Unbalanced Low-rank Optimal Transport Solvers

Meyer Scetbon*
Microsoft Research
t-mscetbon@microsoft.com

Michael Klein*
Apple
michalk@apple.com

Giovanni Palla
Helmholtz Center Munich
giovanni.palla@helmholtz-muenchen.de

Marco Cuturi
Apple
cuturi@apple.com

Abstract

Two salient limitations have long hindered the relevance of optimal transport methods to machine learning. First, the $O(n^3)$ computational cost of standard sample-based solvers (when used on batches of n samples) is prohibitive. Second, the mass conservation constraint makes OT solvers too rigid in practice: because they must match *all* points from both measures, their output can be heavily influenced by outliers. A flurry of recent works has addressed these computational and modeling limitations, but has resulted in two separate strains of methods: While the computational outlook was much improved by entropic regularization, more recent $O(n)$ linear-time *low-rank* solvers hold the promise to scale up OT further. In terms of modeling flexibility, the rigidity of mass conservation has been eased for entropic regularized OT, thanks to unbalanced variants of OT that can penalize couplings whose marginals deviate from those specified by the source and target distributions. The goal of this paper is to merge these two strains, low-rank and unbalanced, to achieve the promise of solvers that are *both* scalable and versatile. We propose custom algorithms to implement these extensions for the linear OT problem and its fused-Gromov-Wasserstein generalization, and demonstrate their practical relevance to challenging spatial transcriptomics matching problems. These algorithms are implemented in the `ott-jax` toolbox [Cuturi et al., 2022].

1 Introduction

Recent machine learning (ML) works have witnessed a flurry of activity around optimal transport (OT) methods. The OT toolbox provides convenient, intuitive and versatile ways to quantify the difference between two probability measures, either to quantify a distance (the Wasserstein and Gromov-Wasserstein distances), or, in more elaborate scenarios, by computing a push-forward map that can transform one measure into the other [Peyré and Cuturi, 2019]. Recent examples include, e.g., single-cell omics [Bunne et al., 2021, 2022, Demetci et al., 2020, Nitzan et al., 2019, Cang et al., 2023, Klein et al., 2023], attention mechanisms [Tay et al., 2020, Sander et al., 2022], self-supervised learning[Caron et al., 2020, Oquab et al., 2023], and learning on graphs [Vincent-Cuaz et al., 2023].

On the challenges of using OT. Despite their long history in ML [Rubner et al., 2000], OT methods have long suffered from various limitations, that arise from their statistical, computational, and modelling aspects. The *statistical* argument is commonly referred to as the curse-of-dimensionality of OT estimators: the Wasserstein distance between two probability densities, and its associated optimal Monge map, is poorly approximated using samples as the dimension d of observation grows [Dudley et al., 1966, Boissard and Le Gouic, 2014]. On the *computational* side, computing OT between a pair of n samples involves solving a (generalized) matching problem, with a price of $O(n^3)$ and above [Kuhn, 1955, Ahuja et al., 1993]. Finally, the original *model* for OT rests on a

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

mass conservation constraint: all observations from either samples must be accounted for, including outliers that are prevalent in machine learning datasets. Combined, these weaknesses have long hindered the use of OT, until a more recent generation of solvers addressed these three crucial issues.

The Entropic Success Story. The winning approach, so far, to carry out that agenda has been entropic regularization methods [Cuturi, 2013]. The computational virtues of the Sinkhorn algorithm when solving OT [Altschuler et al., 2017, Peyré et al., 2016, Solomon et al., 2016, Le et al., 2021] come with statistical efficiency [Genevay et al., 2019, Mena and Niles-Weed, 2019, Chizat et al., 2020], and can also be seamlessly combined with *unbalanced* formulations by penalizing – rather than constraint – mass conservation, both for the linear [Frogner et al., 2015, Chizat et al., 2018, Séjourné et al., 2022, Fatras et al., 2021, Pham et al., 2020] and quadratic [Séjourné et al., 2021] problems. These developments have all been implemented in popular OT packages [Feydy et al., 2019, Flamary et al., 2021, Cuturi et al., 2022].

The Low-Rank Alternative. A recent strain of solvers relies instead on *low-rank* (LR) properties of cost and coupling matrices [Forrow et al., 2018, Scetbon and Cuturi, 2020, Scetbon et al., 2021]. Much like entropic solvers, these LR solvers have a better statistical outlook [Scetbon and Cuturi, 2022] and extend to GW problems [Scetbon et al., 2022]. In stark contrast to entropic solvers, however, LR solvers benefit from linear complexity $O(nrd)$ w.r.t sample size n (using rank r and cost dimension d) that can scale to ambitious tasks where entropic solvers fail [Klein et al., 2023].

The Need for Unbalanced Low-Rank Solvers. LR solvers do suffer, however, from a major practical limitation: their inability to handle unbalanced problems. Yet, unbalancedness is a crucial ingredient for OT to be practically relevant. This is exemplified by the fact that unbalancedness played a crucial role in the seminal reference [Schiebinger et al., 2019], where it is used to model cell birth and death.

Our Contributions We propose in this work to lift this last limitation for LR solvers to:

- Incorporate unbalanced regularizers to define a LR linear solver (§ 3.1);
- Provide accelerated algorithms, inspired by some of the recent corrections proposed by [Séjourné et al., 2022], to isolate translation terms that appear in dual subroutines (§ 3.2);
- Carry over and adapt these approaches to the GW (§ 3.3) and Fused-GW problems (§ 3.4);
- Carry out an exhaustive hyperparameter selection procedure within large scale OT tasks (spatial transcriptomics, brain imaging), and demonstrate the benefits of our approach (§ 4).

2 Reminders on Low-Rank Transport and Unbalanced Transport

We consider two metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$, as well as a cost function $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty[$. The simplex Δ_n^+ holds all positive n-vectors summing to 1. For $n, m \geq 1, a \in \Delta_n^+$, and $b \in \Delta_m^+$, given points $x_1, \dots, x_n \in \mathcal{X}$ and $y_1, \dots, y_m \in \mathcal{Y}$, we define two discrete probability measures μ and ν as $\mu := \sum_{i=1}^n a_i \delta_{x_i}$, $\nu := \sum_{j=1}^m b_j \delta_{y_j}$ where δ_z is the Dirac mass at z.

Cost matrices. For $q \geq 1$, consider first two square pairwise *cost* matrices, each encoding the geometries of points *within* μ and ν, and a rectangular matrix that studies that *across* their support:

$$A := [d_\mathcal{X}^q(x_i, x_{i'})]_{1\leq i,i' \leq n},\ B := [d_\mathcal{Y}^q(y_j, y_{j'})]_{1\leq j,j'\leq m}\ ,\ C := [c(x_i, y_j)]_{\substack{1\leq i\leq n,\\ 1\leq j\leq m}}\ .$$

The Kantorovich Formulation of OT is defined as the following parameterized linear program:

$$\mathrm{OT}(\mu, \nu) := \min_{P \in \Pi_{a,b}} \langle C, P\rangle\ , \quad \text{where} \quad \Pi_{a,b} := \left\{P \in \mathbb{R}_+^{n\times m},\ \text{s.t.}\ P\mathbf{1}_m = a,\ P^T\mathbf{1}_n = b\right\}\ . \quad (1)$$

The Low-Rank Formulation of OT is best understood as a variant of (1) that rests on a low-rank *property* for cost matrix C, and low-rank *constraints* for couplings P. More precisely, Scetbon et al. [2021] propose to constraint the set of admissible couplings to those, within $\Pi_{a,b}$, that have a non-negative rank of $r \geq 1$. That set can be equivalently reparameterized as

$$\Pi_{a,b}(r) = \{P \in \mathbb{R}_+^{n\times m} | P = Q\,\mathrm{diag}(1/g)R^T,\ \ Q \in \Pi_{a,g},\ \ R \in \Pi_{b,g},\ \ \text{and}\ \ g \in \Delta_r^+\}.$$

The low-rank optimal transport (LOT) problem simply uses that restriction in (1) to define :

$$\mathrm{LOT}_r(\mu, \nu) := \min_{P\in\Pi_{a,b}(r)} \langle C, P\rangle = \min_{Q\in\Pi_{a,g}, R\in\Pi_{a,g}, g\in\Delta_r^+} \langle C, Q\,\mathrm{diag}(1/g)R\rangle\ . \quad (2)$$

Scetbon et al. [2021] propose and prove the convergence of a mirror-descent scheme to solve (2), and obtain linear time and memory complexities with respect to the number of samples, where each iteration in that descent scales as $(n+m)rd$, where d is the rank of C.

The Unbalanced Formulation of OT starts from (1) as well, but proposes to do without $\Pi_{a,b}$ and its marginal constraints [Frogner et al., 2015, Chizat et al., 2018], and rely instead on two regularizers:

$$\mathrm{UOT}(\mu,\nu) := \min_{P\in\mathbb{R}_+^{n\times m}} \langle C, P\rangle + \tau_1 \mathrm{KL}(P\mathbf{1}_m|a) + \tau_2 \mathrm{KL}(P^T\mathbf{1}_n|b) \tag{3}$$

where $\tau_1, \tau_2 > 0$ and $\mathrm{KL}(p|q) := \sum_i p_i \log(p_i/q_i) + q_i - p_i$. This formulation is solved using entropic regularization, with modified Sinkhorn updates [Frogner et al., 2015]. *Proposing an efficient algorithm able to merge* (2) *with* (3) *is the first goal of this paper.*

Gromov-Wasserstein (GW) Considerations. The GW problem [Mémoli, 2011] is a generalization of (1) where the energy $\mathcal{Q}_{A,B}$ is a quadratic function of P defined through inner cost matrices A, B:

$$\mathcal{Q}_{A,B}(P) := \sum_{i,j,i',j'} (A_{ii'} - B_{jj'})^2 P_{ij} P_{i'j'} = \mathbf{1}_m^T P^T A^{\odot 2} P \mathbf{1}_m + \mathbf{1}_n^T P B^{\odot 2} P^T \mathbf{1}_n - 2\langle APB, P\rangle \tag{4}$$

where $\odot$ is the Hadamard product. To minimize (4), the default approach rests on entropic regularization [Solomon et al., 2016, Peyré et al., 2016] and variants [Sato et al., 2020, Blumberg et al., 2020, Xu et al., 2019, Li et al., 2023]. Scetbon et al. [2022] adapted the low-rank framework to minimize $\mathcal{Q}_{A,B}$ over low-rank matrices P, achieving a linear-time complexity when A and B are themselves low-rank. Independently, [Séjourné et al., 2021] proposed an unbalanced generalization that also applies to GW and which can be implemented practically using entropic regularization. Finally, the minimization of a composite objective involving the sum of $\mathcal{Q}_{A,B}$ with $\langle C, \cdot\rangle$ is known as the *fused* GW problem [Vayer et al., 2018].

3 Unbalanced Low-Rank Transport

3.1 Unbalanced Low-rank Linear Optimal Transport

We incorporate unbalancedness to low-rank solvers [Scetbon et al., 2021, 2022], moving gradually from the linear problem to the more involved GW and FGW problem. Using the framework of [Frogner et al., 2015, Chizat et al., 2018], we extend first the definition of LOT, introduced in (2), to the unbalanced case by considering the following optimization problem:

$$\mathrm{ULOT}_r(\mu,\nu) := \min_{P:\, \mathrm{rk}_+(P)\leq r} \langle C, P\rangle + \tau_1 \mathrm{KL}(P\mathbf{1}_m|a) + \tau_2 \mathrm{KL}(P^T\mathbf{1}_n|b), \tag{5}$$

where $\mathrm{rk}_+(P)$ denotes the non-negative rank of P. Therefore by denoting $\Pi_r := \{(Q,R,g) \in \mathbb{R}_+^{n\times r} \times \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^r \colon Q^T\mathbf{1}_n = R^T\mathbf{1}_m = g\}$, and using the reparameterization of low-rank couplings, we obtain the following equivalent formulation of ULOT:

$$\mathrm{ULOT}_r(\mu,\nu) = \min_{(Q,R,g)\in\Pi_r} \underbrace{\langle C, Q\,\mathrm{diag}(1/g)R^T\rangle}_{\mathcal{L}_C(Q,R,g)} + \underbrace{\tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2 \mathrm{KL}(R\mathbf{1}_r|b)}_{\mathcal{G}_{a,b}(Q,R,g)}\ . \tag{6}$$

We introduce the more compact notation $\mathcal{G}_{a,b}(Q,R,g) := F_{\tau_1,a}(Q\mathbf{1}_r) + F_{\tau_2,b}(R\mathbf{1}_r)$, where $F_{\tau,z}(s) := \tau\mathrm{KL}(s|z)$ for $\tau > 0$ and $z \geq 0$ coordinate-wise. To solve (6), and using this split, we move away from mirror-descent and apply instead proximal gradient-descent for the KL divergence. At each iteration, we consider a linear approximation of $\mathcal{L}_C$ where a KL penalization is added to the objective (as in the classical mirror descent scheme). However, we leave $\mathcal{G}_{a,b}$ intact at each iteration. Borrowing notations from [Scetbon et al., 2021], we must solve at each iteration the convex optimization problem:

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta}\in\Pi_r} \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k) + \tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2 \mathrm{KL}(R\mathbf{1}_r|b)\,, \tag{7}$$

where $(Q_0, R_0, g_0) \in \Pi_r$ is the initialization, and the triplet $\boldsymbol{\xi}_k := (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)})$ holds costs matrices that are updated at each iteration k:

$$\xi_k^{(1)} := Q_k \odot e^{-\gamma_k C R_k \,\mathrm{diag}(1/g_k)}, \xi_k^{(2)} := R_k \odot e^{-\gamma_k C^T Q_k \,\mathrm{diag}(1/g_k))}, \xi_k^{(3)} := g_k \odot e^{\gamma_k \omega_k / g_k^2},$$

with $[\omega_k]_i := [Q_k^T C R_k]_{i,i}$ for all $i \in \{1, \dots, r\}$, and $(\gamma_k)_{k \geq 0}$ is a sequence of positive step sizes.

Reformulation using Duality. To solve (7), we apply Dykstra's algorithm [1983], whose iterations correspond to an alternating maximization on the dual formulation of (7):

Proposition 1. *The convex optimization problem defined in (7) admits the following dual:*

$$\begin{aligned} \sup_{f_1, h_1, f_2, h_2} \mathcal{D}_k(f_1, h_1, f_2, h_2) := -F^\star_{\tau_1, a}(-f_1) - \frac{1}{\gamma_k} \langle e^{\gamma_k (f_1 \oplus h_1)} - 1, \xi_k^{(1)} \rangle \\ - F^\star_{\tau_2, b}(-f_2) - \frac{1}{\gamma_k} \langle e^{\gamma_k (f_2 \oplus h_2)} - 1, \xi_k^{(2)} \rangle - \frac{1}{\gamma_k} \langle e^{-\gamma_k (h_1 + h_2)} - 1, \xi_k^{(3)} \rangle \end{aligned} \tag{8}$$

where $h_1, h_2 \in \mathbb{R}^r$, $f_1 \in \mathbb{R}^n$, $f_2 \in \mathbb{R}^m$, $F^\star_{\tau,z}(\cdot) := \sup_y \{\langle y, \cdot \rangle - F_{\tau,z}(y)\}$ is the convex conjugate of $F_{\tau,z}$. In addition strong duality holds and the primal problem admits a unique minimizer.

Remark 1. *While we stick to KL regularizers in this work for simplicity, it is worth noting that this can be extended to more generic regularizers $F_{\tau_1,a}$ and $F_{\tau_2,b}$, as considered by Chizat et al. [2018].*

We use an alternating maximization scheme to solve (8). Starting from $h_1^{(0)} = h_2^{(0)} = \mathbf{0}_r$, we apply for $\ell \geq 0$ the following updates (dropping iteration number k in (7) for simplicity):

$$\begin{aligned} f_1^{(\ell+1)} := \arg\sup_z \mathcal{D}(z, h_1^{(\ell)}, f_2^{(\ell)}, h_2^{(\ell)}),\ f_2^{(\ell+1)} := \arg\sup_z \mathcal{D}(f_1^{(\ell+1)}, h_1^{(\ell)}, z, h_2^{(\ell)}), \\ (h_1^{(\ell+1)}, h_2^{(\ell+1)}) := \arg\sup_{z_1, z_2} \mathcal{D}(f_1^{(\ell+1)}, z_1, f_2^{(\ell+1)}, z_2). \end{aligned}$$

These maximizations can all be obtained in closed form, to result in the closed-form updates:

$$\begin{aligned} &\exp(\gamma f_1^{(\ell+1)}) = \left(\frac{a}{\xi^{(1)} \exp(\gamma h_1^{(\ell)})} \right)^{\frac{\tau_1}{\tau_1 + 1/\gamma}}, \quad \exp(\gamma f_2^{(\ell+1)}) = \left(\frac{b}{\xi^{(2)} \exp(\gamma h_2^{(\ell)})} \right)^{\frac{\tau_2}{\tau_2 + 1/\gamma}} \\ &g_{\ell+1} := \left(\xi^{(3)} \odot (\xi^{(1)})^T \exp(\gamma f_1^{(\ell+1)}) \odot (\xi^{(2)})^T \exp(\gamma f_2^{(\ell+1)}) \right)^{1/3} \\ &\exp(\gamma h_1^{(\ell+1)}) = \frac{g_{\ell+1}}{(\xi^{(1)})^T \exp(\gamma f_1^{(\ell+1)})}, \quad \exp(\gamma h_2^{(\ell+1)}) = \frac{g_{\ell+1}}{(\xi^{(2)})^T \exp(\gamma f_2^{(\ell+1)})} \end{aligned}$$

When using "scaling" representations for these dual variables, $\ell \geq 0$, $u_i^{(\ell)} := \exp(\gamma f_i^{(\ell)})$ and $v_i^{(\ell)} := \exp(\gamma h_i^{(\ell)})$ for $i \in \{1, 2\}$, we obtain a simple update, provided in the appendix (Alg. 5).

Initialization and Termination. We use the stopping criterion proposed in [Scetbon et al., 2021] to terminate the algorithm, $\Delta(\boldsymbol{\zeta}, \tilde{\boldsymbol{\zeta}}, \gamma) := \frac{1}{\gamma^2}(\mathrm{KL}(\boldsymbol{\zeta}, \tilde{\boldsymbol{\zeta}}) + \mathrm{KL}(\tilde{\boldsymbol{\zeta}}, \boldsymbol{\zeta}))$. Finding an efficient initialization is a research topic in itself, explored for instance in [Cuturi et al., 2022]. Here we adopt the practical choices proposed in [Scetbon and Cuturi, 2022], and follow them in adapting the choice of γ_k at each iteration k of the outer loop. We summarize our proposal in Algorithm 1, which can be seen as an extension of [Scetbon et al., 2021, Alg.2].

Convergence. The convergence proof for Dykstra's algorithm (Alg. 5) can be found in [Bauschke and Combettes, 2008]). In addition, [Scetbon et al., 2021] show the convergence of their scheme towards a stationary points w.r.t to the criterion $\Delta(\cdot, \cdot, \gamma)$ for γ fixed along the iterations of the outer loop. The stationary convergence of our proposed algorithm can be directly derived from their results.

Complexity. Given $\boldsymbol{\xi}$, solving Eq. (7) requires a time and memory complexity of $\mathcal{O}((n+m)r)$. However computing $\boldsymbol{\xi}$ requires in general $\mathcal{O}((n^2+m^2)r)$ time and $\mathcal{O}(n^2+m^2)$ memory. Scetbon et al. [2021] propose to consider low-rank approximation of the cost matrix C of the form $C \simeq C_1 C_2^T$ where $C_1 \in \mathbb{R}^{n \times d}$ and $C_2 \in \mathbb{R}^{m \times d}$. In that case computing $\boldsymbol{\xi}$ can be done in $\mathcal{O}((n+m)rd)$ time and $\mathcal{O}((n+m)(r+d))$ memory. Such approximations can be obtained using the algorithm in [Indyk et al., 2019] which guarantees that for any distance matrix $C \in \mathbb{R}^{n \times m}$ and $\alpha > 0$ it can outputs matrices $C_1 \in \mathbb{R}^{n \times d}$, $C_2 \in \mathbb{R}^{m \times d}$ in $\mathcal{O}((m+n)\mathrm{poly}(\frac{d}{\alpha}))$ algebraic operations such that with probability at least 0.99, $\|C - C_1 C_2^T\|_F^2 \leq \|C - C_d\|_F^2 + \alpha \|C\|_F^2$, where C_d denotes the best rank-d approximation to C.

```
Algorithm 1 ULOT$(C, a, b, r, \gamma_0, \tau_1, \tau_2, \delta)$
Inputs: $C, a, b, r, \gamma_0, \tau_1, \tau_2, \delta$
$Q, R, g \leftarrow$ Initialization as proposed in [Scetbon and Cuturi, 2022]
repeat
    $\tilde{Q} = Q, \quad \tilde{R} = R, \quad \tilde{g} = g,$
    $\nabla_Q = CR\,\mathrm{diag}(1/g), \quad \nabla_R = C^\top Q\,\mathrm{diag}(1/g),$
    $\omega \leftarrow \mathcal{D}(Q^T CR), \quad \nabla_g = -\omega/g^2,$
    $\gamma \leftarrow \gamma_0/\max(\|\nabla_Q\|_\infty^2, \|\nabla_R\|_\infty^2, \|\nabla_g\|_\infty^2),$
    $\xi^{(1)} \leftarrow Q \odot \exp(-\gamma\nabla_Q), \ \xi^{(2)} \leftarrow R \odot \exp(-\gamma\nabla_R), \ \xi^{(3)} \leftarrow g \odot \exp(-\gamma\nabla_g),$
    $Q, R, g \leftarrow$ ULR-Dykstra$(a, b, \boldsymbol{\xi}, \gamma, \tau_1, \tau_2, \delta)$ (Alg. 5)
until $\Delta((Q, R, g), (\tilde{Q}, \tilde{R}, \tilde{g}), \gamma) < \delta$;
Result: $Q, R, g$
```

3.2 Improvements on the Unbalanced Dykstra Algorithm

A well documented source of instability of unbalanced formulations of OT lies in capturing efficiently what optimal mass is targeted by such formulations. Séjourné et al. [2022] have proposed a technique to address this issue and lower significantly computational costs. They propose first a dual objective that is *translation* invariant. We take inspiration from this strategy and adapt it to our problem, to propose the following variant of (8):

$$\sup_{\tilde{f}_1, \tilde{h}_1, \tilde{f}_2, \tilde{h}_2} \left(\mathcal{D}_{\mathrm{TI}}(\tilde{f}_1, \tilde{h}_1, \tilde{f}_2, \tilde{h}_2) := \sup_{\lambda_1, \lambda_2 \in \mathbb{R}} \mathcal{D}(\tilde{f}_1 + \lambda_1, \tilde{h}_1 - \lambda_1, \tilde{f}_2 + \lambda_2, \tilde{h}_2 - \lambda_2) \right) \tag{9}$$

It is clear from the reparameterization that both problems (8) and (9) have the same value and also that $(\tilde{f}_1, \tilde{h}_1, \tilde{f}_2, \tilde{h}_2)$ is solution of (9) if and only if $(\tilde{f}_1 + \lambda_1^\star, \tilde{h}_1 - \lambda_1^\star, \tilde{f}_2 + \lambda_2^\star, \tilde{h}_2 - \lambda_2^\star)$ is solution of (8) where $(\lambda_1^\star, \lambda_2^\star)$ solves $\mathcal{D}_{\mathrm{TI}}(\tilde{f}_1, \tilde{h}_1, \tilde{f}_2, \tilde{h}_2)$. To solve (9), we show that the variational formulation of the translation invariant dual objective targeted inside (9) can be obtained in closed form.

Proposition 2. *Let $\tilde{f}_1 \in \mathbb{R}^n$, $\tilde{f}_2 \in \mathbb{R}^m$ and $\tilde{h}_1, \tilde{h}_2 \in \mathbb{R}^r$, then the inner problem defined in* (9) *by $\mathcal{D}_{TI}(\tilde{f}_1, \tilde{h}_1, \tilde{f}_2, \tilde{h}_2)$ admits a unique solution $(\lambda_1^\star, \lambda_2^\star)$ and we have that*

$$\lambda_1^\star := \left(1 - \frac{\tau_1\tau_2}{(1/\gamma + \tau_1)(1/\gamma + \tau_2)}\right)^{-1} \left(\frac{\tau_1/\gamma}{1/\gamma + \tau_1} c_1 - \frac{\tau_1/\gamma}{1/\gamma + \tau_1}\frac{\tau_2}{1/\gamma + \tau_2} c_2\right) \tag{10}$$

$$\lambda_2^\star := \left(1 - \frac{\tau_1\tau_2}{(1/\gamma + \tau_1)(1/\gamma + \tau_2)}\right)^{-1} \left(\frac{\tau_2/\gamma}{1/\gamma + \tau_2} c_2 - \frac{\tau_1/\gamma}{1/\gamma + \tau_1}\frac{\tau_2}{1/\gamma + \tau_2} c_1\right) \tag{11}$$

where

$$c_1 := \log\left(\frac{\langle \exp(-\tilde{f}_1/\tau_1), a\rangle}{\langle \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2)), \xi^{(3)}\rangle}\right), \quad \textit{and} \quad c_2 := \log\left(\frac{\langle \exp(-\tilde{f}_2/\tau_2), a\rangle}{\langle \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2)), \xi^{(3)}\rangle}\right).$$

Using Proposition 2, we perform an alternate maximization scheme on the translation invariant formulation of the dual $\mathcal{D}_{\mathrm{TI}}$. Indeed using Danskin's theorem (under the assumption that $\lambda_1^\star, \lambda_2^\star$ do not diverge), one obtains a variant of Algorithm 5, summarized in Algorithm 3.

```
Algorithm 2 compute-lambdas$(a, b, \xi^{(3)}, u_1, v_1, u_2, v_2, \gamma, \tau_1, \tau_2)$
Inputs: $a, b, \xi^{(3)}, u_1, v_1, u_2, v_2, \gamma, \tau_1, \tau_2$
$\tilde{u}_1 \leftarrow u_1^{-1/\gamma/\tau_1}, \ \tilde{u}_2 \leftarrow u_2^{-1/\gamma/\tau_2}$
$c_1 \leftarrow \log(\langle \tilde{u}_1, a\rangle) - \log(\langle \xi^{(3)}, v_1^{-1} \odot v_2^{-1}\rangle), \ c_2 \leftarrow \log(\langle \tilde{u}_2, b\rangle) - \log(\langle \xi^{(3)}, v_1^{-1} \odot v_2^{-1}\rangle)$
Result: $\lambda_1^\star, \ \lambda_2^\star$ as in (10), (11)
```

Algorithm 3 ULR-TI-Dykstra$(a, b, \boldsymbol{\xi}, \gamma, \tau_1, \tau_2, \delta)$

```
Inputs: $a, b, \boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)}), \gamma, \tau_1, \tau_2, \delta$
$v_1 = v_2 = \mathbf{1}_r,\ u_1 = \mathbf{1}_n,\ u_2 = \mathbf{1}_m$
repeat
    $\tilde{v}_1 = v_1,\ \tilde{v}_2 = v_2, \tilde{u}_1 = u_1, \tilde{u}_2 = u_2$
    $\lambda_1, \lambda_2 \leftarrow \text{compute-lambdas}(a, b, \xi^{(3)}, u_1, v_1, u_2, v_2, \gamma, \tau_1, \tau_2)$ (Alg. 2)
    $u_1 = \left(\frac{a}{\xi^{(1)} v_1}\right)^{\frac{\tau_1}{\tau_1 + 1/\gamma}} \exp(-\lambda_1/\tau_1)^{\frac{\tau_1}{1/\gamma + \tau_1}}, \quad u_2 = \left(\frac{b}{\xi^{(2)} v_2}\right)^{\frac{\tau_2}{\tau_2 + 1/\gamma}} \exp(-\lambda_2/\tau_2)^{\frac{\tau_2}{1/\gamma + \tau_2}},$
    $\lambda_1, \lambda_2 \leftarrow \text{compute-lambdas}(a, b, \xi^{(3)}, u_1, v_1, u_2, v_2, \gamma, \tau_1, \tau_2)$ (Alg. 2)
    $g = \exp(\gamma(\lambda_1 + \lambda_2))^{1/3} \left(\xi^{(3)} \odot (\xi^{(1)})^T u_1 \odot (\xi^{(2)})^T u_2\right)^{1/3},\ v_1 = \frac{g}{(\xi^{(1)})^T u_1},\ v_2 = \frac{g}{(\xi^{(2)})^T u_2}$
until $\frac{1}{\gamma} \max(\| \log(u_i/\tilde{u}_i)\|_\infty, \| \log(v_i/\tilde{v}_i)\|_\infty) < \delta$;
Result: $\operatorname{diag}(u_1)\xi_k^{(1)} \operatorname{diag}(v_1), \quad \operatorname{diag}(u_2)\xi_k^{(2)} \operatorname{diag}(v_2), \quad g$
```

3.3 Unbalanced Low-rank Gromov-Wasserstein

The low-rank Gromov-Wasssertein (LGW) problem [Scetbon et al., 2022] between the two discrete metric measure spaces $(\mu, d_{\mathcal{X}})$ and $(\nu, d_{\mathcal{Y}})$, written for compactness using (a, A) and (b, B), reads

$$\mathrm{LGW}_r((a, A), (b, B)) = \min_{P \in \Pi_{a,b}(r)} \mathcal{Q}_{A,B}(P), \tag{12}$$

Building upon § 3.1, we introduce the unbalanced low-rank Gromov-Wasserstein (ULGW) problem. There is, however, a significant challenge that appears when introducing unbalanced regularizers in (12): When P is constrained to be in $\Pi_{a,b}$, the first two terms of the RHS in (12) simplify to $a^T A^{\odot 2} a + b^T B^{\odot 2} b$. Hence, they are constant and discarded when optimizing. In an unbalanced setting, these terms vary and must be accounted for:

$$\begin{aligned} \mathrm{ULGW}_r((a, A), (b, B)) := \min_{(Q,R,g) \in \Pi_r} \langle A^{\odot 2} Q \mathbf{1}_r, Q \mathbf{1}_r \rangle + \langle B^{\odot 2} R \mathbf{1}_r, R \mathbf{1}_r \rangle \\ - 2 \langle AQ \operatorname{diag}(1/g) R^T B, Q \operatorname{diag}(1/g) R^T \rangle + \tau_1 \mathrm{KL}(Q \mathbf{1}_r | a) + \tau_2 \mathrm{KL}(R \mathbf{1}_r | b) \end{aligned} \tag{13}$$

To solve the problem, we apply the same scheme as proposed for ULOT, that is a proximal gradient descent where we linearize $\mathcal{Q}_{A,B}$ and add a KL penalization while leaving the soft marginal constraints unchanged. Therefore the algorithm to solve ULGW is the same as that solving ULOT, however, the kernels $\boldsymbol{\xi}_k$ now take into account the quadratic terms of the original LGW problem. More formally, at each iteration k of the outer loop, we propose to solve

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta} \in \Pi_r} \frac{1}{\gamma_k} \mathrm{KL}(\boldsymbol{\zeta} | \boldsymbol{\xi}_k) + \tau_1 \mathrm{KL}(Q \mathbf{1}_r | a) + \tau_2 \mathrm{KL}(R \mathbf{1}_r | b), \tag{14}$$

where $(Q_0, R_0, g_0) \in \Pi_r$ is an initial point, $(\gamma_k)_{k \geq 0}$ is a sequence of positive step sizes, $P_k = Q_k \operatorname{diag}(1/g_k) R_k^T$, $\boldsymbol{\xi}_k := (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)})$ and

$$\begin{aligned} \xi_k^{(1)} &:= Q_k \odot \exp(-2\gamma_k A^{\odot 2} Q_k \mathbf{1}_r \mathbf{1}_r^T) \odot \exp(-4\gamma_k A P_k B R_k \operatorname{diag}(1/g_k))) \\ \xi_k^{(2)} &:= R_k \odot \exp(-2\gamma_k B^{\odot 2} R_k \mathbf{1}_r \mathbf{1}_r^T) \odot \exp(-4\gamma_k B P_k^T A Q_k \operatorname{diag}(1/g_k))) \\ \xi_k^{(3)} &:= g_k \odot \exp(4\gamma_k \omega_k / g_k^2) \quad \text{with} \quad [\omega_k]_i := [Q_k^T A P_k B R_k]_{i,i} \ \ \forall\, i \in \{1, \ldots, r\}. \end{aligned}$$

Note that (14) is the exact same optimization problem as (7), where only $\boldsymbol{\xi}_k$ has changed and therefore can be solved using Algorithm 3. Algorithm 4 summarizes our strategy to solve (13).

Algorithm 4 ULGW$(A, B, a, b, r, \gamma_0, \tau_1, \tau_2, \delta)$

```
Inputs: $A, B, a, b, r, \gamma_0, \tau_1, \tau_2, \delta$
$Q, R, g \leftarrow$ Initialization as proposed in [Scetbon and Cuturi, 2022]
repeat
    $\tilde{Q} = Q, \;\; \tilde{R} = R, \;\; \tilde{g} = g,$
    $\nabla_Q = 4AQ\,\mathrm{diag}(1/g)R^TBR\,\mathrm{diag}(1/g) + 2A^{\odot 2}Q\mathbf{1}_r\mathbf{1}_r^T,$
    $\nabla_R = 4BR\,\mathrm{diag}(1/g)Q^TAQ\,\mathrm{diag}(1/g) + 2B^{\odot 2}R\mathbf{1}_r\mathbf{1}_r^T,$
    $\omega \leftarrow \mathcal{D}(Q^TAQ\,\mathrm{diag}(1/g)R^TBR), \;\; \nabla_g = -\omega/g^2,$
    $\gamma \leftarrow \gamma_0/\max(\|\nabla_Q\|_\infty^2, \|\nabla_R\|_\infty^2, \|\nabla_g\|_\infty^2),$
    $\xi^{(1)} \leftarrow Q \odot \exp(-\gamma\nabla_Q), \; \xi^{(2)} \leftarrow R \odot \exp(-\gamma\nabla_R), \; \xi^{(3)} \leftarrow g \odot \exp(-\gamma_k\nabla_g),$
    $Q, R, g \leftarrow$ ULR-TI-Dykstra$(a, b, \boldsymbol{\xi}, \gamma, \tau_1, \tau_2, \delta)$ (Alg. 3)
until $\Delta((Q, R, g), (\tilde{Q}, \tilde{R}, \tilde{g}), \gamma) < \delta$;
Result: $Q, R, g$
```

Convergence and Complexity. Similarly to linear ULOT, the unbalanced Dykstra algorithm is guaranteed to converge [Bauschke and Lewis, 2000]. in addition, [Scetbon et al., 2022] prove the convergence of their scheme to a stationary point of the problem. Because we use Algorithm 5, we retain exactly the same complexity, both in terms of time of memory, to solve these inner problems. The slight variation in kernel $\boldsymbol{\xi}$ compared to ULOT still retains the same $\mathcal{O}((n^2 + m^2)r)$ time and $\mathcal{O}(n^2 + m^2)$ memory complexities. However, as in ULOT, we can take advantage of low-rank approximations of the costs matrices A and B to reach linear complexity. Indeed, assuming $A \simeq A_1A_2^T$ and $B \simeq B_1B_2$ where $A_1, A_2 \in \mathbb{R}^{n\times d_X}$ and $B_1, B_2 \in \mathbb{R}^{m\times d_Y}$, then the total time and memory complexities become respectively $\mathcal{O}(mr(r + d_Y) + nr(r + d_X))$ and $\mathcal{O}((n + m)(r + d_X + d_Y))$. Again, when A and B are distance matrices, we use the algorithms from [Indyk et al., 2019].

3.4 Unbalanced Low-rank Fused-Gromov-Wasserstein

We finally focus on the increasingly popular [Klein et al., 2023] fused-Gromov-Wasserstein problem, which merges linear and quadratic objectives [Vayer et al., 2018]:

$$\mathrm{FGW}(\mu, \nu) := \min_{P\in\Pi_{a,b}} \alpha\langle C, P\rangle + \bar{\alpha}\mathcal{Q}_{A,B}(P) \tag{15}$$

where $\alpha \in [0, 1]$ and $\bar{\alpha} := 1-\alpha$ allows interpolating between the GW and linear OT geometries. This problem remains a GW problem, where one replaces the 4-way cost $M[i, i', j, j'] := (A_{i,i'} - B_{j,j'})^2$ appearing in (4) by a composite interpolated cost between the OT and GW geometries, redefined as $M[i, i', j, j'] = \alpha C_{i,j} + \bar{\alpha}(A_{i,i'} - B_{j,j'})^2$. Our proposed unbalanced and low-rank version of the FGW problem includes $|P| := \|P\|_1$ the mass of P, to homogenize linear and quadratic terms,

$$\mathrm{ULFGW}_r(\mu, \nu) := \min_{P\colon\, \mathrm{rk}_+(P)\leq r} \alpha|P|\langle C, P\rangle + \bar{\alpha}\mathcal{Q}_{A,B}(P) + \tau_1\mathrm{KL}(P\mathbf{1}_m|a) + \tau_2\mathrm{KL}(P^T\mathbf{1}_n|b)\,, \tag{16}$$

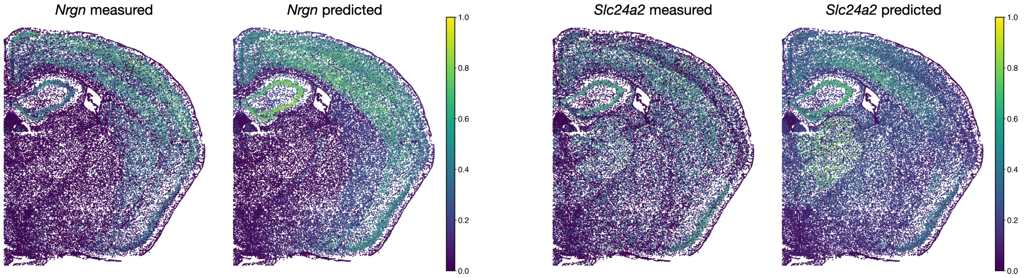
which is expanded through the explicit factorization of P, noticing that $|P| = |g| := \|g\|_1$:

$$\mathrm{ULFGW}_r(\mu, \nu) := \min_{(Q,R,g)\in\Pi_r} \alpha|g|\mathcal{L}_C(Q, R, g) + \bar{\alpha}\mathcal{Q}_{A,B}(Q, R, g) + \mathcal{G}_{a,b}(Q, R, g) \tag{17}$$

Then by linearizing again $\mathcal{H} : (Q, R, g) \to \alpha|g|\mathcal{L}_C(Q, R, g) + \bar{\alpha}\mathcal{Q}_{A,B}(Q, R, g)$ with an added KL penalty and leaving $\mathcal{G}_{a,b}$ unchanged, we obtain at each iteration, the same optimization problem as in (14) where the kernels $\boldsymbol{\xi}_k$ are now defined as

$$\begin{aligned}
\boldsymbol{\xi}_k &:= (\xi_k^{(1)}, \xi_k^{(2)}, \xi_k^{(3)}),\\
\xi_k^{(1)} &:= Q_k \odot \exp(-\gamma_k\nabla_Q\mathcal{H}_k), \; \xi_k^{(2)} := R_k \odot \exp(-\gamma_k\nabla_Q\mathcal{H}_k), \; \xi_k^{(3)} := g_k \odot \exp(-\gamma_k\nabla_g\mathcal{H}_k)\\
\nabla_Q\mathcal{H}_k &:= \alpha|g_k|CR_k\,\mathrm{diag}(1/g_k) + \bar{\alpha}\left(2A^{\odot 2}Q_k\mathbf{1}_r\mathbf{1}_r^T + 4AP_kBR_k\,\mathrm{diag}(1/g_k)\right)\\
\nabla_R\mathcal{H}_k &:= \alpha|g_k|C^TQ_k\,\mathrm{diag}(1/g_k) + \bar{\alpha}\left(2B^{\odot 2}R_k\mathbf{1}_r\mathbf{1}_r^T + 4BP_k^TAQ_k\,\mathrm{diag}(1/g_k)\right)\\
\nabla_g\mathcal{H}_k &:= \alpha\left(\langle C, P_k\rangle\mathbf{1}_r - |g_k|\omega_k^{\mathrm{lin}}/g_k^2\right) - 4\bar{\alpha}\omega_k^{\mathrm{quad}}/g_k^2\\
[\omega_k^{\mathrm{lin}}]_i &:= [Q_k^TCR_k]_{i,i}, \;\; [\omega_k^{\mathrm{quad}}]_i := [Q_k^TAP_kBR_k]_{i,i} \;\; \forall\, i \in \{1, \ldots, r\}\,.
\end{aligned}$$

These steps are summarized in Algorithm 6, proposed in the appendix. These steps result usually in a quadratic complexity, both in time and memory, with respect to the number of points n and m. These complexities become linear as soon as all three matrices C, A, B admit a low-rank factorization.



(a) Visualization of measured and predicted gene expression of *Nrgn*.

(b) Visualization of measured and predicted gene expression of *Slc24a2*.

Figure 1: Spatial visualization of the two mouse brain sections used in **Exp. 2**

4 Experiments

We focus first in **Exp. 1** on demonstrating the empirical benefits of the translation invariant (TI) variant of our algorithms, as implemented in Algorithm 3, and which is subsequently used as an inner routine to solve ULR problems. We compare in **Exp. 2** *unbalanced* low-rank (ULR) solvers to *balanced* low-rank (LR) counterparts, and follow in **Exp. 3** by comparing ULR solvers to entropic (E) counterparts. We conclude in **Exp. 4** by comparing ULR solvers to [Thual et al., 2022], which can learn a sparse transport coupling, in the unbalanced FGW setting.

Datasets. We consider two real-world datasets, described in B.1, and two synthetic datasets, that are large enough to showcase our solvers. The real-world datasets consist of both a shared feature space, used to compute the costs matrices for the linear term in the OT and FGW settings, as well as geometries that are specific to each source s and target t measures, and which are used to compute the costs matrices for the quadratic term in the GW and FGW settings. In **Exp. 1**, we simply consider high dimensional Gaussians and mixture of Gaussians to evaluate the performance of the TI variant. We use the mouse brain STARmap spatial transcriptomics data from [Shi et al., 2022] for **Exp. 2** and **Exp. 3**. We use data from the Individual Brain Charting dataset [Pinho et al., 2018], to replicate the settings of [Thual et al., 2022], in **Exp. 4**.

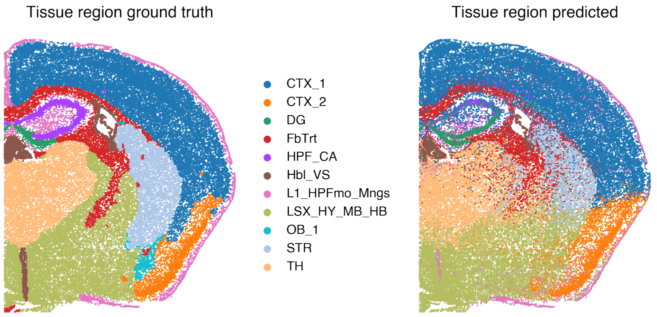


Figure 2: Visualization of measured and predicted tissue regions in the mouse brain in **Exp. 2**

Metrics. Following Klein et al. [2023], we evaluate maps by focusing on the two following metrics: (i) **pearson correlation** ρ computed between the source s feature matrix F^s and the barycentric projection of the target t to the source scaled by the target marginals b^t: $T^T_{t\to s}\,(F^t\ \frac{1}{b^t})$; (ii) **F1 score** computed between the original source s labels l^s and the inferred source labels, computed by taking the $\operatorname{argmax}_j B_{i,j}$ of the barycentric projection of the target t one hot encoded labels L^t, scaled by the target marginal b^t, to the source $T^T_{t\to s}\,(L^t\ \frac{1}{b^t})$.

Experiment 1: Benchmarking The Translation Invariant Variant. We evaluate the effect of the proposed TI procedure on the computational cost of ULR solvers: We compare the time taken when solving unbalanced LR problems, with or without using the TI objective. In Figure 3, we compare the execution time (using our `ott-jax` implementation) of unbalanced LR Sinkhorn on large and high dimensional Gaussian distributions. The results presented are averaged over 10 random seeds with error bars. We use a $1e-9$ convergence threshold and 1000 maximal number of iterations for Dykstra, in 64-bit precision. We observe that the use of our proposed TI objective is consistently beneficial when solving ULR problems. See also Appendix B.3 for additional experiments.

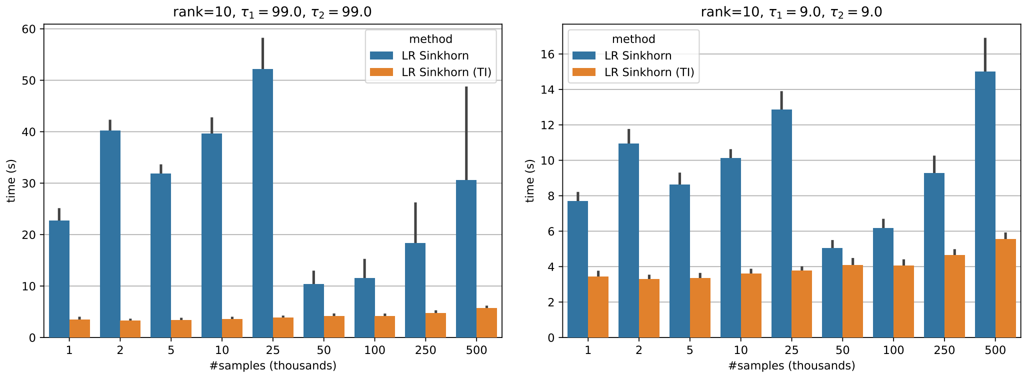


Figure 3: Execution time of unbalanced LR Sinkhorn, with (Alg. 3) or without (Alg. 5) the TI variant. We fix the rank to $r = 10$; n points (displayed in thousands) are sampled from two Gaussian distributions in $d = 30$ of means respectively -1.2 and 1.3, and standard deviations 1 and 0.2. (left) displays large τ (close to balanced), (right) is smaller τ (more unbalanced). We use the *same convergence threshold* for the outer loop, for all sample sizes. As n gets bigger, this results in a relatively *looser* threshold, explaining why timings can slightly decrease w.r.t. n. What matters is, therefore, the comparative performance of TI vs non-TI for a fixed n, *not the behaviour w.r.t. n*.

Experiment 2: ULOT vs. LOT on Gene Expression / Cell Type Annotation. We evaluate the accuracy of ULOT solvers for a large-scale spatial transcriptomics task, using gene expression mapping and cell type annotation. We compare it to the balanced LR alternative using the Pearson correlation ρ as described in the metrics section. We leverage two coronal sections of the mouse brain profiled by STARmap spatial transcriptomics by [Shi et al., 2022]. They consist of $n \approx 40,000$ cells in both the source and target brain section. Each cell is described by 1000 gene features, in addition to 2D spatial coordinates. As a result A, B are $\approx 40k \times 40k$, and the fused term C is a squared-Euclidean distance matrix on 30D PCA space computed on the gene expression space. We selected 10 marker genes for the validation and test sets from the *HPF_CA* cluster. We run an extensive grid search as reported in B.2, we pick the best hyperparameters combination using performance on the 10 validation genes as a criterion, and we report that metric on the other genes in Table 1, as well as qualitative results in Figure 1 and Figure 2. Clearly, ULFGW is the best performing solver across all metrics. Interestingly, the ULOT does not consistently outperforms its balanced version, and unbalancedness seems to hurt performance for the LGW solvers. Nevertheless, both solvers display inconsistent performance across metrics, whereas the ULFGW and LFGW are consistently superior to the rest of the solvers. These results highlight how the flexibility given by the FGW formulation to leverage common and disparate geometries, paired with the unbalancedness relaxation, can provide state of the art algorithms for matching problems in large-scale, real world biological problems.

solver	mass %	val ρ	test ρ	F1 mac.	F1 mic.	F1 weig.
LOT	1.000	0.282	0.386	0.210	0.411	0.360
ULOT	0.889	0.301	0.409	0.200	0.425	0.363
LGW	1.000	0.227	0.288	0.487	0.716	0.692
ULGW	1.001	0.222	0.287	0.463	0.701	0.665
LFGW	1.000	0.365	0.443	0.576	0.720	0.714
ULFGW	0.443	**0.379**	**0.463**	**0.582**	**0.733**	**0.724**

Table 1: **Exp.2**, Results for spatial transcriptomics dataset (brain coronal section from Shi et al. [2022]).

Experiment 3: ULOT vs. UEOT. We compare the performance of ULOT solvers to their unbalanced entropic alternatives (UEOT). We use the same datasets as in **Exp. 2**, but must pick a smaller subset (Olfactory bulb), to avoid OOM errors for entropic UGW solvers, since they cannot handle the $40k$ sizes considered in **Exp. 2** (see B.1). This results in $n \approx 20,000$ source and $\approx 15,000$ target cells, and 1000 genes. Similar to **Exp. 2**, the fused term C is a squared-Euclidean distance matrix

solver	mass %	val ρ	test ρ	F1 mac.	F1 mic.	F1 weig.
UEOT	1.012	0.368	0.479	0.511	0.763	0.751
LOT	1.000	0.335	0.440	0.511	0.760	0.751
ULOT	0.998	0.356	0.461	0.518	0.770	0.762
UEFGW	1.015	0.343	0.475	**0.564**	**0.839**	**0.831**
LFGW	1.000	0.348	0.453	0.512	0.762	0.753
ULFGW	0.339	**0.368**	**0.491**	0.556	0.826	0.818

Table 2: **Exp. 3**: Results for spatial transcriptomics dataset (Olfactory bulb section from Shi et al. [2022]).

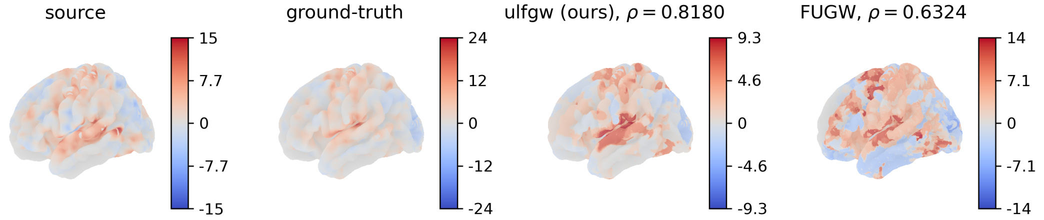


Figure 4: Visualization of measured and predicted *right auditory click* contrast map in **Exp.4**.

on 30-D PCA space, computed on gene expressions. As done in **Exp. 2**, we select 10 marker genes for the validation and 10 genes for the test set, from cluster *OB_1*. We run an extensive grid search, as in **Exp. 2** and B.2. In Table 2, shows that ULFGW outperforms entropic solvers w.r.t. ρ, but is worse when considering F1 scores. On the other hand, ULFGW confirms its superiority compared to the balanced alternative LFGW. Taken together, these results suggest that while unbalanced LR solvers are on par with unbalanced entropic solvers in terms of performance, in small data regimes, they unlock the applications of unbalanced OT to larger scales.

Experiment 4: ULOT to align brain meshes. In this experiment, we compare the performance of our ULFGW solver to FUGW-sparse, a new approach of the unbalanced FGW problem based on a full-rank formulation proposed in Thual et al. [2022]. This method was demonstrated to be effective in aligning brain anatomies, encompassing both mesh structures and functional signals associated with each vertex. For their empirical analysis, they utilized the Individual Brain Charting dataset Pinho et al. [2018].

The dataset uses the *fsaverage7* mesh, which describes $n \approx 160,000$ vertices. We embed them into a 30-dimensional embedding space using an approximation of the geodesic distances with landmark multi-dimensional scaling [De Silva and Tenenbaum, 2004] where 2048 points were used as anchors. Each vertex has an associated functional signal that entails 22 features. For both the quadratic and linear terms, we compute the costs based on the squared Euclidean distance. We evaluate the performance of the method by comparing each best hyperparameter combinations based on the average correlation between the barycentric projection and ground-truth value of 5 features, across a test set of 5 contrast maps. See also Appendix B.2 for additional experimental details and results. In Table 3, we observe that ULFGW and LFGW outperforms FUGW-sparse. In this setting, there is no clear evidence that the unbalanced version performs better than its balanced counterpart for low-rank methods.

solver	**mass**	**val ρ**	**test ρ**
FUGW-sparse	0.999	0.492	0.472
LFGW	1.000	0.513	**0.663**
ULFGW	0.981	**0.533**	0.643

Table 3: Results on the brain anatomy with functional signal data from Pinho et al. [2018] in **Exp.4**.

Conclusion. Recent practical successes of OT methods to natural sciences have demonstrated the relevance of OT to their analysis pipelines, but have also shown, repeatedly, that a certain degree of freedom to depart from the rigid assumption of mass conservation is needed in practice. On the other hand, and across the same range of applications, low-rank approaches can hold the promise of scaling OT methods to relevant sample sizes for natural sciences. This paper merges these two strains and demonstrate the practical relevance of these novel algorithms.

References

Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice hall, 1993.

Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *arXiv preprint arXiv:1705.09634*, 2017.

Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016.

Heinz H Bauschke and Patrick L Combettes. A Dykstra-like algorithm for two monotone operators. *Pacific Journal of Optimization*, 4(3):383–391, 2008.

Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

Andrew J Blumberg, Mathieu Carriere, Michael A Mandell, Raul Rabadan, and Soledad Villar. Mrec: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*, 2020.

Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021. doi: 10.1101/2021.12.15.472775. URL `https://www.biorxiv.org/content/early/2021/12/15/2021.12.15.472775`.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised Training of Conditional Monge Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zixuan Cang, Yanxiang Zhao, Axel A Almet, Adam Stabell, Raul Ramos, Maksim V Plikus, Scott X Atwood, and Qing Nie. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods*, 20(2):218–228, 2023.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123, 2018.

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

Samir Chowdhury, David Miller, and Tom Needham. Quantized gromov-wasserstein. *arXiv preprint arXiv:2104.02013*, 2021.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.

Vin De Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, technical report, Stanford University, 2004.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. Gromov-wasserstein optimal transport to align single-cell multi-omics data. *bioRxiv*, 2020. doi: 10.1101/2020.04.28.066787.

Richard Mansfield Dudley et al. Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. *Illinois Journal of Mathematics*, 10(1):109–126, 1966.

Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

Kilian Fatras, Thibault Sejourne, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3186–3197. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/fatras21a.html`.

Jean Feydy, Pierre Roussillon, Alain Trouvé, and Pietro Gori. Fast and scalable optimal transport for brain tractograms. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 636–644. Springer, 2019.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL `http://jmlr.org/papers/v22/20-451.html`.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings, 2018.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample Complexity of Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

Piotr Indyk, Ali Vakilian, Tal Wagner, and David Woodruff. Sample-optimal low-rank approximation of distance matrices, 2019.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimee Bastidas-Ponce, Marta Tarquis-Medina, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, and Fabian J. Theis. Mapping cells through time and space with moscot. *bioRxiv*, 2023. doi: 10.1101/2023.05.11.540374. URL `https://www.biorxiv.org/content/early/2023/05/11/2023.05.11.540374`.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Khang Le, Huy Nguyen, Quang M Nguyen, Tung Pham, Hung Bui, and Nhat Ho. On robust optimal transport: Computational complexity and barycenter computation. *Advances in Neural Information Processing Systems*, 34:21947–21959, 2021.

Jiajin Li, Jianheng Tang, Lemin Kong, Huikang Liu, Jia Li, Anthony Man-Cho So, and Jose Blanchet. A convergent single-loop algorithm for relaxation of gromov-wasserstein in graph data. In *The Eleventh International Conference on Learning Representations*, 2023.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

Mor Nitzan, Nikos Karaiskos, Nir Friedman, and Nikolaus Rajewsky. Gene expression cartography. *Nature*, 576(7785), 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019. ISSN 1935-8245.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2020.

Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, Véronique Joly-Testault, Gaëlle Médiouni-Cloarec, Christine Doublé, Bernadette Martins, Philippe Pinel, Evelyn Eger, Gael Varoquaux, Christophe Pallier, Stanislas Dehaene, Lucie Hertz-Pannier, and Bertrand Thirion. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. *Scientific Data* , 5:180105, June 2018. doi: 10.1038/sdata.2018.105. URL https://hal.science/hal-01817528.

R Tyrrell Rockafellar. *Convex analysis*. Princeton university press, 1970.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

Ryoma Sato, Marco Cuturi, Makoto Yamada, and Hisashi Kashima. Fast and robust comparison of probability measures in heterogeneous spaces. *arXiv preprint arXiv:2002.01615*, 2020.

Meyer Scetbon and Marco Cuturi. Linear time sinkhorn divergences using positive features. *Advances in Neural Information Processing Systems*, 33:13468–13480, 2020.

Meyer Scetbon and Marco Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 6802–6814. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2d69e771d9f274f7c624198ea74f5b98-Paper-Conference.pdf.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR, 2021.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov-Wasserstein distances using low rank couplings and costs. *ICML*, 2022.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4995–5021. PMLR, 2022.

Hailing Shi, Yichun He, Yiming Zhou, Jiahao Huang, Brandon Wang, Zefang Tang, Peng Tan, Morgan Wu, Zuwan Lin, Jingyi Ren, Yaman Thapa, Xin Tang, Albert Liu, Jia Liu, and Xiao Wang. Spatial atlas of the mouse central nervous system at molecular resolution. *bioRxiv*, 2022. doi: 10.1101/2022.06.20.496914. URL https://www.biorxiv.org/content/early/2022/06/22/2022.06.20.496914.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):1–13, 2016.

Yi Tay, Dara Bahri, Liu Yang, Donald Metzler, and Da-Cheng Juan. Sparse Sinkhorn attention. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9438–9447. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/tay20a.html.

Alexis Thual, Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with fused unbalanced Gromov-Wasserstein. *arXiv*, 2022.

Titouan Vayer, Laetita Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects: theoretical foundations and mathematical properties. *arXiv preprint arXiv:1811.02834*, 2018.

Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed gromov-wasserstein divergence and applications on graphs. In *International Conference on Learning Representations*, 2023.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 2018.

Hongteng Xu, Dixin Luo, and Lawrence Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/6e62a992c676f611616097dbea8ea030-Paper.pdf.

Appendix

A Algorithms

Algorithm 5 ULR-Dykstra$(a, b, \boldsymbol{\xi}, \gamma, \tau_1, \tau_2, \delta, \alpha)$

```
Inputs: $a, b, \boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)}, \xi^{(3)}), \gamma, \tau_1, \tau_2, \delta$
$v_1 = v_2 = \mathbf{1}_r, u_1 = \mathbf{1}_n, u_2 = \mathbf{1}_m$
repeat
    $\tilde{v}_1 = v_1,\ \tilde{v}_2 = v_2, \tilde{u}_1 = u_1, \tilde{u}_2 = u_2$
    $u_1 = \left(\frac{a}{\xi^{(1)} v_1}\right)^{\frac{\tau_1}{\tau_1 + 1/\gamma}}, \quad u_2 = \left(\frac{b}{\xi^{(2)} v_2}\right)^{\frac{\tau_2}{\tau_2 + 1/\gamma}},$
    $g = \left(\xi^{(3)} \odot (\xi^{(1)})^T u_1 \odot (\xi^{(2)})^T u_2\right)^{1/3},\ v_1 = \frac{g}{(\xi^{(1)})^T u_1},\ v_2 = \frac{g}{(\xi^{(2)})^T u_2}$
until $\frac{1}{\gamma} \max(\| \log(u_i/\tilde{u}_i)\|_\infty, \| \log(v_i/\tilde{v}_i)\|_\infty) < \delta$;
Result: $\operatorname{diag}(u_1)\xi_k^{(1)} \operatorname{diag}(v_1), \quad \operatorname{diag}(u_2)\xi_k^{(2)} \operatorname{diag}(v_2), \quad g$
```

Algorithm 6 ULFGW$(A, B, a, b, r, \gamma_0, \tau_1, \tau_2, \delta)$

```
Inputs: $A, B, C, a, b, r, t, \gamma_0, \tau_1, \tau_2, \delta, \alpha$
$Q, R, g \leftarrow$ Initialization as proposed in [Scetbon and Cuturi, 2022]
repeat
    $\tilde{Q} = Q,\ \ \tilde{R} = R,\ \ \tilde{g} = g,$
    $\nabla_Q = \alpha |g| C R \operatorname{diag}(1/g) + \bar{\alpha} \left(2A^{\odot 2} Q \mathbf{1}_r \mathbf{1}_r^T + 4AQ \operatorname{diag}(1/g) R^T B R \operatorname{diag}(1/g)\right),$
    $\nabla_R = \alpha |g| C^T Q \operatorname{diag}(1/g) + \bar{\alpha} \left(2B^{\odot 2} R \mathbf{1}_r \mathbf{1}_r^T + 4BR \operatorname{diag}(1/g) Q^T A Q \operatorname{diag}(1/g)\right),$
    $\omega^{\text{lin}} \leftarrow \mathcal{D}(Q^T C R),\ \ \omega^{\text{quad}} \leftarrow \mathcal{D}(Q^T A Q \operatorname{diag}(1/g) R^T B R)$
    $\nabla_g = \alpha \left(\langle C, Q \operatorname{diag}(1/g) R^T \rangle \mathbf{1}_r - |g_k| \omega^{\text{lin}}/g^2\right) - 4\bar{\alpha}\omega^{\text{quad}}/g^2,$
    $\gamma \leftarrow \gamma_0 / \max(\|\nabla_Q\|_\infty^2, \|\nabla_R\|_\infty^2, \|\nabla_g\|_\infty^2),$
    $\xi^{(1)} \leftarrow Q \odot \exp(-\gamma \nabla_Q),\ \xi^{(2)} \leftarrow R \odot \exp(-\gamma \nabla_R),\ \xi^{(3)} \leftarrow g \odot \exp(-\gamma_k \nabla_g),$
    $Q, R, g \leftarrow$ ULR-TI-Dykstra$(a, b, \boldsymbol{\xi}, \gamma, \tau_1, \tau_2, \delta)$ (Alg. 3)
until $\Delta((Q, R, g), (\tilde{Q}, \tilde{R}, \tilde{g}), \gamma) < \delta$;
Result: $Q, R, g$
```

B Experiments

B.1 Datasets and preprocessing

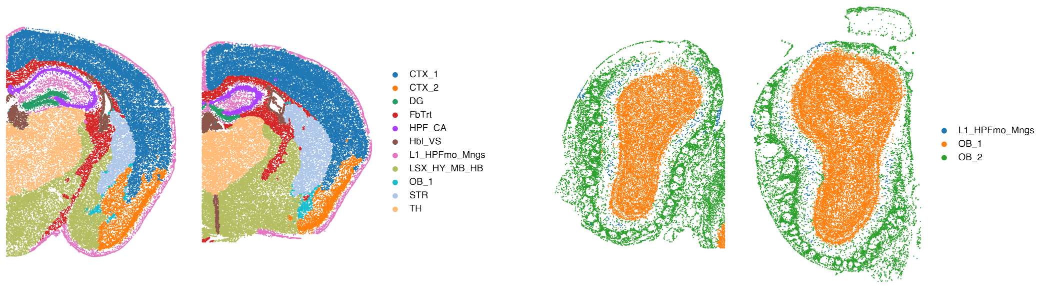
We downloaded the two publicly available datasets from the respective publications:

- STARmap mouse brain sections from [Shi et al., 2022]
- Brain mesh anatomy and functional signal from [Pinho et al., 2018]

We reprocessed the datasets using standard tools from the SCANPY pipeline [Wolf et al., 2018]. Specifically, we log-normalized gene expression of all genes present in dataset. We selected two brain coronal sections for **Exp.1** and two Coronal Olfactory Bulb (OB) sections for **Exp.2**, from the STARmap dataset. For **Exp.3**, we used the meshes together with their functional signal of the brains to recapitulate **Exp.1** in [Thual et al., 2022]. A visualization of the STARmap dataset for the two subsets used in **Exp.1** and **Exp.2** can be seen in Figure 5 and an overview of the cell type proportions present in each of the section pairs can be see in Figure 6. These visualization highlight the differences in terms of spatial organization and cell type proportions of the brain sections used in the experiment.

B.2 Experimental settings

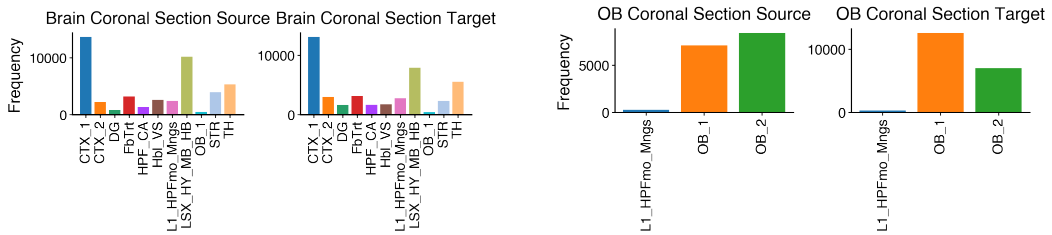
For **FUGW-sparse** presented in Table 3, we compute the coupling in 2 stages: (i) similarly as in Thual et al. [2022], we subsample the mesh to 10% of the points using Ward's algorithm and compute



(a) Visualization of the brain coronal sections used in **Exp.1**.

(b) Visualization of the OB sections used in **Exp.2**.

Figure 5: Spatial visualization of the two mouse brain sections used in **Exp.1**.



(a) Visualization of cell type frequencies for the brain coronal sections used in **Exp.1**.

(b) Visualization of cell type frequencies for the OB sections used in **Exp.2**.

Figure 6: Cell type frequencies of the datasets used in **Exp.1** and **Exp.2**.

the coarse optimal transport coupling. And (ii) we then use this coarse coupling to define a sparsity mask on the full mesh by selecting for each source (target) vertex the most coupled target (source) vertex and its neighbors within $\frac{4}{max\_distance}$ radius using the approximation of the geodesic distances. This mask is then used to compute the fine-grained sparse coupling.

For all experiments, we ran the grid search as defined by 4 and selected the best set of hyperparameters based on the validation correlation. We report results of top performing hyperparameters for the evaluated algorithms in Table 5 for **Exp.1**, Table 6 for **Exp.2** and Table 7 for **Exp.3**

	values
rank	10, 50, 100
reg (ours)	0.0, 0.001, 0.01
reg (fugw-sparse)	0.0001, 0.001, 0.01
tau1	0.1, 1.0, 100.0
tau2	0.1, 1.0, 100.0

Table 4: Hyperparameters considered in our grid-search.

solver	**rank**	**tau1**	**tau2**	**temp**	**reg**	**mass**	**val ρ**	**test ρ**	**F1-mac**	**F1-mic**	**F1-wei**
lot	10	-	-	0.200	0.010	1.000	0.282	0.386	0.210	0.411	0.360
ulot	10	1.000	1.000	0.200	0.010	0.889	0.301	0.409	0.200	0.425	0.363
lgw	100	-	-	0.200	0.001	1.000	0.227	0.288	0.487	0.716	0.692
ulgw	100	100.000	100.000	0.200	0.010	1.001	0.222	0.287	0.463	0.701	0.665
lfgw	50	-	-	0.400	0.010	1.000	0.365	0.443	0.576	0.720	0.714
ulfgw	100	0.100	0.100	0.400	0.001	0.443	0.379	0.463	0.582	0.733	0.724

Table 5: Results on the large spatial transcriptomics dataset (brain coronal section from [Shi et al., 2022]).

solver	rank	tau1	tau2	temp	reg	mass	val ρ	test ρ	F1-mac	F1-mic	F1-wei
uot	-	0.909	0.999	0.400	0.100	1.012	0.368	0.479	0.511	0.763	0.751
lot	10	-	-	0.200	0.010	1.000	0.335	0.440	0.511	0.760	0.751
ulot	10	1.000	100.000	0.200	0.010	0.998	0.356	0.461	0.518	0.770	0.762
ufgw	-	0.500	0.999	0.600	0.100	1.015	0.343	0.475	0.564	0.839	0.831
lfgw	10	-	-	0.600	0.010	1.000	0.348	0.453	0.512	0.762	0.753
ulfgw	10	0.100	0.100	0.600	0.001	0.339	0.368	0.491	0.556	0.826	0.818

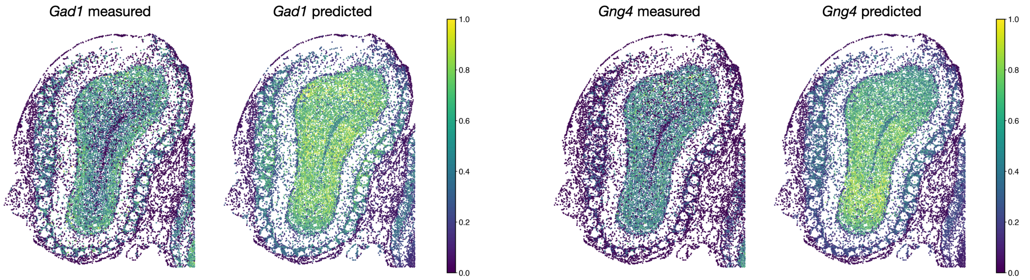
Table 6: Results on the small subset STARmap dataset (OB section from [Shi et al., 2022]).

solver	rank	tau1	tau2	reg	reg	mass	val ρ	test ρ
fugw-sparse	-	1.000	0.100	0.200	0.01	0.999	0.492	0.472
lfgw	100	-	-	0.600	0.000	1.000	0.513	0.663
ulfgw	100	1.000	0.100	0.600	0.001	0.981	0.533	0.643

Table 7: Results on the brain anatomy and functional signal from [Pinho et al., 2018]).

experiment		val ρ	tst ρ	F1-mac	F1-mic	F1-wei
Exp. 1	mean	0.362	0.449	0.546	0.687	0.677
	std	0.027	0.022	0.054	0.062	0.061
Exp. 2	mean	0.356	0.463	0.538	0.800	0.791
	std	0.008	0.018	0.021	0.031	0.032

Table 8: Effect of k-means initialization [Scetbon and Cuturi, 2022]. We report mean and standard deviation of *test* criterion for ULFGW, with the best hyperparameter on validation data for each experiment. We use 5 initial seeds for **Exp. 1**. We observe more variability in validation performance for **Exp. 2**, and therefore start with 10 seeds, pruning the lowest performing 5 seeds.



(a) Visualization of measured and predicted gene expression of *Gad1*.

(b) Visualization of measured and predicted gene expression of *Gnrg4*.

Figure 7: Measured and predicted gene expression for the small subset STARmap dataset (OB section from [Shi et al., 2022]) for ULRFGW.

B.3 Additional Experiments on the TI procedure

Here, we provide additional experiments in order to measure the effect of the TI version on the computational performance of LR solvers.

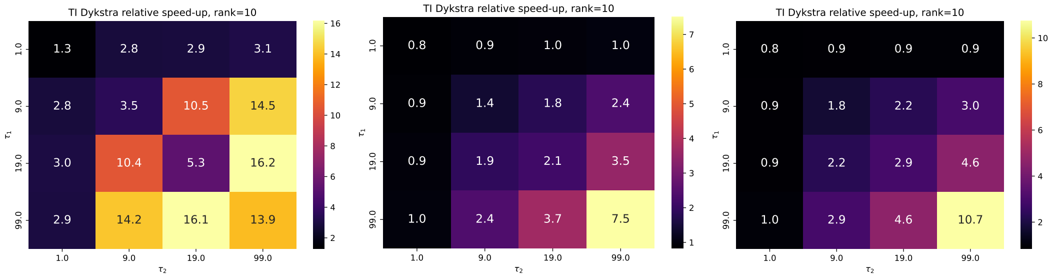


Figure 8: Speed-up of TI variant when varying τ_1, τ_2. *(left)* ULR **Sinkhorn** for $n = 25k$ points in 30d, rank=10, as in Figure 3. *(middle)* ULR-**GW** for $n = 50k$ points in src/tgt in 30d / 40d, means -1.2 / 1.3, std 1/0.2 between Gaussians.*(right)* ULR-**GW** as in middle, but data comes from **GMMs** (sklearn's blobs) with 10/15 clusters.

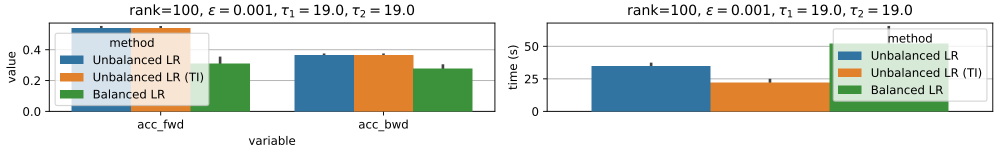


Figure 9: We used the Lobby room from Stanford 3D Indoor Scene Dataset (S3DIS) [Armeni et al., 2016] that consists of 1M points. *(left)* source-to-target, and target-to-source accuracies on the scene data, in a pure GW setting of balanced, unbalanced and unbalanced (TI variant). We use random initializer for all, 150 max outer iterations. The backward accuracy is comparable to what is mentioned in `https://2021.ecmlpkdd.org/wp-content/uploads/2021/07/sub_949.pdf` (≈ 0.41). *(right)* same as left, but showing timings, also comparable to those mentioned in the Quantized GW paper [Chowdhury et al., 2021] (10 min.)

C Proofs

C.1 Proof of Proposition 1

Let $n, m \geq r \geq 1, \gamma > 0, , \boldsymbol{\xi} := (\xi^{(1)}, \xi^{(2)}, \xi^{(3)})$ where $\xi^{(1)} \in \mathbb{R}_+^{n\times r}$, $\xi^{(2)} \in \mathbb{R}_+^{m\times r}$ and $\xi^{(3)} \in \mathbb{R}_+^r$ and let us recall that $\mathrm{KL}(\cdot, \cdot)$ is the generalized Kullback-Leibler divergence defined as $\mathrm{KL}(p|q) := \sum_i p_i \log(p_i/q_i) + q_i - p_i$. Then observe that

$$\min_{(Q,R,g)\in\Pi_r} \frac{1}{\gamma}\left[\mathrm{KL}(Q,\xi^{(1)}) + \mathrm{KL}(R,\xi^{(2)}) + \mathrm{KL}(g,\xi^{(3)})\right] + \tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2 \mathrm{KL}(R\mathbf{1}_r|b)$$

is a convex problem satisfying the Slater's condition and therefore strong duality holds. Therefore we have:

$$\begin{aligned}
&\min_{(Q,R,g)\in\Pi_r} \frac{1}{\gamma}\left[\mathrm{KL}(Q,\xi^{(1)}) + \mathrm{KL}(R,\xi^{(2)}) + \mathrm{KL}(g,\xi^{(3)})\right] + \tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2 \mathrm{KL}(R\mathbf{1}_r|b)\\
&= \sup_{\lambda_1,\lambda_2} \min_{Q,R,g} \langle \lambda_1, g - Q^\top \mathbf{1}_n\rangle + \langle \lambda_2, g - R^\top \mathbf{1}_m\rangle + \frac{1}{\gamma}\left[\mathrm{KL}(Q,\xi^{(1)}) + \mathrm{KL}(R,\xi^{(2)}) + \mathrm{KL}(g,\xi^{(3)})\right]\\
&+ \tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2 \mathrm{KL}(R\mathbf{1}_r|b)\\
&= \sup_{\lambda_1,\lambda_2} \min_{Q} \frac{1}{\gamma}\mathrm{KL}(Q,\xi^{(1)}) + \tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \langle \lambda_1, -Q^\top \mathbf{1}_n\rangle\\
&+ \min_R \frac{1}{\gamma}\mathrm{KL}(R,\xi^{(2)}) + \tau_2 \mathrm{KL}(R\mathbf{1}_r|b) + \langle -\lambda_2, R^\top \mathbf{1}_m\rangle + \min_g \frac{1}{\gamma}\mathrm{KL}(g,\xi^{(3)}) + \langle g, \lambda_1 + \lambda_2\rangle.
\end{aligned}$$

Now consider

$$\min_g \frac{1}{\gamma}\mathrm{KL}(g,\xi^{(3)}) + \langle g, \lambda_1 + \lambda_2\rangle$$

and observe that this problem can be solved explicitly. The first-order optimality condition gives us that $g^* = \exp(-\gamma(\lambda_1 + \lambda_2)) \odot \xi^{(3)}$ solves the problem and

$$\min_g \frac{1}{\gamma}\mathrm{KL}(g,\xi^{(3)}) + \langle g, \lambda_1 + \lambda_2\rangle = -\frac{1}{\gamma}\langle \exp(-\gamma(\lambda_1 + \lambda_2)), \xi^{(3)}\rangle + \langle \xi^{(3)}, \mathbf{1}\rangle.$$

Let us now focus on the following convex optimization problem,

$$\min_{Q} \frac{1}{\gamma}\mathrm{KL}(Q, \xi^{(1)}) + \tau_1 \mathrm{KL}(Q\mathbf{1}_r|a) + \langle -\lambda_1, Q^\top \mathbf{1}_n\rangle \tag{18}$$

and note that it admits a unique solution due to the strict convexity of $Q \to \mathrm{KL}(Q, \xi^{(1)})$. Then by denoting $F_{\tau,z}(s) := \tau\mathrm{KL}(s|z)$ and $G_\lambda(s) := \langle s, -\lambda\rangle$, and by applying the Fenchel-Rockafellar theorem [Rockafellar, 1970], we obtain that strong duality holds, the dual problem of (18) is

$$\sup_{f_1,h_1} -F^*_{\tau_1,a}(-f_1) - G^*_{\lambda_1}(-h_1) - \frac{1}{\gamma}\langle \exp(\gamma(f_1+h_1)), \xi^{(1)}\rangle$$

and that (f_1, h_1) solves the dual if and only if $-f_1 \in \partial F_{\tau_1,a}(Q\mathbf{1}_r)$,$-h_1 \in \partial G_{\lambda_1}(Q^\top\mathbf{1}_n)$ and $Q = \mathrm{diag}(\exp(\gamma f_1))\xi^{(1)}\,\mathrm{diag}(\exp(\gamma h_1))$ where Q is the solution of (18). Recall that here we denote for any convex set $X \in \mathbb{R}^q$ and function $f : X \to \mathbb{R}\cup\{+\infty\}$, f^* its convex conjugate defined for any $y \in X^* := \{x^* \text{ s.t. } \sup_{x\in X}\langle x, x^*\rangle - f(x) < +\infty\}$ by $f^*(y) := \sup_{x\in X}\langle x, y\rangle - f(x)$ and $\partial f(x) := \{y \text{ s.t. } f(x') - f(x) \geq \langle y, x - x'\rangle\ \forall x' \in X\}$. Now remarks that

$$G^*_{\lambda_1}(-h_1) = \sup_s\langle s, \lambda_1 - h_1\rangle = \begin{cases} +\infty & \text{if } \lambda_1 \neq h_1 \\ 0 & \text{otherwise .}\end{cases}$$

therefore $G^*_{\lambda_1}$ ensures that $\lambda_1 = h_1$. Similarly we obtain that

$$\min_{R} \frac{1}{\gamma}\mathrm{KL}(R, \xi^{(2)}) + \tau_2 \mathrm{KL}(r\mathbf{1}_r|b) + \langle -\lambda_2, R^\top \mathbf{1}_m\rangle \tag{19}$$

is equal to its dual defined as

$$\sup_{f_2,h_2} -F^*_{\tau_2,b}(-f_2) - G^*_{\lambda_2}(-h_2) - \frac{1}{\gamma}\langle \exp(\gamma(f_2+h_2)), \xi^{(2)}\rangle$$

where again

$$G^*_{\lambda_2}(-h_2) = \begin{cases} +\infty & \text{if } \lambda_2 \neq h_2 \\ 0 & \text{otherwise}\end{cases}$$

and with the primal-dual relationship $R = \mathrm{diag}(\exp(\gamma f_2))\xi^{(2)}\,\mathrm{diag}(\exp(\gamma h_2))$ such that $-f_2 \in \partial F_{\tau_2,b}(R\mathbf{1}_r)$,$-h_2 \in \partial G_{\lambda_2}(R^\top\mathbf{1}_m)$. Finally the dual can be written as

$$\begin{aligned}
&\sup_{\lambda_1,\lambda_2}\sup_{f_1,h_1} -F^*_{\tau_1,a}(-f_1) - G^*_{\lambda_1}(-h_1) - \frac{1}{\gamma}\langle \exp(\gamma(f_1+h_1)), \xi^{(1)}\rangle \\
&+ \sup_{f_2,h_2} -F^*_{\tau_2,b}(-f_2) - G^*_{\lambda_2}(-h_2) - \frac{1}{\gamma}\langle \exp(\gamma(f_2+h_2)), \xi^{(2)}\rangle \\
&- \frac{1}{\gamma}\langle \exp(-\gamma(\lambda_1+\lambda_2)), \xi^{(3)}\rangle + \langle \xi^{(3)}, \mathbf{1}\rangle
\end{aligned}$$

and using the definition of $G^*_{\lambda_1}(-h_1)$ and $G^*_{\lambda_2}(-h_2)$, we obtain the desired dual up to an additive constant ($\langle\xi^{(3)}, \mathbf{1}\rangle$) which does not affect the solution of the problem and conclude the proof.

C.2 On the Iterations of the Dykstra's Algorithm

Recall that we propose to consider an alternate maximization scheme to solve (8). Starting from $h_1^{(0)} = h_2^{(0)} = \mathbf{0}_r$, we apply for $\ell \geq 0$ the following updates (dropping iteration number k in (7) for simplicity):

$$\begin{aligned}
f_1^{(\ell+1)} := \arg\sup_z \mathcal{D}(z, h_1^{(\ell)}, f_2^{(\ell)}, h_2^{(\ell)}),\ f_2^{(\ell+1)} := \arg\sup_z \mathcal{D}(f_1^{(\ell+1)}, h_1^{(\ell)}, z, h_2^{(\ell)}),\\
(h_1^{(\ell+1)}, h_2^{(\ell+1)}) := \arg\sup_{z_1,z_2} \mathcal{D}(f_1^{(\ell+1)}, z_1, f_2^{(\ell+1)}, z_2).
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{D}(f_1, h_1, f_2, h_2) = -F^*_{\tau_1,a}(-f_1) - \frac{1}{\gamma}\langle e^{\gamma(f_1\oplus h_1)} - 1, \xi^{(1)}\rangle - F^*_{\tau_2,b}(-f_2) - \frac{1}{\gamma}\langle e^{\gamma(f_2\oplus h_2)} - 1, \xi^{(2)}\rangle \\
- \frac{1}{\gamma}\langle e^{-\gamma(h_1+h_2)} - 1, \xi^{(3)}\rangle.
\end{aligned}$$

Let us consider the first update of the scheme that consists in solving

$$f_1^{(\ell+1)} := \arg\sup_z \mathcal{D}(z, h_1^{(\ell)}, f_2^{(\ell)}, h_2^{(\ell)})$$

To solve this problem, we again apply the Fenchel-Rockafellar theorem [Rockafellar, 1970] and obtain that

$$\sup_{f_1} -F^*_{\tau_1,a}(-f_1) - \frac{1}{\gamma}\langle \exp(\gamma(f_1 + h_1)), \xi^{(1)}\rangle = \min_s F_{\tau_1,a}(s) + \frac{1}{\gamma}\mathrm{KL}(s|\xi^{(1)}\exp(\gamma h_1))$$

and the optimality condition gives that f_1^* is solution of the LHS if and only if s^* solves the RHS and belongs to the subdifferential of $f_1 \to \exp(\gamma(f_1 + h_1)), \xi^{(1)}\rangle$ at f_1^*, that is $s^* = \exp(\gamma f_1^*) \odot \xi^{(1)}\exp(\gamma h_1)$. However the RHS problem can can be solved exactly and one obtained that $s^* = a^{(\tau_1/(1/\gamma+\tau_1))} \odot \xi^{(1)}\exp(\gamma h_1)^{(1/(1/1+\gamma\tau_1))}$, therefore when combined with the previous equation on s^* we obtain that

$$\exp(\gamma f_1^*) = \frac{s^*}{\xi^{(1)}\exp(\gamma h_1)} = \left(\frac{a}{\xi^{(1)}\exp(\gamma h_1)}\right)^{\frac{\tau_1}{1/\gamma+\tau_1}},$$

Similarly, the solution of $\arg\sup_z \mathcal{D}(f_1, h_1, z, h_2)$ is

$$\exp(\gamma f_2^*) = \left(\frac{b}{\xi^{(2)}\exp(\gamma h_2)}\right)^{\frac{\tau_2}{1/\gamma+\tau_2}}.$$

Let us now consider the following optimization problem corresponding to the last update if the alternate maximization scheme, that is

$$(h_1^{(\ell+1)}, h_2^{(\ell+1)}) := \arg\sup_{z_1,z_2} \mathcal{D}(f_1^{(\ell+1)}, z_1, f_2^{(\ell+1)}, z_2).$$

In fact this problem can be solved directly using simply the first-order condition of optimality that gives the two following equations:

$$\exp(\gamma h_1) \odot (\xi^{(1)})^\top \exp(\gamma f_1) - \exp(-\gamma h_1) \odot (\xi^{(3)}) \odot \exp(-\gamma h_2) = 0 \quad \text{and}$$
$$\exp(\gamma h_2) \odot (\xi^{(2)})^\top \exp(\gamma f_2) - \exp(-\gamma h_2) \odot (\xi^{3)}) \odot \exp(-\gamma h_1) = 0$$

leading to

$$g = (\xi^{3)} \odot (\xi^{(1)})^\top \exp(\gamma f_1) \odot (\xi^{(2)})^\top \exp(\gamma f_2))^{1/3}$$

and

$$\exp(\gamma h_1) = \frac{g}{(\xi^{(1)})^\top \exp(\gamma f_1)}, \quad \exp(\gamma h_2) = \frac{g}{(\xi^{(2)})^\top \exp(\gamma f_2)}.$$

C.3 Proof of Proposition 2

Let us consider the following optimization problem

$$\mathcal{D}_{\text{TI}}(\tilde{f}_1, \tilde{h}_1, \tilde{f}_2, \tilde{h}_2) := \sup_{\lambda_1,\lambda_2\in\mathbb{R}} \mathcal{D}(\tilde{f}_1 + \lambda_1, \tilde{h}_1 - \lambda_1, \tilde{f}_2 + \lambda_2, \tilde{h}_2 - \lambda_2)$$

Therefore we have

$$\sup_{\lambda_1,\lambda_2\in\mathbb{R}} \mathcal{D}(\tilde{f}_1 + \lambda_1, \tilde{h}_1 - \lambda_1, \tilde{f}_2 + \lambda_2, \tilde{h}_2 - \lambda_2)$$
$$= -F^*_{\tau_1,a}(-(\tilde{f}_1 + \lambda_1)) - F^*_{\tau_2,b}(-(\tilde{f}_2 + \lambda_2)) - \frac{1}{\gamma}\langle e^{-\gamma(\tilde{h}_1+\tilde{h}_2)} \odot e^{\gamma(\lambda_1+\lambda_2)}, \xi^{(3)}\rangle + C$$

where C does not depends on λ_1 and λ_2. Now observe that

$$F^*_{\tau_1,a}(s) = \sup_x \langle x, s\rangle - \tau_1 \mathrm{KL}(s|a)$$

and by applying the first-order optimality condition, we obtain that $x^* = \exp(s/\tau_1) \odot a$ solves the above optimization problem and

$$F^*_{\tau_1,a}(s) = \tau_1 \langle \exp(s/\tau_1), a\rangle.$$

Similarly,

$$F^*_{\tau_2,b}(s) = \tau_2 \langle \exp(s/\tau_2), b\rangle,$$

Then by appyling the first-order optimality condition we obtain the two following equations

$$\exp(-\lambda_1/\tau_1)\langle \exp(-\tilde{f}_1/\tau_1), a\rangle - \exp(\gamma\lambda_1)\langle \exp(\gamma\lambda_2), \xi^{(3)} \odot \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2))\rangle = 0 \quad \text{and}$$
$$\exp(-\lambda_2/\tau_2)\langle \exp(-\tilde{f}_2/\tau_2), b\rangle - \exp(\gamma\lambda_2)\langle \exp(\gamma\lambda_1), \xi^{(3)} \odot \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2))\rangle = 0.$$

which is equivalent to

$$\exp\left(\lambda_1 \frac{1/\gamma + \tau_1}{\tau_1/\gamma}\right) = \frac{\langle \exp(-\tilde{f}_1/\tau_1), a\rangle}{\langle \xi^{(3)}, \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2))\rangle} \exp(-\gamma\lambda_2) \quad \text{and}$$
$$\exp\left(\lambda_2 \frac{1/\gamma + \tau_2}{\tau_2/\gamma}\right) = \frac{\langle \exp(-\tilde{f}_2/\tau_2), b\rangle}{\langle \xi^{(3)}, \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2))\rangle} \exp(-\gamma\lambda_1)$$

Then applying $\log$ to the system, we obtain that

$$\lambda_1 \gamma \frac{1/\gamma + \tau_1}{\tau_1} = c_1 - \gamma\lambda_2 \quad \text{and}$$
$$\lambda_2 \gamma \frac{1/\gamma + \tau_2}{\tau_2} = c_2 - \gamma\lambda_1$$

where

$$c_1 := \log\left(\frac{\langle \exp(-\tilde{f}_1/\tau_1), a\rangle}{\langle \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2)), \xi^{(3)}\rangle}\right), \quad \text{and} \quad c_2 := \log\left(\frac{\langle \exp(-\tilde{f}_2/\tau_2), a\rangle}{\langle \exp(-\gamma(\tilde{h}_1 + \tilde{h}_2)), \xi^{(3)}\rangle}\right).$$

Finally we obtain a simple linear system and the solution follows.

C.4 Double Regularizations: Low-rank Structure and Entropy

Our proposed procedure can be easily extended to the case where one wants to add entropy in addition to the low-rank constraint to solve unbalanced low-rank and entropic optimal transport problems. More precisely, let us consider the general case where one aims at solving for any $\varepsilon > 0$

$$\mathrm{ULOT}_{r,\varepsilon}(\mu,\nu) := \min_{(Q,R,g)\in\Pi_r} \underbrace{\langle C, Q\,\mathrm{diag}(1/g)R^T\rangle}_{\mathcal{L}_C(Q,R,g)} + \underbrace{\tau_1\mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2\mathrm{KL}(R\mathbf{1}_r|b) - \varepsilon H(Q,R,g)}_{\mathcal{G}_{a,b,\varepsilon}(Q,R,g)} \tag{20}$$

where $H(Q,R,g) = H(Q) + H(R) + H(g)$ and $H(p) := -\sum_i p_i(\log(p_i) - 1)$. Note that here, compared to (6), we have simply add en entropic term to the objective to smooth the matrices Q, R and the barycenter g. To solve this problem, we propose to consider the exact same strategy as the one proposed to solve (6) where we slightly modify $\mathcal{G}_{a,b,\varepsilon}$ and explicitely show the dependency w.r.t. ε. Now by applying the linearzation step of $\mathcal{L}_C(Q,R,g)$, we now aim to solve at iteration k the following optimization problem:

$$(Q_{k+1}, R_{k+1}, g_{k+1}) := \operatorname*{argmin}_{\boldsymbol{\zeta}\in\Pi_r} \frac{1}{\gamma_k}\mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_k) + \varepsilon H(\boldsymbol{\zeta}) + \tau_1\mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2\mathrm{KL}(R\mathbf{1}_r|b) \tag{21}$$

In fact, this problem can be reformulated as a problem of the form (14) where we simply have to modify $\boldsymbol{\xi}_k$ and γ. Indeed observe that we have

$$\frac{1}{\gamma}\mathrm{KL}(Q|\xi^{(1)}) - \varepsilon H(Q) = \frac{1}{\gamma_\varepsilon}\mathrm{KL}(Q|\xi^{(1)}_\varepsilon)$$

where $\gamma_\varepsilon = \frac{1}{1/\gamma + \varepsilon}$ and $\xi^{(1)}_\varepsilon := (\xi^{(1)})^{\gamma_\varepsilon/\gamma}$. Therefore we obtain that

$$\operatorname*{argmin}_{\boldsymbol{\zeta}\in\Pi_r} \frac{1}{\gamma}\mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}) + \varepsilon H(\boldsymbol{\zeta}) + \tau_1\mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2\mathrm{KL}(R\mathbf{1}_r|b)$$
$$= \operatorname*{argmin}_{\boldsymbol{\zeta}\in\Pi_r} \frac{1}{\gamma_\varepsilon}\mathrm{KL}(\boldsymbol{\zeta}, \boldsymbol{\xi}_\varepsilon) + \tau_1\mathrm{KL}(Q\mathbf{1}_r|a) + \tau_2\mathrm{KL}(R\mathbf{1}_r|b)$$

where $\boldsymbol{\xi}_\varepsilon := (\xi_\varepsilon^{(1)}, \xi_\varepsilon^{(2)}, \xi_\varepsilon^{(3)})$. Therefore the entropic version of our problem can be solved using the exact same solver as the one proposed in the main paper where only simple updates of the gradient-step γ and the kernels $\boldsymbol{\xi}$ are required at each iteration. We summarize the proposed algorithm below.

Algorithm 7 $\mathrm{ULOT}_\varepsilon(C, a, b, r, \gamma_0, \tau_1, \tau_2, \delta)$

```
**Inputs:** $C, a, b, \varepsilon, \gamma_0, \tau_1, \tau_2, \delta$
$Q, R, g \leftarrow$ Initialization as proposed in [Scetbon and Cuturi, 2022]
**repeat**
    $\tilde{Q} = Q,\ \ \tilde{R} = R,\ \ \tilde{g} = g,$
    $\nabla_Q = CR\,\mathrm{diag}(1/g),\ \ \nabla_R = C^\top Q\,\mathrm{diag}(1/g),$
    $\omega \leftarrow \mathcal{D}(Q^T CR),\ \ \nabla_g = -\omega/g^2,$
    $\gamma \leftarrow \gamma_0 / \max(\|\nabla_Q\|_\infty^2, \|\nabla_R\|_\infty^2, \|\nabla_g\|_\infty^2),$
    $\gamma \leftarrow \frac{1}{1/\gamma + \varepsilon}$
    $\xi^{(1)} \leftarrow Q \odot \exp(-\gamma \nabla_Q),\ \xi^{(2)} \leftarrow R \odot \exp(-\gamma \nabla_R),\ \xi^{(3)} \leftarrow g \odot \exp(-\gamma \nabla_g),$
    $\xi^{(1)} \leftarrow (\xi^{(1)})^{\gamma_\varepsilon/\gamma},\ \xi^{(2)} \leftarrow (\xi^{(2)})^{\gamma_\varepsilon/\gamma},\ \xi^{(3)} \leftarrow (\xi^{(3)})^{\gamma_\varepsilon/\gamma},$
    $Q, R, g \leftarrow$ ULR-Dykstra$(a, b, \boldsymbol{\xi}, \gamma, \tau_1, \tau_2, \delta)$ (Alg. 5)
**until** $\Delta((Q, R, g), (\tilde{Q}, \tilde{R}, \tilde{g}), \gamma) < \delta$;
**Result:** $Q, R, g$
```