamma_k\boldsymbol{\lambda}_1})
\end{aligned}$$

Thus, we identify $e^{\gamma_k\boldsymbol{h}_1} = e^{-\gamma_k\boldsymbol{\lambda}_1}$ as the right dual vector, and identify the following relationship in terms of the left dual vector:

$$\left(\frac{\boldsymbol{a}}{e^{\gamma_k\boldsymbol{f}_1}\odot \boldsymbol{K}_{\boldsymbol{Q}}e^{\gamma_k\boldsymbol{h}_1}}\right)^{\tau\gamma_k} = e^{\gamma_k\boldsymbol{f}_1} \implies e^{\gamma_k\boldsymbol{f}_1} = \left(\frac{\boldsymbol{a}}{\boldsymbol{K}_{\boldsymbol{Q}}e^{\gamma_k\boldsymbol{h}_1}}\right)^{\frac{\tau}{\tau+1/\gamma_k}}$$

From 40, the condition that $(\boldsymbol{Q}, \boldsymbol{R}_k, \boldsymbol{g}_k) \in \mathcal{C}_1 \cap \mathcal{C}_2$ implies that $\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_m = \boldsymbol{g}_k := \boldsymbol{g}$. As such, we find that:

$$\boldsymbol{Q}^{\mathrm{T}}\mathbf{1}_m = \mathrm{diag}(e^{\gamma_k\boldsymbol{h}_1})\boldsymbol{K}_{\boldsymbol{Q}}^{\mathrm{T}}\,\mathrm{diag}(e^{\gamma_k\boldsymbol{f}_1})\mathbf{1}_m = \mathrm{diag}(e^{\gamma_k\boldsymbol{h}_1})\boldsymbol{K}_{\boldsymbol{Q}}^{\mathrm{T}}e^{\gamma_k\boldsymbol{f}_1} = \boldsymbol{g}$$

Implying an update for $e^{\gamma_k\boldsymbol{h}_1}$ in the form:

$$e^{\gamma_k\boldsymbol{h}_1} = \left(\frac{\boldsymbol{g}}{\boldsymbol{K}_{\boldsymbol{Q}}^{\mathrm{T}}e^{\gamma_k\boldsymbol{f}_1}}\right)$$

Analogous reasoning applies for a relaxation of the other marginal, yielding the SR^{R}-projection and SR^{L}-projection (i.e. semi-relaxed OT).

Algorithm 7 SR^{L}-projection *(semi-relaxed OT, left marginal relaxed)*

Input $\boldsymbol{K}, \gamma, \tau, \boldsymbol{a}, \boldsymbol{b}, \delta$
$\boldsymbol{u} \leftarrow \mathbf{1}_n$
$\boldsymbol{v} \leftarrow \mathbf{1}_r$
repeat
  $\tilde{\boldsymbol{u}} \leftarrow \boldsymbol{u}$
  $\tilde{\boldsymbol{v}} \leftarrow \boldsymbol{v}$
  $\boldsymbol{u} \leftarrow (\boldsymbol{a}/\boldsymbol{K}\boldsymbol{v})^{\tau/(\tau+\gamma^{-1})}$
  $\boldsymbol{v} \leftarrow \left(\boldsymbol{b}/\boldsymbol{K}^{\mathrm{T}}\boldsymbol{u}\right)$
until $\gamma^{-1}\max\{\|\log\tilde{\boldsymbol{u}}/\boldsymbol{u}\|_\infty, \|\log\tilde{\boldsymbol{v}}/\boldsymbol{v}\|_\infty\} < \delta$
return $\mathrm{diag}(\boldsymbol{u})\boldsymbol{K}\,\mathrm{diag}(\boldsymbol{v})$

H Discussion of Complexity

For $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \in \mathbb{R}_+^{n\times r_1} \times \mathbb{R}_+^{m\times r_2} \times \mathbb{R}_+^{r_1\times r_2}$, the space complexity $O(nr_1 + r_1r_2 + mr_2)$ is linear if the ranks $r_1, r_2 = o(1)$ are taken to be small constants. The time-complexity of Algorithm 4 is $O(BLr^2(n+m))$ for B the number of inner Sinkhorn iterations, L the number of mirror-descent steps, n, m the number of samples in the first and second dataset, and $r = \max\{r_1, r_2, d\}$ for r_1, r_2 the ranks of the latent coupling and d the rank of the factorized distance matrix $\boldsymbol{C}$ (generally chosen to be a constant near r_1, r_2). Each matrix-multiplication is of max order $(n+m)r^2$, which happens a constant number of times in the computation of each gradient ∇_i, and for the respective Sinkhorn matrix-vector multiplications $\boldsymbol{K}\boldsymbol{v}$ and $\boldsymbol{K}^{\mathrm{T}}\boldsymbol{u}$. The L outer steps follow from the mirror-descent convergence rate and the number of iterations B required for each projection follow from the convergence of Sinkhorn. In particular, for ε a fixed error tolerance and η the entropy constant, one finds a $\pm\varepsilon D$ approximation for D the diameter of the data in $B = \mathrm{poly}(1/\eta\varepsilon)$ iterations using the Sinkhorn algorithm Charikar et al. (2023); Cuturi (2013a).

I Review of Background Material

I.1 Low-Rank Approximation of Pairwise Distance Matrices

As mentioned previously, works such as Charikar et al. (2023) have developed algorithms with linear $O((n+m)^{1+o(1)}\text{poly}(1/\epsilon))$ time-complexity and $O((n+m)d)$ space-complexity for sketching the optimal transport cost *value*. Recent works on low-rank factorization of the optimal transport coupling matrix $\boldsymbol{P}$ (the matrix associated to the coupling $\gamma \in \Pi(\mu,\nu)$) Scetbon & Cuturi (2022); Scetbon et al. (2023, 2021) have achieved per-iteration time-complexities of $O(T(n+m)dr)$ for some constant non-negative rank $r \geq 1$, d the dimension of the metric space, and T the number of iterations. By the JL-lemma one can simply embed the points in dimension $d = O(\log(nm)/\epsilon^2)$ while preserving pairwise distances, however, currently no proofs exist which offer the number of iterations T until convergence to some tolerance ε. This is partially due to how recent these works are, and also to the non-convexity of the objective which is sensitive to initial conditions Scetbon et al. (2021). However, the space complexity of the algorithm is $O((n+m)dr)$, which is noteworthy for being *linear* in the number of points and avoids storing the potentially intractable $O(nm)$ coupling matrix $\boldsymbol{P}$. To accomplish this, however, these works rely on a low-rank approximation of the pairwise distance matrix. A number of works by Indyk and Woodruff have concerned algorithms for finding low-rank approximations for such distance matrices. A seminal work Bakshi & Woodruff (2018) developed an algorithm which, given two point sets $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_j\}_{j=1}^m$ in some metric space $\mathcal{X}$, finds a rank r approximation in $O((n+m)^{1+\gamma}\text{poly}(r,1/\varepsilon))$ for $\gamma > 0$ an arbitrarily small constant and $\varepsilon > 0$ an error parameter. A more recent work Indyk et al. (2019) improves on this one, by reading a sample-optimal $O((n+m)r/\varepsilon)$ entries of the input matrix with a run-time which removes dependence on γ that is merely $O((n+m)\text{poly}(r,1/\varepsilon))$. This algorithm is used by all of the low-rank optimal transport works. These works, by finding a low-rank approximation to the coupling matrix $\boldsymbol{P} \approx \boldsymbol{A}\boldsymbol{B}^{\mathrm{T}} \in \mathbb{R}_+^{n\times m}$ due to space limitations on the coupling, necessarily cannot store the full distance matrix $\boldsymbol{C} \in \mathbb{R}_+^{n\times m}$ of the same size in memory either. As such, it must also be approximated as $\boldsymbol{C} \approx \boldsymbol{V}\boldsymbol{U}$ before input to the algorithm, where we necessarily require very effective approximations of $\boldsymbol{U}$ and $\boldsymbol{V}$ to tolerate the additional source of error from coarse-graining the distance matrix to be low-rank. As such, we present some of the details and algorithm of Indyk et al. (2019) as an essential component of the existing low-rank optimal transport solvers. We begin by summarizing the main theorems in Indyk et al. (2019), which provide an algorithm (upper-bound) on the low-rank distance-matrix approximation problem and a lower-bound on the number of entries which must be read.

Theorem I.1. *Indyk et al. (2019) There is a randomized algorithm that, given a distance matrix $\boldsymbol{C} \in \mathbb{R}^{n\times m}$, reads $O((n+m)r/\varepsilon)$ entries of $\boldsymbol{C}$, runs in time $\tilde{O}(n+m)\cdot\text{poly}(r,1/\varepsilon)$[1] and computes low-rank factors $\boldsymbol{V} \in \mathbb{R}^{n\times r}$, $\boldsymbol{U} \in \mathbb{R}^{r\times m}$ that with probability 0.99 satisfy:*

$$\|\boldsymbol{C} - \boldsymbol{V}\boldsymbol{U}\|_F^2 \leq \|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + \varepsilon\|\boldsymbol{C}\|_F^2 \tag{42}$$

For $\boldsymbol{C}_r$ the optimal rank-r approximation of $\boldsymbol{C}$.

This is a remarkable result, especially in light of the next theorem.

Theorem I.2. *Indyk et al. (2019) Let $r \leq m \leq n$ and $\varepsilon > 0$ such that $r/\varepsilon = O(\min\{m, n^{1/3}\})$. Any randomized and possibly adaptive algorithm that given a distance matrix $\boldsymbol{C} \in \mathbb{R}^{n\times m}$ computes $\boldsymbol{V} \in \mathbb{R}^{n\times r}$, $\boldsymbol{U} \in \mathbb{R}^{r\times m}$ satisfying $\|\boldsymbol{C} - \boldsymbol{V}\boldsymbol{U}\|_F^2 \leq \|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + \varepsilon\|\boldsymbol{C}\|_F^2$ must read $\Omega((n+m)r/\varepsilon)$ entries of $\boldsymbol{C}$ in expectation. This lower bound also holds for symmetric distance matrices $\boldsymbol{C} \in \mathcal{S}_n$.*

The lower-bound follows from a difficult argument which involves constructing a hard distribution over distance matrices, involving the use of random matrix theory. The upper-bound, however, follows relatively straightforwardly from a number of previous algorithms and their associated guarantees, along with the algorithm presented below. We introduce the algorithm and also offer a proof of I.1 for completeness.

[1] Where $\tilde{O}(\cdot)$ hides poly-log factors.

Algorithm 8 Low-Rank approximation for distance matrix $\boldsymbol{C}$

Input point sets $\{\boldsymbol{x}_i\}_{i=1}^n$, $\{\boldsymbol{y}_j\}_{j=1}^M$ in metric space $\mathcal{X}$ and metric d
Pick indices $i^* \in [n], j^* \in [m]$ uniformly at random
for $i = 1$ to n **do**
  Update sample probability $p_i = d(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + d(\boldsymbol{x}_{i^*}, \boldsymbol{y}_{j^*})^2 + \frac{1}{m}\sum_{j=1}^m d(\boldsymbol{x}_{i^*}, \boldsymbol{y}_j)^2$
end for
Sample $O(r/\varepsilon)$ rows $\boldsymbol{C}_{i,\cdot} \sim Categorical\left(\frac{p_i}{\sum_i p_i}\right)$
Compute $\boldsymbol{U}$ using Frieze et al. (2004)
Compute $\boldsymbol{V}$ using Chen & Price (2017)
return $\boldsymbol{V}, \boldsymbol{U}$

This algorithm relies on two previous works, whose main results we summarize here.

Theorem I.3. *Frieze et al. (2004) Let $\boldsymbol{C} \in \mathbb{R}^{n\times m}$. For a sample of $O(r/\varepsilon)$ rows according to a distribution $\boldsymbol{p} \in \Delta_n$ which satisfies $p_i \geq \Omega(1)\|\boldsymbol{C}_{i,\cdot}\|_2^2/\|\boldsymbol{C}\|_F^2$ for $i \in [n]$. Then in $O(mr/\epsilon + poly(r, 1/\varepsilon))$ time one may compute a matrix $\boldsymbol{U} \in \mathbb{R}^{r\times m}$ from this sample which satisfies:*

$$\|\boldsymbol{C} - \boldsymbol{C}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\|_F^2 \leq \|\boldsymbol{C} - \boldsymbol{C}_k\|_F^2 + \varepsilon\|\boldsymbol{C}\|_F^2$$

With probability 0.99.

Thus, to compute the first low-rank factor $\boldsymbol{U}$, we need to ensure the p_i generated from the algorithm satisfies this requirement and offer the (short) proof below.

Proof. First, it is helpful to note $d(x,y)^2 \leq (d(x,z) + d(z,y))^2 = d(x,z)^2 + 2d(x,z)d(y,z) + d(y,z)^2 \leq 2(d(x,z)^2 + d(y,z)^2)$. Where in the last step one uses AM-GM where $\prod_i a_i^{1/n} \leq \frac{\sum_i a_i}{n}$. Rewriting the norm of row i we have:

$$\begin{aligned}
\|\boldsymbol{C}_{i,\cdot}\|_2^2 &= \sum_{j=1}^m d(\boldsymbol{x}_i, \boldsymbol{y}_j)^2 \leq 2\sum_{j=1}^m d(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + d(\boldsymbol{y}_{j^*}, \boldsymbol{y}_j)^2 \\
&= 2md(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + 2\sum_{j=1}^m d(\boldsymbol{y}_{j^*}, \boldsymbol{y}_j)^2 \\
&\leq 2md(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + 4\sum_{j=1}^m d(\boldsymbol{y}_{j^*}, \boldsymbol{x}_{i^*})^2 + d(\boldsymbol{y}_j, \boldsymbol{x}_{i^*})^2 \\
&= 2md(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + 4md(\boldsymbol{y}_{j^*}, \boldsymbol{x}_{i^*})^2 + 4\sum_{j=1}^m d(\boldsymbol{y}_j, \boldsymbol{x}_{i^*})^2 \\
&= 4m\left(\frac{1}{2}d(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + d(\boldsymbol{y}_{j^*}, \boldsymbol{x}_{i^*})^2 + \frac{1}{m}\sum_{j=1}^m d(\boldsymbol{y}_j, \boldsymbol{x}_{i^*})^2\right) \leq 4mp_i
\end{aligned}$$

As we have the re-normalization $p_i \leftarrow \frac{p_i}{\sum_i p_i}$ before sampling, we need to consider the value of the expectation $\mathbb{E}[\sum_i p_i]$ to conclude.

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} p_i\right] &= \sum_{i=1}^{n} \mathbb{E}\left[d(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2 + d(\boldsymbol{x}_{i^*}, \boldsymbol{y}_{j^*})^2 + \frac{1}{m}\sum_{j=1}^{m} d(\boldsymbol{x}_{i^*}, \boldsymbol{y}_j)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_{j^*\sim[m]}\left[d(\boldsymbol{x}_i, \boldsymbol{y}_{j^*})^2\right] + \mathbb{E}_{i^*\sim[n], j^*\sim[m]}\left[d(\boldsymbol{x}_{i^*}, \boldsymbol{y}_{j^*})^2\right] \\
&\qquad + \frac{1}{m}\sum_{j=1}^{m} \mathbb{E}_{i^*\sim[n]}\left[d(\boldsymbol{x}_{i^*}, \boldsymbol{y}_j)^2\right] \\
&= \sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m} d(\boldsymbol{x}_i, \boldsymbol{y}_j) + n\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{nm} d(\boldsymbol{x}_i, \boldsymbol{y}_j) + \frac{n}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} d(\boldsymbol{x}_i, \boldsymbol{y}_j) \\
&= \frac{3}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} d(\boldsymbol{x}_i, \boldsymbol{y}_j) = \frac{3}{m}\|\boldsymbol{C}\|_F^2
\end{aligned}$$

By an application of Markov's inequality we have that:

$$P\left[\left(\sum_{i=1}^{n} p_i\right)^{-1} \geq \left(\frac{3\|\boldsymbol{C}\|_F^2}{m\delta}\right)^{-1}\right] = P\left[\sum_{i=1}^{n} p_i \leq \frac{3\|\boldsymbol{C}\|_F^2}{m\delta}\right] \geq 1 - \frac{\mathbb{E}[\sum_{i=1}^{n} p_i]}{\frac{3\|\boldsymbol{C}\|_F^2}{m\delta}} = 1 - \delta$$

Thus with probability $1-\delta$ we have:

$$\frac{p_i}{\sum_{i'=1}^{n} p_{i'}} \geq \frac{mp_i\delta}{3\|\boldsymbol{C}\|_F^2} = \frac{4mp_i\delta}{12\|\boldsymbol{C}\|_F^2} \geq \frac{\|\boldsymbol{C}_{i,.}\|_2^2\delta}{12\|\boldsymbol{C}\|_F^2} = \Omega(\delta)\frac{\|\boldsymbol{C}_{i,.}\|_2^2}{\|\boldsymbol{C}\|_F^2}$$

This indicates the algorithm presented has probabilities with an appropriate bound for using the algorithm of Frieze et al. (2004) to sample the $O(r/\varepsilon)$ rows of $\boldsymbol{C}$ and generate a rank-r factor $\boldsymbol{U}$. □

To conclude the result of Indyk et al. (2019) requires reference to an additional work which solves a regression problem for $\boldsymbol{V}$ given $\boldsymbol{C}$ and $\boldsymbol{U}$.

Theorem I.4. *Chen & Price (2017) There is a randomized algorithm $\mathcal{A}$, given matrices $\boldsymbol{C} \in \mathbb{R}^{n\times m}$, $\boldsymbol{U} \in \mathbb{R}^{r\times m}$ reads only $O(r/\varepsilon)$ columns of $\boldsymbol{C}$ with time-complexity $O(mr) + poly(r/\varepsilon)$ and returns $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ which satisfies*

$$\|\boldsymbol{C} - \boldsymbol{V}\boldsymbol{U}\|_F^2 \leq (1+\varepsilon)\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times r}}\|\boldsymbol{C} - \boldsymbol{Z}\boldsymbol{U}\|_F^2$$

with probability 0.99.

Thus, using the result of Chen and Price Chen & Price (2017), one may find a satisfying $\boldsymbol{V}$ easily for a fixed $\boldsymbol{U}, \boldsymbol{C}$. In particular, Indyk et al. (2019) concludes by tying together the low-rank distance matrix algorithm and the guarantees of the algorithms from Frieze et al. (2004) Chen & Price (2017) as follows

$$\begin{aligned}
\|\boldsymbol{C} - \boldsymbol{V}\boldsymbol{U}\|_F^2 &\leq (1+\varepsilon)\min_{\boldsymbol{Z}}\|\boldsymbol{C} - \boldsymbol{Z}\boldsymbol{U}\|_F^2 \\
&\leq (1+\varepsilon)\|\boldsymbol{C} - \boldsymbol{C}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\|_F^2 \\
&\leq (1+\varepsilon)(\|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + \varepsilon\|\boldsymbol{C}\|_F^2) \\
&= \|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + \varepsilon\|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + (1+\varepsilon)\varepsilon\|\boldsymbol{C}\|_F^2 \\
&\leq \|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + (1+2\varepsilon)\varepsilon\|\boldsymbol{C}\|_F^2 \leq \|\boldsymbol{C} - \boldsymbol{C}_r\|_F^2 + \tilde{\varepsilon}\|\boldsymbol{C}\|_F^2
\end{aligned}$$

Which achieves the bound up to a constant scaling of ε and shows the result of Indyk et al. (2019). We next investigate low-rank optimal transport solvers, which assume the result and algorithm of Indyk et al. (2019) to tractably scale to massive datasets.

J Connection of Optimal Transport to Projection Problems

Before discussing works which have address the problem of finding low-rank decompositions of the coupling matrix $\boldsymbol{P}$, we discuss a few relevant preliminaries. The Bregman-divergence of some function $F(\boldsymbol{x})$, defined by the first-order Taylor expansion of F:

$$D_F(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}) - F(\boldsymbol{y}) + \nabla F(\boldsymbol{y})^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{y})$$

For the negative entropy function $-\mathrm{H}(\boldsymbol{\xi})$, this corresponds to the KL-divergence $\mathrm{KL}(\boldsymbol{\zeta} \mid \boldsymbol{\xi})$. We introduce the iterative-Bregman projection algorithm in the context of the KL-divergence owing to the direct connection with entropically-regularized optimal transport.

Definition J.1. Iterative Bregman ProjectionsBregman (1967)

Suppose $\mathcal{C} = \cap_{l=1}^{L} \mathcal{C}_l$ is an intersection of closed convex sets $\{\mathcal{C}_l\}_{l=1}^{L}$. For $n > L$, let the indexing be L-periodic as $\mathcal{C}_n := \mathcal{C}_{n \mod L}$. Suppose we want to find a minimizer $\boldsymbol{\zeta}$ of the KL-divergence with some positive vector $\boldsymbol{\xi} \in \mathbb{R}_+^{n \times m}$ such that $\boldsymbol{\zeta} \in \mathcal{C}$. This is to say, we hope to solve the problem of minimizing a distance subject to the condition that one remains in this intersection of convex sets:

$$\min_{\boldsymbol{\zeta} \in \mathcal{C}} \mathrm{KL}(\boldsymbol{\zeta} \| \boldsymbol{\xi})$$

Where the projection of $\boldsymbol{\xi}$ onto the set $\mathcal{C}$ is denoted by

$$\boldsymbol{\zeta}^* = \arg\min_{\boldsymbol{\zeta} \in \mathcal{C}} \mathrm{KL}(\boldsymbol{\zeta} \| \boldsymbol{\xi}) := \mathcal{P}_{\mathcal{C}}^{KL}(\boldsymbol{\xi})$$

Supposing that each $\mathcal{C}_l$ forms a quotient space $\mathcal{C}_l = V/U$ for a subspace $U \subset V$, defined as $V/U = \{\boldsymbol{v} + U \mid \boldsymbol{v} \in V\}$[2], the iterative Bregman projection algorithm alternates projections onto each set $\mathcal{C}_n$ as

$$\boldsymbol{\zeta}^{(n)} \leftarrow \mathcal{P}_{C_n}^{KL}(\boldsymbol{\zeta}^{(n-1)}) \tag{43}$$

starting from $\boldsymbol{\zeta}^{(0)} = \boldsymbol{\xi}$.

One may show Bregman (1967) the convergence of Bregman projections to the unique minimizer in $\mathcal{C}$, $\boldsymbol{\zeta}^*$, where we have the guarantee that $\boldsymbol{\zeta}^{(n)} \to \boldsymbol{\zeta}^*$ as $n \to \infty$. These iterative projections only have convergence guarantees when the constraint sets are quotient spaces–this is clearly not the case for the constraints of the optimal transport LP, and a few more notions are required.

Definition J.2. Dykstra's AlgorithmDykstra (1983)

Given a point $\boldsymbol{x}_0 \in E$ for E a Euclidean space [3], to find the unique point in $\mathcal{C} = \cap_{l=1}^{L} \mathcal{C}_l$ for closed, convex sets $\mathcal{C}_l$ that minimize the distance to $\boldsymbol{x}_0$ as

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{C}} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$$

One may initialize the residuals $\boldsymbol{q}_{-(L-1)} = ... = \boldsymbol{q}_0 = 0$ and apply the algorithm:

$$\boldsymbol{x}_n = \mathcal{P}_{\mathcal{C}_n}(\boldsymbol{x}_{n-1} + \boldsymbol{q}_{n-1})$$
$$\boldsymbol{q}_n = (\boldsymbol{x}_{n-1} - \boldsymbol{x}_n) + \boldsymbol{q}_{n-L}$$

Where $\mathcal{C}_n := \mathcal{C}_{n \mod L}$ as before and $\mathcal{P}_{\mathcal{C}_n}$ denotes the projection operator onto the convex set $\mathcal{C}_n$.

One can note that Dykstra's algorithm for projections onto intersections of convex sets no longer relies on the assumption that the set is a quotient space, and applies in the case that it is merely closed under convex-combinations. To generalize Dykstra's for more general functions that than the ℓ_2-norm, one may define the projection with respect to the Bregman divergence of a cost function F and define the projection by the minimization: $\mathcal{P}_{\mathcal{C}_n} := \arg\min_{\boldsymbol{x}} D_F(\boldsymbol{x}, \boldsymbol{y})$. It was proven in Bauschke & Lewis (2000) that the generalized form of Dykstra's iterations are given as:

$$\boldsymbol{x}_n = \mathcal{P}_{\mathcal{C}_n}(\nabla F^*(\nabla F(\boldsymbol{x}_{n-1}) + \boldsymbol{q}_{n-1}))$$
$$\boldsymbol{q}_n = (\nabla F(\boldsymbol{x}_{n-1}) - \nabla F(\boldsymbol{x}_n)) + \boldsymbol{q}_{n-L}$$

For F^* denoting the Fenchel-conjugate of F, which we define later in connection to the low-rank dual problem. Notably, Bauschke & Lewis (2000) also provided guarantees of convergence to the optimal solution which extend to the case that F is the negative entropy and D_F the KL-divergence. These constitute Dykstra's algorithm with cyclic Bregman projections, and project a point to the closest point in the intersection of convex sets $\mathcal{C} = \cap_{l=1}^{L} \mathcal{C}_l$ for an arbitrary cost function F and its associated Bregman-divergence D_F.

[2] This is to say, $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}_l \implies c_1\boldsymbol{x} + c_2\boldsymbol{y} \in V/U$ for any $c_1, c_2 \in \mathbb{R}$.

[3] e.g. $\mathbb{R}, \mathbb{R}^d, \mathbb{R}^{n \times m}$, etc.

J.1 Dykstra's algorithm with cyclic Bregman projections

As such, we can see the updates for F being the negative entropy and D_F the KL-divergence. Without proof, the conjugate of the negative entropy is simply given as $F^*(\boldsymbol{\zeta}) = \exp\{\boldsymbol{\zeta} - \mathbf{1}\}$. Thus, for the minimization problem:

$$\boldsymbol{\zeta}^* = \arg\min_{\boldsymbol{\zeta} \in \mathcal{C}} \mathrm{KL}(\boldsymbol{\zeta} \| \boldsymbol{\xi}) := \mathcal{P}_{\mathcal{C}}^{\mathrm{KL}}(\boldsymbol{\xi})$$

Letting $\log \boldsymbol{q}_{-(L-1)} = ... = \log \boldsymbol{q}_0 = 0$, one may combine the two algorithms above to solve this minimization using generalized Dykstra's iterations. In particular, we have:

$$\begin{aligned}
\boldsymbol{\zeta}^{(n)} &= \mathcal{P}_{\mathcal{C}_n}\left(\nabla F^*(\nabla F(\boldsymbol{\zeta}^{(n-1)}) + \log \boldsymbol{q}_{n-1})\right) \\
&= \mathcal{P}_{\mathcal{C}_n}\left(\exp\left(\nabla F(\boldsymbol{\zeta}^{(n-1)}) + \log \boldsymbol{q}_{n-1} - \mathbf{1}\right)\right) \\
&= \mathcal{P}_{\mathcal{C}_n}\left(\exp\left(\log \boldsymbol{\zeta}^{(n-1)} + \mathbf{1} + \log \boldsymbol{q}_{n-1} - \mathbf{1}\right)\right) \\
&= \mathcal{P}_{\mathcal{C}_n}\left(\boldsymbol{\zeta}^{(n-1)} \odot \boldsymbol{q}_{n-1}\right)
\end{aligned}$$

And:

$$\begin{aligned}
\log \boldsymbol{q}_n &= (\nabla F(\boldsymbol{\zeta}^{(n-1)}) - \nabla F(\boldsymbol{\zeta}^{(n)})) + \log \boldsymbol{q}_{n-N} \\
&= \left(\log \boldsymbol{\zeta}^{(n-1)} + \mathbf{1} - (\log \boldsymbol{\zeta}^{(n)} + \mathbf{1}) + \log \boldsymbol{q}_{n-L}\right) \\
&= \log\left(\boldsymbol{q}_{n-L} \odot \frac{\boldsymbol{\zeta}^{(n-1)}}{\boldsymbol{\zeta}^{(n)}}\right)
\end{aligned}$$

So that:

$$\boldsymbol{\zeta}^{(n)} \leftarrow \mathcal{P}_{\mathcal{C}_n}\left(\boldsymbol{\zeta}^{(n-1)} \odot \boldsymbol{q}_{n-1}\right) \tag{44}$$

$$\boldsymbol{q}_n \leftarrow \boldsymbol{q}_{n-L} \odot \frac{\boldsymbol{\zeta}^{(n-1)}}{\boldsymbol{\zeta}^{(n)}} \tag{45}$$

Where division is interpreted to be element-wise, $\odot$ refers to the Hadamard product, and the logarithm is applied elementwise.

J.2 Connection to Sinkhorn distances

Interestingly, the Sinkhorn algorithm described in Algorithm 5 can be alternatively derived in the context of Bregman iterations Benamou et al. (2015) as a minimization of the form:

$$\mathrm{W}_\epsilon(\mu, \nu) = \epsilon \min_{\boldsymbol{P} \in \Pi(\mu,\nu)} \mathrm{KL}(\boldsymbol{P} \| \boldsymbol{\xi})$$

Where $\boldsymbol{\xi}$ is the kernel $\boldsymbol{\xi} = e^{-\boldsymbol{C}/\epsilon}$ and $\Pi(\mu, \nu) = \mathcal{C}_1 \cap \mathcal{C}_2$ for the convex constraint sets $\mathcal{C}_1 = \{\boldsymbol{P} : \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}\}$ and $\mathcal{C}_2 = \{\boldsymbol{P} : \boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n = \boldsymbol{b}\}$. To cast this into the Bregman-projection framework, one alternates between the two updates:

$$\begin{aligned}
\boldsymbol{P}^{(l)} &= \mathcal{P}_{\mathcal{C}_1}(\boldsymbol{P}^{(l-1)}) \\
\boldsymbol{P}^{(l+1)} &= \mathcal{P}_{\mathcal{C}_2}(\boldsymbol{P}^{(l)})
\end{aligned}$$

Where for the first projection one has the following first-order KKT condition:

$$\begin{aligned}
\nabla\left(\epsilon \mathrm{KL}(\boldsymbol{P} \| \boldsymbol{P}^{(l-1)}) + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{P}\mathbf{1}_m - \boldsymbol{a})\right) = \epsilon \log\left(\frac{\boldsymbol{P}}{\boldsymbol{P}^{(l-1)}}\right) + \boldsymbol{\lambda}\mathbf{1}_m^{\mathrm{T}} = \mathbf{0} \\
\implies \boldsymbol{P} = \mathrm{diag}(e^{-\boldsymbol{\lambda}/\epsilon})\boldsymbol{P}^{(l-1)}
\end{aligned}$$

With the constraint of $\mathcal{C}_1$ that $\boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}$, this implies $\mathrm{diag}(\boldsymbol{a}/\boldsymbol{P}^{(l-1)}\mathbf{1}_m) = \mathrm{diag}(e^{-\boldsymbol{\lambda}/\epsilon})$ and recovers the first update of Sinkhorn $\boldsymbol{P}^{(l)} \leftarrow \mathrm{diag}(\boldsymbol{a}/\boldsymbol{P}^{(l-1)}\mathbf{1}_m)\boldsymbol{P}^{(l-1)}$. An analogous argument gives the second, where all iterates satisfy $\boldsymbol{P}^{(l)} = \mathrm{diag}(\boldsymbol{u}^{(l)})e^{-\boldsymbol{C}/\epsilon}\,\mathrm{diag}(\boldsymbol{v}^{(l)})$ for $\boldsymbol{u}$, $\boldsymbol{v}$ as defined in Algorithm 5.

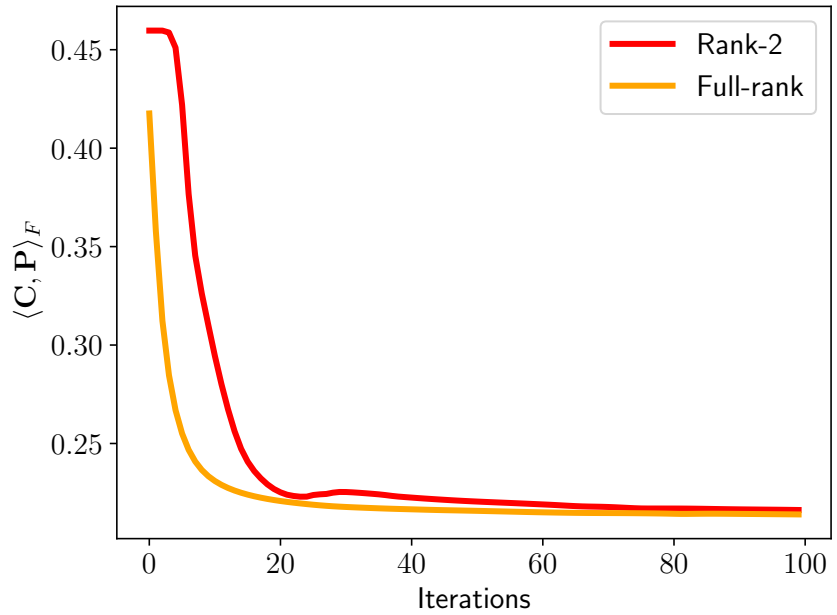


Figure 5: Transport cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ against number of iterations for FRLC with rank 200 on the synthetic dataset of two moons and eight Gaussians. Smooth convergence is observed for both rank-2 and full-rank random initialization.

K Additional Simulations

We tested on two additional synthetic datasets, both used as benchmarking datasets in Scetbon et al. (2021). We follow exactly the parameters provided in Scetbon et al. (2021) to simulate these datasets. For the first one, we simulated $n = m = 10,000$ points from two Gaussian mixtures in 2D (Fig. 6). The first Gaussian mixture is a mixture of three Gaussian distributions with means $(0, 0), (0, 1), (1, 1)$ respectively. The mixture proportion is $\frac{1}{3}$ and the covariance is 0.05 $\times$ identity for each Gaussian. The second Gaussian mixture is a mixture of two Gaussian distributions with means $(0.5, 0.5), (-0.5, 0.5)$ respectively. The mixture proportions is $\frac{1}{2}$ and the covariance is 0.05 $\times$ identity for each Gaussian.

For the second dataset, we simulated $n = m = 5,000$ points from two Gaussian mixtures in 10D. The first Gaussian mixture is a mixture of three Gaussian distributions with means $(0, 0, 0, \cdots, 0), (0, 1, 0, \cdots, 0), (1, 1, 0, \cdots, 0)$ respectively. The mixture proportions is $\frac{1}{3}$ and the covariance is 0.05 $\times$ identity for each Gaussian. The second Gaussian mixture is a mixture of two Gaussian distributions with means $(0.5, 0.5, 0, \cdots, 0), (-0.5, 0.5, 0, \cdots, 0)$ respectively. The mixture proportions is $\frac{1}{2}$ and the covariance is 0.05 $\times$ identity for each Gaussian.

For each dataset, we repeat the same procedure as in § 4.1, running FRLC and LOT with Euclidean distance as cost to find a low-rank coupling matrix between the two Gaussian mixtures with rank between 50 and 200. We observe the same pattern as Fig. 2b. FRLC obtains lower transport cost with increasing rank, and achieves lower cost for each rank than LOT under all initializations (Fig. 23c, Fig. 7).

L Graph Partitioning

L.1 Evaluation on a Graph Partitioning Task

We next evaluate FRLC on an unsupervised graph partitioning (node clustering) problem described by Chowdhury & Needham (2021). Specifically, given a graph $G = (V, E)$ of n nodes, we represent the graph as $G = (\boldsymbol{A}, \boldsymbol{h})$, where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ encodes the intra-graph node relationship (e.g. adjacency matrix) and $\boldsymbol{h} \in \Delta_n$ is a uniform measure. We cluster the nodes of G by estimating a GW coupling

Figure 6: Plot of the two simulated mixtures of Gaussians in 2D, following the same parameters as Scetbon et al. (2021).

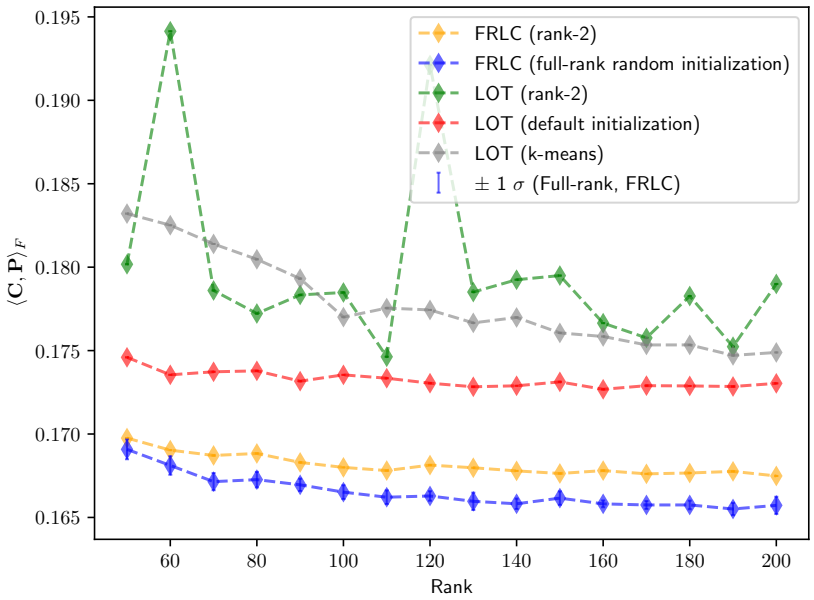


Figure 7: Transport cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ achieved by LOT Scetbon et al. (2021) and FRLC across different ranks and different initializations on the Wasserstein problem on the synthetic dataset of two mixtures of Gaussians in 2D.

Method	Dataset	Time in seconds (CPU)	Cost Value $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$	$\|\boldsymbol{P}\mathbf{1}_n - \boldsymbol{a}\|_2$	$\|\boldsymbol{P}^T\mathbf{1}_m - \boldsymbol{b}\|_2$
FRLC	2-moons,	**0.379**	**0.207**	1.56E-5	6.1E-18
LOT	8-Gaussians	0.751	0.210	1.90E-5	1.89E-5
FRLC	Gaussian mixture	**0.354**	**0.178**	1.05E-5	7.0E-18
LOT	(2D)	0.735	0.181	1.65E-5	1.70E-5
FRLC	Gaussian mixture	**0.323**	**0.294**	2.48E-6	8.9E-18
LOT	(10D)	0.677	0.307	1.39E-5	1.46E-5

Table 2: Runtime of FRLC and LOT on the synthetic datasets of 1000 samples, as well as cost value $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ and marginal tightness for context. This was done with the FRLC setting `max_inneriters_balanced`=1000, `max_inneriters_relaxed`=50, `min_iter`=7 and rank $r = 100$. This time excludes the extra time incurred for `ott-jax` problem setup and includes the setup time for FRLC.

between G and a smaller graph $\overline{G} = (\overline{\boldsymbol{B}}, \overline{\boldsymbol{h}})$, where $\overline{\boldsymbol{B}} \in \mathbb{R}^{m \times m}$ represents the relationship between each of the m clusters and $\overline{\boldsymbol{h}} \in \Delta_m$ is the proportion of nodes of G in each cluster. Without a priori knowledge, $\overline{\boldsymbol{h}}$ is set to be uniform and $\overline{\boldsymbol{B}}$ is set as the identity matrix. Vincent-Cuaz et al. (2022) notes that instead of solving a balanced GW problem, semi-relaxed GW with the right marginal $\overline{\boldsymbol{h}}$ relaxed learns $\overline{\boldsymbol{h}}$ from data and leads to more accurate node clustering.

Dataset		GWL	FRLC SR-GW	SpecGWL	SpecFRLC SR-GW
Wikipedia	sym, raw	0.314	0.387	0.372	**0.444**
(1998 nodes, 2700 edges)	sym, noisy	0.250	0.361	0.293	**0.400**
	asym, raw	0.263	0.276	0.194	**0.304**
	asym, noisy	**0.208**	0.201	0.141	0.177
EU-email	sym, raw	0.434	**0.464**	0.009	0.040
(1005 nodes, 25571 edges)	sym, noisy	0.392	**0.422**	0.009	0.014
	asym, raw	0.388	**0.398**	0.012	0.028
	asym, noisy	**0.385**	0.348	0.008	0.012
Amazon	raw	0.322	0.338	**0.505**	0.479
(1501 nodes, 4626 edges)	noisy	0.274	0.257	0.438	**0.453**
Village	raw	0.531	**0.710**	0.553	0.579
(1991 nodes, 8423 edges)	noisy	0.413	0.536	0.397	**0.829**

Table 3: Performance (measured using Adjusted Mutual Information (AMI)) in graph partitioning for *full-rank* OT algorithms GWL, SpecGWL and full-rank semi-relaxed FRLC. The top performing method for each dataset is highlighted in bold.

We benchmark FRLC for semi-relaxed GW on four real-world graphs: a Wikipedia hyperlink network with 15 webpage categories Yang & Leskovec (2012); an email interaction network within a European institute with 42 departments Yin et al. (2017); an Amazon product network with 12 product categories Yang & Leskovec (2012); and a network of interactions between 12 Indian villages Banerjee et al. (2013). We also test on the symmetric and noisy versions of each graph provided by Chowdhury & Needham (2021). We compare with two OT-based methods: (1) GWL Xu et al. (2019), which solves a balanced GW problem between G and $\overline{G}$ with the adjacency matrix of G as the intra-domain cost matrix $\boldsymbol{A}$; (2) SpecGWL Chowdhury & Needham (2021) which uses the heat kernel on the graph Laplacian as $\boldsymbol{A}$. We similarly run our FRLC algorithm using with both the adjacency matrix (denoted FRLC-SR-GW) and heat kernel (denoted SpecFRLC-SR-GW). Since the number of clusters in each dataset is not large, we compute the full-rank coupling matrix in each case.

Overall, FRLC achieves the best clustering performance on 9 out of 12 datasets (Table 3). When using the adjacency matrix, our semi-relaxed algorithm achieves better clustering result than GWL on 9 out of 12 datasets. When using the heat kernel, our semi-relaxed algorithm achieves better

clustering result than SpecGWL on 11 out of 12 datasets. These results show the importance of semi-relaxed OT on real-world problems, as well as the accuracy of FRLC.

L.2 Problem Statement

As discussed in Vincent-Cuaz et al. (2022); Chowdhury & Needham (2021), it is possible to achieve unsupervised graph partitioning (node clustering) through Gromov-Wasserstein (GW) OT. Given a graph (V, E) of n nodes, we encode it as $G = (\boldsymbol{A}, \boldsymbol{h})$, where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ encodes the intra-graph node relationship (e.g. adjacency matrix, Laplacian) and $\boldsymbol{h} \in \Delta n$ is a uniform measure over the nodes. If we want to cluster the nodes of G into m clusters, we can define a new graph $\overline{G} = (\overline{\boldsymbol{B}}, \overline{\boldsymbol{h}})$, where $\overline{\boldsymbol{B}} \in \mathbb{R}^{m \times m}$ is a diagonal matrix representing the cluster's connections and $\overline{\boldsymbol{h}}$ is a distribution over clusters estimating the proportion of nodes in G in each cluster. Since we don't know the density of clusters a priori, we can set $\overline{\boldsymbol{h}}$ to be uniform. Usually $\overline{\boldsymbol{B}}$ is set as the identity matrix.

To cluster the nodes in G, we can solve a GW problem matching nodes in G with nodes in $\overline{G}$, with intra-domain cost matrices $\boldsymbol{A}$ and $\overline{\boldsymbol{B}}$:

$$\begin{aligned} \min \quad & \sum_{ij'kl'} (\boldsymbol{A}_{ik} - \overline{\boldsymbol{B}}_{j'l'})^2 \boldsymbol{P}_{ij'} \boldsymbol{P}_{kl'} \\ s.t. \quad & \boldsymbol{P} \in \Pi_{\boldsymbol{h}, \overline{\boldsymbol{h}}} \end{aligned} \tag{46}$$

The cluster assignment of each node in G can then be recovered from $\boldsymbol{P}$ by finding the node in $\overline{G}$ mapped to it with the maximum weight.

However, since the proportion of nodes in G in each cluster is not known a priori, solving a balanced GW problem fixing the marginal of $\boldsymbol{P}$ on $\overline{G}$ to be a uniform $\overline{\boldsymbol{h}}$ significantly constrains the expressivity of the algorithm. Therefore, as proposed in Vincent-Cuaz et al. (2022), we can instead solve a semi-relaxed GW problem with the right marginal relaxed, fixing the marginal on G to be $\boldsymbol{h}$ but allowing the marginal on $\overline{G}$ to deviate from $\overline{\boldsymbol{h}}$:

$$\begin{aligned} \min \quad & \sum_{ij'kl'} (\boldsymbol{A}_{ik} - \overline{\boldsymbol{B}}_{j'l'})^2 \boldsymbol{P}_{ij'} \boldsymbol{P}_{kl'} + \tau \mathrm{KL}(\boldsymbol{P}^{\mathrm{T}} \mathbf{1}_n \mid \overline{\boldsymbol{h}}) \\ s.t. \quad & \boldsymbol{P} \in \Pi_{\boldsymbol{h}, \cdot} \end{aligned} \tag{47}$$

The learned $\overline{\boldsymbol{h}}$ from semi-relaxed GW estimates the posterior proportion of nodes in G in each of the m clusters.

L.3 Dataset and Preprocessing

We run our semi-relaxed FRLC algorithm on four real-world graph datasets: a Wikipedia hyperlink network with 1998 nodes and 15 clusters Yang & Leskovec (2012), a directed graph of email interactions in a European research institute with 1005 nodes and 42 clusters Yin et al. (2017), an Amazon product network with 1501 nodes and 12 clusters Yang & Leskovec (2012), and a network of interactions between Indian villages with 1991 nodes and 12 clusters Banerjee et al. (2013). The Wikipedia and EU-email graphs are directed, so we also use undirected versions of them. We also use noisy version of each graph by adding up to 10% additional edges following Chowdhury & Needham (2021), leading to a total of 12 different graphs to cluster.

L.4 Experiment Settings

We compare our algorithm with two baseline methods, GWL and SpecGWL. GWL Xu et al. (2019) solves the entropy-regularized version of the balanced GW problem of (46) with the adjacency matrix of G as $\boldsymbol{A}$. We set $\boldsymbol{h}$ such that the density of each node is proportional to its degree. We set $\overline{\boldsymbol{h}}$ to be a distribution estimated by sorting the weights of $\boldsymbol{h}$ and sampling m values via linear interpolation, following Chowdhury & Needham (2021). We set $\overline{\boldsymbol{B}} = diag(\overline{\boldsymbol{h}})$. We set the entropy regularization

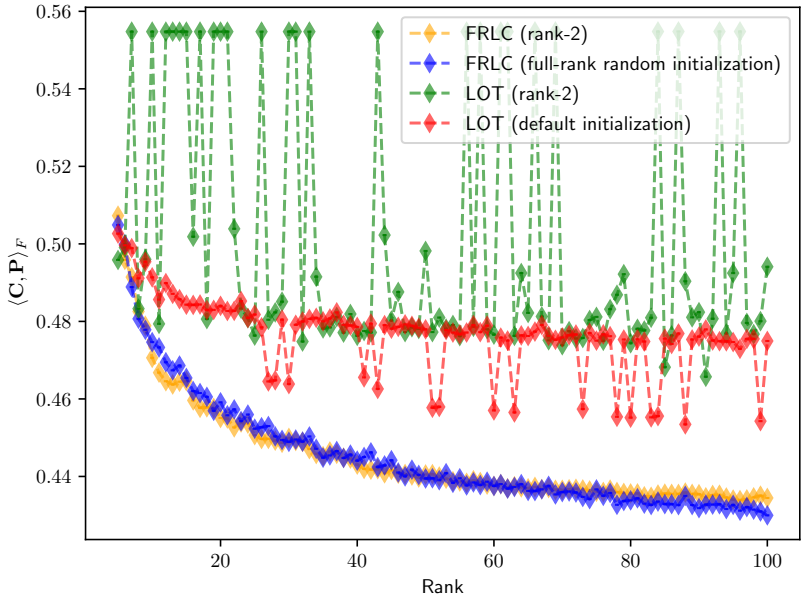


Figure 8: FRLC achieves lower primal cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ for $\boldsymbol{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}$ than Scetbon et al. (2021) on a spatial-transcriptomics dataset of mouse embryonic development Chen et al. (2022). FRLC demonstrates a more robust trend of improved performance with higher rank.

$\eta = 10^{-6}$. SpecGWL Chowdhury & Needham (2021) solves the same problem as GWL, but instead of using the adjacency matrix as $\boldsymbol{A}$, it uses the heat kernel on the normalized graph Laplacians Chung (2005). We set the heat parameter $t = 10$.

For our method, we solve the semi-relaxed GW problem of (47) with full rank solution. We use both adjacency matrix and heat kernel as $\boldsymbol{C}$ and label the result of the two representations as FRLC-SR-GW and SpecFRLC-SR-GW. We set $\tau = 0.01$ as to minimize the conformation to the right marginal. Since our method depends on random initialization, we run our method 10 times on each dataset and report the mean performance. We evaluate the resulting clusterings of all methods by computing the Adjusted Mutual Information (AMI) between the computed clustering and the ground truth clustering. This experiment, and the experiments on mouse embryo spatial transcriptomics, were conducted on cluster GPUs.

M Spatial Transcriptomics Alignment

M.1 Problem Statement

In this problem, we use FRLC to find a low-rank alignment matrix between cells from two spatial transcriptomics (ST) Ståhl et al. (2016) slices, collected at two timepoints, then use the computed alignment matrix for two downstream prediction tasks. An ST experiment on a 2D tissue slice yields a pair (X, Z). $X \in \mathbb{R}^{n \times p}$ is the gene expression matrix, where n is the number of cells on the slice and p is the number of genes measured. $X_{ij} \in \mathbb{R}$ is the gene expression level of gene j in cell i, where a higher number indicates stronger expression. $Z \in \mathbb{R}^{n \times 2}$ is the spatial coordinate matrix, where each row i stores the x-y coordinate of cell i on the slice. Therefore, each cell on the slice has a gene expression vector of length p, which encodes the feature of the cell, as well as a coordinate vector of length two, which encodes the geometrical information of the cell on the slice.

Our input data is a pair of ST slices $\mathcal{S}_0 = (X^0, Z^0), \mathcal{S}_1 = (X^1, Z^1)$, with n and m cells, of the same tissue region. We assume $\mathcal{S}_0$ is collected at timepoint $t = 0$ and $\mathcal{S}_1$ is collected at timepoint $t = 1$, hence the transition from $\mathcal{S}_0$ to $\mathcal{S}_1$ reflects the biological development of the tissue during the time period. We would like to find the ancestor-descendant relationship between cells from $\mathcal{S}_0$ and $\mathcal{S}_1$ by computing an optimal transport coupling matrix between cells from $\mathcal{S}_0$ and cells from $\mathcal{S}_1$. The state-of-the-art spatial transcriptomics alignment method moscot Klein et al. (2023) claims that

unbalanced OT is the most appropriate setup for this problem. Specifically, given discrete uniform measure $\boldsymbol{a}$ and $\boldsymbol{b}$ over cells from $\mathcal{S}_0$ and $\mathcal{S}_1$, we solve the unbalanced Wasserstein problem

$$\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \langle \boldsymbol{C}, \boldsymbol{P}\rangle + \tau \mathrm{KL}(\boldsymbol{P}\mathbf{1}_m \| \boldsymbol{a}) + \tau \mathrm{KL}(\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n \| \boldsymbol{b}) \tag{48}$$

$\boldsymbol{C} \in \mathbb{R}_+^{n\times m}$ has entries $\boldsymbol{C}_{ij} = c(X_i^0, X_j^1)$, where c is the Euclidean distance between the features of cell i from $\mathcal{S}_0$ and cell j from $\mathcal{S}_1$.

The above formulation only considers the feature information of the two slices, but not the geometrical information. Therefore, we can also solve the unbalanced GW problem

$$\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \sum_{i,j',k,l'} (\boldsymbol{A}_{ik} - \boldsymbol{B}_{j'l'})^2 \boldsymbol{P}_{ij'}\boldsymbol{P}_{kl'} + \tau \mathrm{KL}(\boldsymbol{P}\mathbf{1}_m \| \boldsymbol{a}) + \tau \mathrm{KL}(\boldsymbol{P}^{\mathrm{T}}\mathbf{1}_n \| \boldsymbol{b}) \tag{49}$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the Euclidean distance matrix between the spatial location of cells within $\mathcal{S}_0$, and $\boldsymbol{B} \in \mathbb{R}^{m\times m}$ is the Euclidean distance matrix between the spatial location of cells within $\mathcal{S}_1$. Similarly, we can combine the above two formulations to solve the unbalanced FGW problem.

Notice the inherent asymmetry stemming from the temporal nature of the problem: all cells on $\mathcal{S}_1$ should have an ancestor from $\mathcal{S}_1$, but not all cells from $\mathcal{S}_1$ need to have a descendent in $\mathcal{S}_0$ because of cell death. Therefore, the most natural OT task for this problem is semi-relaxed OT with the left marginal (the marginal on the first/ancestor slice) relaxed. Specifically, we can also solve the semi-relaxed Wasserstein problem

$$\min_{\boldsymbol{P}\in\Pi_{\cdot,\boldsymbol{b}}} \langle \boldsymbol{C}, \boldsymbol{P}\rangle + \tau \mathrm{KL}(\boldsymbol{P}\mathbf{1}_m \| \boldsymbol{a}) \tag{50}$$

as well as semi-relaxed GW and FGW problem.

Gene expression prediction task Given the alignment matrix $\boldsymbol{P}$ linking cells from $\mathcal{S}_0$ to $\mathcal{S}_1$, we can predict properties of cells in $\mathcal{S}_1$ from properties of cells in $\mathcal{S}_0$. Let the expression of a gene j in $\mathcal{S}_0$ be a vector $\boldsymbol{f}_j \in \mathbb{R}^n$, such that $\boldsymbol{f}_{ji}$ is the expression level of gene j in cell i, we can predict the expression of gene j in $\mathcal{S}_1$ as $\widetilde{\boldsymbol{f}}_j = m \times \boldsymbol{P}^{\mathrm{T}} \times \boldsymbol{f}_j \in \mathbb{R}^m$. The accuracy of the prediction can be measured by the Spearman correlation between the predicted expression and the ground truth expression $\overline{\boldsymbol{f}}_j$ of gene j in $\mathcal{S}_1$: $\rho(\widetilde{\boldsymbol{f}}_j, \overline{\boldsymbol{f}}_j)$. In this work, we test the prediction accuracy on 10 test genes: `Tubb2b`, `Pantr1`, `Actc1`, `Tnni1`, `Afp`, `Hbb-bh1`, `Fez1`, `Crabp1`, `Crabp2`, `Col3a1`, which are markers genes for various cell types in mouse embryo.

Cell type prediction task We can also use the cell type labels of cells in $\mathcal{S}_0$ to predict the cell type labels of cells in $\mathcal{S}_1$. Specifically, for each cell j in $\mathcal{S}_1$, we can assign it the type of the cell $\mathrm{argmax}_i \boldsymbol{P}_{ij}$ in $\mathcal{S}_0$. We can measure the accuracy of the cell type prediction by computing the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) between the predicted clustering of cells in $\mathcal{S}_1$ and the ground truth clustering.

M.2 Dataset and Preprocessing

We use the large-scale real-world dataset of mouse embryo development Chen et al. (2022), consisting of eight timepoints of ST slices during the whole process of mouse embryo development. In this work, we align the pair of adjacent timepoints of E11.5 and E12.5 embryos, consisting of 30,124 cells and 51,365 cells, respectively. We preprocess the dataset using the standard SCANPY Wolf et al. (2018) pipeline. We first filter the two slices to have the same set of genes, resulting in 26,436 genes for all cells from both slices. We then log-normalize the gene expression of all cells from the two slices, and apply Principle Component Analysis (PCA) to reduce the dimensionality of gene expressions to 30. We take the Euclidean distance between the gene expression in the PCA space as the cost matrix $\boldsymbol{C}$. We take the Euclidean distance between the 2D coordinate of each cell within each slice as the intra-domain cost matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

Fig. 9 visualizes the two slices in this dataset, with each cell annotated with a cell type from the original publication.

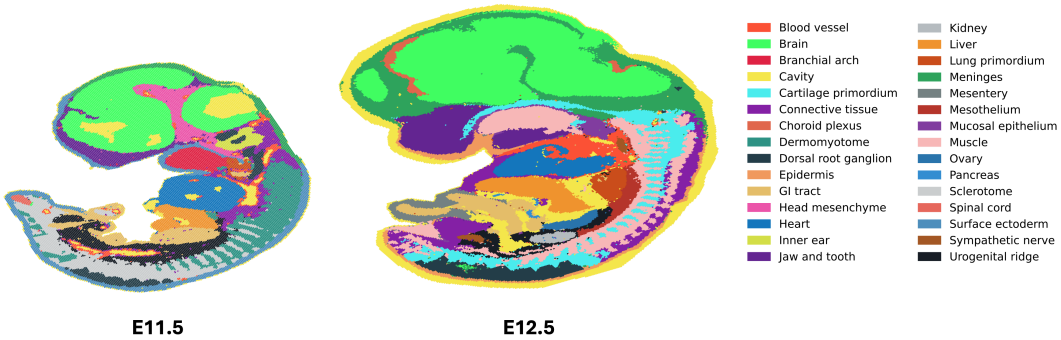


Figure 9: Visualization of the E11.5 and E12.5 mouse embryos, with each cell colored by the cell type annotated by Chen et al. (2022).

M.3 Experiment Settings

We compare with the unbalanced low-rank optimal transport algorithm of Scetbon et al. (2023), the backbone of the spatial transcriptomics alignment method moscot Klein et al. (2023), which was shown by Scetbon et al. (2023); Klein et al. (2023) to achieve state-of-the art performance on spatial transcriptomics alignment. We use Scetbon et al. (2023) to solve three unbalanced problems for spatial transcriptomics alignment: the unbalanced Wasserstein problem of (48) (result denoted as LOT-U-W), the unbalanced GW problem of (49) (LOT-U-GW), and the unbalanced FGW problem with a convex combination of the previous two costs (LOT-U-FGW). We use FRLC to solve the same three unbalanced problems (results denoted as FRLC-U-W, FRLC-U-GW, FRLC-U-FGW). Since the two slices contain $>$ 30,000 cells and $>$ 50,000 cells respectively, a full-rank solution is not feasible, hence we solve for low-rank solutions with the rank validated. We also solve semi-relaxed versions of Wasserstein (result denoted as FRLC-SR-W), GW (FRLC-SR-GW), FGW (FRLC-SR-FGW) problems using our FRLC algorithm, as well as using a particular setting of LOT-U that is equivalent to semi-relaxed solver (results denoted as LOT-SR-W, LOT-SR-GW, LOT-SR-FGW).

We perform extensive grid search to find the best hyperparameter combinations for each method and each problem. The grid of hyperparameters searched for each method is reported in Table. 6. The best performing hyperparameter combination for each method is reported in Table. 7 along with the performance on the validation genes. We pick the best hyperparameters using the Spearman correlation on the gene expression prediction task for 10 validation genes: `Ckb`, `Fabp7`, `Myl4`, `Tnnt2`, `Apoa2`, `Hba-x`, `Tubb3`, `Epha7`, `Ldha`, `Col1a2`, which are marker genes of various cell types as well. We report the performance of the alignment computed by each method using the Spearman correlation on the gene expression prediction task for 10 test genes, as well as the ARI and AMI on the cell type prediction task. Fig. 10 visualizes the ground truth cell type classification versus the classification predicted by our method FRLC-SR-W.

M.4 Runtime

We report the runtime and OT cost of FRLC and LOT Scetbon et al. (2021) on this dataset (mouse embryo E11.5-12.5) as well as two other datasets (E9.5-10.5, E10.5-11.5) from Chen et al. (2022) in Table 4. For all three datasets, FRLC achieves a better OT cost in a shorter time.

Dataset	LOT (seconds)	FRLC (seconds)	LOT (OT Cost)	FRLC (OT Cost)
Mouse embryo (E9.5–10.5)	2.545	1.112	0.440	0.385
Mouse embryo (E10.5–11.5)	4.209	1.190	0.371	0.344
Mouse embryo (E11.5–12.5)	8.667	1.889	0.478	0.439

Table 4: Comparison of methods on Stereo-Seq mouse embryo spatial transcriptomics datasets using GPU and default settings: `min_iter=10`, `max_iter=100`, rank $r = 50$.

N n^{th}-roots of unity

N.1 Problem Statement

To test if FRLC can effectively coarse-grain transport between two datasets with obvious cluster structure, we generate a pair of two-dimensional datasets $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, with ten and five clusters respectively. We run FRLC with $\boldsymbol{T} \in \mathbb{R}_+^{10\times 5}$ to see if the barycentric projections for the LC factorization (Definition 4.1) can recover the cluster structure. These projections induce 10 and 5 barycenters for the first and second dataset, and are defined by

$$\mathbf{Y}^{(1)} := \operatorname{diag}(1/\boldsymbol{g}_Q)\boldsymbol{Q}^{\mathrm{T}}\mathbf{Z}^{(1)}, \quad \mathbf{Y}^{(2)} := \operatorname{diag}(1/\boldsymbol{g}_R)\boldsymbol{R}^{\mathrm{T}}\mathbf{Z}^{(2)}.$$

We examine, visually, whether these barycenters are good representatives of the clusters in each dataset, and whether the latent coupling depicts a reasonable transfer of mass between barycenters. We also run FRLC with $\boldsymbol{T} \in \mathbb{R}_+^{10\times 10}$ and plot the barycenters from the resulting factorization for comparison. As discussed in § 4.2, the barycentric projections defined above, and in Definition 4.1 can be applied to factored couplings Forrow et al. (2019); Scetbon et al. (2021), yielding projections of the form:

$$\mathbf{Y}^{(1)} := \operatorname{diag}(1/\boldsymbol{g})\boldsymbol{Q}^{\mathrm{T}}\mathbf{Z}^{(1)}, \quad \mathbf{Y}^{(2)} := \operatorname{diag}(1/\boldsymbol{g})\boldsymbol{R}^{\mathrm{T}}\mathbf{Z}^{(2)}.$$

Thus, we also ran the method of Scetbon et al. (2021) on this data, called LOT throughout the experiment, with $\boldsymbol{g} \in (\mathbb{R}_+^*)^5$ and $\boldsymbol{g} \in (\mathbb{R}_+^*)^{10}$, plotting its barycenters in each case along with the diagonal latent coupling $\operatorname{diag}(\boldsymbol{g})$.

N.2 Dataset and Preprocessing

We instantiate a pair of two-dimensional datasets $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ as follows. Let $\mathcal{U}_n$ denote the n^{th}-roots of unity:

$$\mathcal{U}_n := \{e^{2\pi ik/n} : k = 0, \ldots, n\}.$$

These complex numbers can be equivalently expressed as ordered pairs on the unit circle:

$$\mathsf{c}_k = \begin{pmatrix} \operatorname{Re}(e^{2\pi ik/n}) \\ \operatorname{Im}(e^{2\pi ik/n}) \end{pmatrix} = \begin{pmatrix} \cos(2\pi k/n)) \\ \sin(2\pi k/n) \end{pmatrix}.$$

We consider n uniformly weighted mixtures of Gaussians, where for each sample X, we first sample a root of unity uniformly,

$$k \sim \operatorname{Uniform}(n),$$

and conditionally on k, we sample X from an isotropic Gaussian centered at this root of unity

$$X \sim \mathcal{N}(\mathsf{c}_k, \sigma^2 \operatorname{Id}_2),$$

using a standard deviation $\sigma = 0.1$. Samples are generated with the `make_blobs` function from `sklearn.datasets`. We generate two datasets in this way, $\mathbf{Z}^{(1)}$ for $n = 10$, and $\mathbf{Z}^{(2)}$ for $n = 5$. To generate dataset $\mathbf{Z}^{(1)}$, we first homogeneously scale the roots of unity to lie on a circle of radius 3 before sampling, so that the two datasets do not overlap but still use the same standard deviation. We do not scale the centers used for $\boldsymbol{Z}^{(2)}$, so they all lie on the unit circle. We generate $\mathbf{Z}^{(1)} = \{\mathbf{z}_1^{(1)}, \ldots, \mathbf{z}_{1000}^{(1)}\}$ and $\mathbf{Z}^{(2)} = \{\mathbf{z}_1^{(2)}, \ldots, \mathbf{z}_{1000}^{(2)}\}$ using 1000 samples each and form the empirical measures

$$\mu := \sum_{i=1}^{1000} \frac{1}{1000}\delta_{\mathbf{z}_i^{(1)}}, \quad \nu := \sum_{j=1}^{1000} \frac{1}{1000}\delta_{\mathbf{z}_j^{(2)}}.$$

supported on $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$, corresponding to uniform probability vectors $\boldsymbol{a}, \boldsymbol{b} \in \Delta_{1000}$.

N.3 Experiment Settings

We used default hyperparameter settings for both methods. In particular, FRLC sets $\tau = 75$ and $\gamma = 90$, with a maximum of 200 iterations and a minimum of 25 iterations, subject to the stopping criterion Δ given in (10). The LOT default settings are $\gamma = 10$ (there is no hyperparameter analogous to τ in LOT). Both methods use a random initialization and were run on CPU.

O Discussion of differences between FRLC and existing low-rank optimal transport algorithms

We provide an extensive comparison of FRLC against the parametrization and objective of Scetbon et al. (2021) in Appendix A. As Latent-OT of Lin et al. (2021) is an extension of the k-Wasserstein barycenter problem, it has a distinct objective and thus performs worse on primal OT cost (Table 5), making a direct experimental comparison on primal cost minimization only appropriate relative to the works of Scetbon et al. (2021, 2023, 2022). This stated, we still list a number of distinctions between FRLC and Latent-OT – noting that most differences between FRLC and Forrow et al. (2019) transfer from this discussion since Lin et al. (2021) extends the k-Barycenter problem of Forrow et al. (2019). (1) Lin et al. optimize two sets of variables: sub-couplings $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ and anchor points (z^x, z^y) on which the sub-couplings depend. FRLC only has $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ as variables of the optimization. (2) Cost matrices used in Lin et al. (2021) are built from distances between each dataset and its representative anchor points for $\boldsymbol{Q}, \boldsymbol{R}$, or the distances between the two sets of anchor points for $\boldsymbol{T}$. In contrast, ground costs used in FRLC to update $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ are always derived from the distance matrix $\boldsymbol{C}$ in the Wasserstein objective $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$. Specifically, the cost matrices used by Lin et al. are:

$$[\boldsymbol{C_Q}]_{ik} = \|x_i - z_k^x\|_2^2, \quad [\boldsymbol{C_R}]_{j\ell} = \|y_j - z_\ell^y\|_2^2, \quad [\boldsymbol{C_T}]_{k\ell} = \|z_k^x - z_\ell^y\|_2^2,$$

optionally using a Wasserstein distance for the entries of $\boldsymbol{C_T}$. (3) FRLC costs are given in the exponents of the Gibbs kernels written above and below Equation 9. These are derived directly from the rank-r Wasserstein problem $\min_{\boldsymbol{P} \in \Pi_r(a,b)} \langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$ and differ substantially from those of the proxy objective in Lin et al. (2021); Forrow et al. (2019). (4) The different objectives and variables lead to very different algorithms: Lin et al. alternate updates to the sub-couplings $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T})$ using Dykstra, with updates to the latent anchor points (z^x, z^y) using first-order conditions. In contrast, FRLC alternates semi-relaxed OT to update $(\boldsymbol{Q}, \boldsymbol{R})$ and balanced OT to update $\boldsymbol{T}$. (5) Because FRLC does not require anchor points to define costs, FRLC can handle cost matrices which are not simple functions of distance. For example, if $\boldsymbol{C}_{ij}$ is the price of transporting good i to warehouse j one may not be able to re-evaluate a price $c(x_i, z_k^x)$ between x_i and latent anchor z_k^x. In such situations, while finding a low-rank plan may make sense (e.g. to approximate an assignment for a massive dataset), an "anchor" may not have clear definition in the setting of general cost matrices. (6) The Lin et al. objective is only a proxy for a Wasserstein-type loss, and Lin et al. do not explore extensions to Gromov-Wasserstein (GW), or Fused GW, which FRLC readily generalizes to. A summary of the existing low-rank OT algorithms and key distinctions between them is given in Table 1.

For completeness, we offer a compare against the work Latent OT Lin et al. (2021), which solves a variation of the k-Wasserstein barycenter problem. As discussed, while their factorization is similar, their problem is distinct from FRLC as it does not solve the primal OT problem for general cost. We report the cost obtained by FRLC and by Latent OT on various simulated datasets in Table 5.

Dataset	**OT-cost (FRLC)**	**OT-cost (Lin et al.)**
5^{th} and 10^{th} roots of unity (rank $r_1, r_2 = 5, 10$)	**1.174**	2.124
Two-moons and 8-Gaussians (rank $r = 20$)	**2.716**	4.291
2D Gaussian mixture (rank $r = 20$)	**0.552**	0.922
10D Gaussian mixture (rank $r = 20$)	**1.038**	1.298

Table 5: Comparison against Lin et al. (2021) in primal OT-cost $\langle \boldsymbol{C}, \boldsymbol{P} \rangle_F$.

P Limitations

Our method introduces an additional hyperparameter τ relative to previous approaches Scetbon et al. (2021), controlling the strength of the KL penalty on the inner marginals when updating $\boldsymbol{Q}$ and $\boldsymbol{R}$.

Empirically, we found FRLC to be robust to different choices of τ, but applying the method optimally requires this additional hyperparamter in any grid-search.

We also note that the non-asymptotic criterion $\Delta(\cdot, \cdot)$ is weak relative to stronger notions of convergence, and that often users might prefer to simply run the method up to some number of maximal iterations by setting the parameter for whether $\Delta(\cdot, \cdot)$ is used to `False`. The W optimization empirically converges to minima smoothly, so for the most part there is not much of a need for Δ except for early stopping. We recommend that users plot the loss over iterations and use it to set the tolerance parameter `tol`, and the minimum and maximum iteration parameters for the time-being. The needs for these parameters might vary widely–the minimum number of iterations should be very low (around 5) for simple datasets and substantially higher for high-dimensional, structured ones.

Although we demonstrate strong performance already, there is massive room for improvement as our implementation is preliminary and not at the level of a high-performance library like `ott-jax`. We use lightweight vanilla implementations of Sinkhorn as a sub-routine, not taking advantage of the momentum-based techniques which could accelerate it massively. Thus, one can imagine that the potential scalability and speed of this method could be much higher than reported in this document.

Q Broader impacts

FRLC is general enough to be used modularly within any ML algorithm using OT as a subroutine to help with scalability. We also note that the LC factorization is similar to a PCA in the context of OT, yielding an optimal low-rank coupling with an interpretable latent coupling factor.

Hyperparameter	Values
rank (Both)	50, 100, 200
τ (Ours)	30, 50, 100
τ (LOT-U)	0.99, 0.9, 0.7
ϵ (LOT-U)	0.001, 0.01, 0.1

Table 6: Hyperparameter grid considered in hyperparameter search for validation. Scetbon et al. (2023); Klein et al. (2023) scales $\tau' = \frac{\tau}{1-\tau}$ and their τ' are in the same range.

Solver	Rank	τ (Ours)	τ (UL)	ϵ (UL)	Spearman ρ (Validation)
FRLC-SR-W (Ours)	200	30	-	-	0.465
FRLC-SR-GW (Ours)	100	100	-	-	0.288
FRLC-SR-FGW (Ours)	200	50	-	-	0.465
FRLC-U-W (Ours)	200	100	-	-	0.471
FRLC-U-GW (Ours)	50	30	-	-	0.282
FRLC-U-FGW (Ours)	200	30	-	-	0.353
LOT-U-W	200	-	0.9	0.001	0.394
LOT-U-GW	100	-	0.99	0.01	0.001
LOT-U-FGW	200	-	0.7	0.001	0.393
LOT-SR-W	200	-	0.7	0.001	0.394
LOT-SR-GW	200	-	0.7	0.1	0.003
LOT-SR-FGW	200	-	0.99	0.001	0.399

Table 7: The best performing hyperparameters for each solver and the performance on the validation genes.

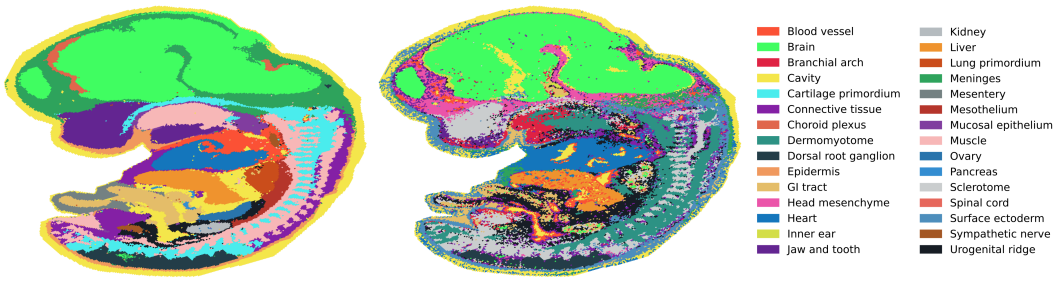


Figure 10: Ground truth and the predicted cell type classification of the E12.5 embryo.

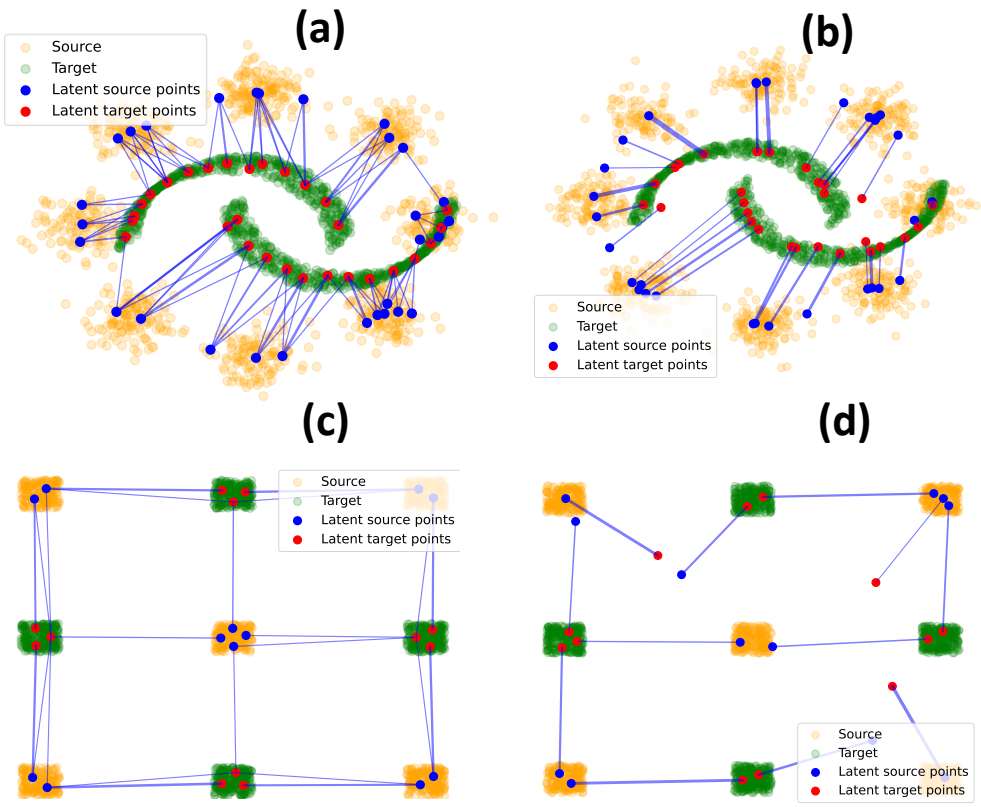


Figure 11: (a) LC-projection barycenters aligned with FRLC latent-coupling $\boldsymbol{T}$ on (a) two moons and eight Gaussians ($r = 30$), (b) LC-projection barycenters aligned with $\mathrm{diag}(\boldsymbol{g})$ from Scetbon et al. (2021) ($r = 30$), (c) the checkerboard dataset with FRLC latent coupling aligned barycenters ($r = 12$), and (d) with diagonal alignment Scetbon et al. (2021) ($r = 12$). We show in A.1 that the output of FRLC can be diagonalized to the factorization of Forrow et al. (2019) (Figure 12).

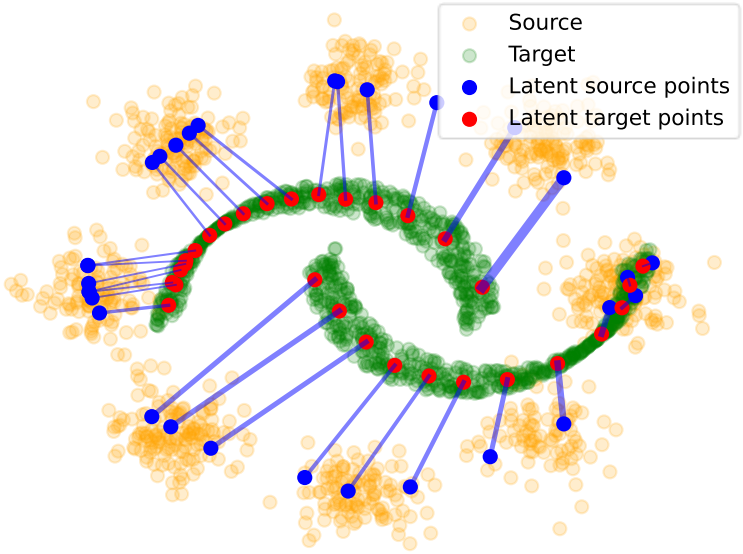


Figure 12: As discussed in A.1, one may recover the factorization of Forrow et al. (2019) as a sub-case of the LC-factorization. Shown is the factorization found by diagonalizing the output of FRLC from $(\boldsymbol{Q}, \boldsymbol{R}, \boldsymbol{T}) \mapsto (\boldsymbol{Q}', \boldsymbol{R}, \mathrm{diag}(\boldsymbol{g}_R))$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction exactly describe the background preceding this paper and placing it into context, and exactly describe the contribution and scope of the paper to the field.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We describe the limitations of our method in Section P. These include an additional hyperparameter that FRLC introduces relative to previous work, and the substantial room for code optimization.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results state their assumptions clearly and all propositions are followed by thorough line-by-line proofs with careful justifications made for each step. They have checked over by all of the authors. All proofs are provided in the supplement with detail and are referenced in the main body where appropriate.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all code used for generating the results of our paper. This not only includes the source code for the optimization, but the experimental code used for benchmarking. The algorithm is implemented exactly as described in the paper, and reviewers can freely consult the code we provide to them.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will upload all code, and all code required to generate the synthetic data experiments used. Any real data experiments using especially large-scale data have publicly available and easily accessible datasets which we provide references for. We provide comprehensive descriptions of how the data was pre-processed, and provide experimental code which can be followed easily.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We introduce all hyper-parameters with rigorous justification of their utility. The values of the default hyper-parameters are accessible in the code we upload and any experiment which does not use the default hyperparameter (e.g. for a validation search) has the full table provided and the code used for the experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The optimization is deterministic, so generally there's no need for reporting statistical significance. One small exception is that one of our proposed initializations is randomized, and for experiments which use it we do include $\pm 1\sigma$ error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We indicate when experiments are run on CPU versus GPU in the experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research was conducted ethnically and conforms to the guidelines of the code of ethics. We do not anticipate any major negative societal impact to result from this work on optimal transport.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We address broader impacts in Section Q: FRLC can be used for scaling an interpretability in any ML method using OT as a subroutine.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not generate any new datasets other than a toy dataset of Gaussians centered at n^{th} roots of unity, which has no societal impact. We do not believe our optimal transport work has significant risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All data that we use and all work that we build on are cited and credited heavily.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We provide no new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper involves no research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.