Unbalanced CO-Optimal Transport

Quang Huy Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, Ritambhara Singh

Unbalanced CO-Optimal Transport

Quang Huy Tran [1,2], **Hicham Janati** [3], **Nicolas Courty** [1], **Rémi Flamary** [2], **Ievgen Redko** [4], **Pinar Demetci** [5, 6], **Ritambhara Singh** [5, 6]

[1]Université Bretagne Sud, IRISA
[2]CMAP, Ecole Polytechnique, IP Paris
[3]LTCI, Télécom Paris, IP Paris
[4]Univ. Lyon, UJM-Saint-Etienne, CNRS, UMR 5516
[5]Center for Computational Molecular Biology, Brown University
[6]Department of Computer Science, Brown University
{quang-huy.tran, nicolas.courty}@univ-ubs.fr, hicham.janati@telecom-paris.fr,
remi.flamary@polytechnique.edu, ievgen.redko@univ-st-etienne.fr,
pinardemetci@gmail.com, ritambhara@brown.edu

Abstract

Optimal transport (OT) compares probability distributions by computing a meaningful alignment between their samples. CO-optimal transport (COOT) takes this comparison further by inferring an alignment between features as well. While this approach leads to better alignments and generalizes both OT and Gromov-Wasserstein distances, we provide a theoretical result showing that it is sensitive to outliers that are omnipresent in real-world data. This prompts us to propose unbalanced COOT for which we provably show its robustness to noise in the compared datasets. To the best of our knowledge, this is the first such result for OT methods in incomparable spaces. With this result in hand, we provide empirical evidence of this robustness for the challenging tasks of heterogeneous domain adaptation with and without varying proportions of classes and simultaneous alignment of samples and features across single-cell measurements.

Introduction

The last decade has witnessed many successful applications of optimal transport (OT) (Monge 1781; Kantorovich 1942) in machine learning, namely in domain adaptation (Courty et al. 2016), generative adversarial networks (Arjovsky, Chintala, and Bottou 2017), classification (Frogner et al. 2015), dictionary learning (Rolet, Cuturi, and Peyré 2016), semi-supervised learning (Solomon et al. 2014). When the supports of the probability measures lie in the same ground metric space, it is natural to use the distance defined by the metric to induce the cost, which leads to the famous Wasserstein distance (Villani 2003). When they do not, one can rely on the idea of Gromov-Hausdorff distance (Gromov 1981) and its equivalent reformulations (Gromov 1999; Kalton and Ostrovskii 1999; Burago, Burago, and Ivanov 2001), and adapt them to the setting of metric measure spaces (Gromov 1999). This results in, for example, the Gromov-Wasserstein (GW) distance (Mémoli 2007, 2011; Sturm 2012), which has been widely used in many applications, namely in shape matching (Mémoli 2011), comparing kernel matrices (Peyré, Cuturi, and Solomon 2016), graphs (Vayer et al. 2019; Xu et al. 2019; Xu, Luo, and Carin 2019), computational biology (Demetci et al. 2022), heterogeneous domain adaptation (Yan et al. 2018), correspondence alignment (Solomon et al. 2016), machine translation (Alvarez-Melis and Jaakkola 2018).

By construction, the GW distance can only provide the sample alignment that best preserves the intrinsic geometry of the distributions and, as such, compares square pairwise relationship matrices. The CO-Optimal transport (COOT) (Redko et al. 2020; Chowdhury et al. 2021) goes beyond these limits by simultaneously learning two independent (feature and sample) correspondences, and thus provides greater flexibility over the GW distance in terms of usage and interpretability. First, it allows us to measure similarity between arbitrary-size matrices. An interesting use case is, for instance, on tabular data, which are usually expressed as a matrix whose rows represent samples and columns represent features. For the GW distance, the similarity or distance matrix (or any square matrix derived from the data) must be calculated in advance and the effect of the individual variables is lost during this computation. On the other hand, COOT can bypass this step as it can use either the tabular data directly or the similarity matrices as inputs. Second, COOT provides both sample and feature correspondences. These feature correspondences are also interpretable and allow to recover relations between the features of two different datasets even when they do not lie in the same space.

Similar to classical OT, COOT enforces hard constraints on the marginal distributions both between samples and features. These constraints lead to two main limitations: (1) imbalanced datasets where samples or features are re-weighted cannot be accurately compared; (2) mass transportation *must* be exhaustive: outliers, if any, must be matched regardless of the cost they induce. To circumvent these limitations, we propose to relax the mass preservation constraints in the COOT distance and study a broadly applicable and general OT framework that includes several well-studied cases presented in Table 1.

Related work. To relax the OT marginal constraints, a straightforward solution is to control the difference between the marginal distributions of the transportation plan and the

data by some discrepancy measure, e.g., Kullback-Leibler divergence. In classical OT, this gives rise to the unbalanced OT (UOT), which was first proposed by (Benamou 2003). The theoretical and numerical aspects of this extension have been studied extensively (Liero, Mielke, and Savaré 2018; Chizat et al. 2018b,a; Pham et al. 2020) and are gaining increasing attention in the machine learning community, with wide-range applications, namely in domain adaptation (Fatras et al. 2021), generative adversarial networks (Balaji, Chellappa, and Feizi 2020; Yang and Uhler 2019), dynamic tracking (Lee, Bertrand, and Rozell 2020), crowd counting (Ma et al. 2021), neuroscience (Janati et al. 2019; Bazeille et al. 2019) or modeling cell developmental trajectories (Schiebinger et al. 2019). Unbalanced OT and its variants are usually sought for their known robustness to outliers (Mukherjee et al. 2021; Balaji, Chellappa, and Feizi 2020; Fatras et al. 2021). This appealing property goes beyond classical OT. For instance, to compare signed and non-negative measures in incomparable spaces, unbalanced OT (Liero, Mielke, and Savaré 2018) can be blended with the L_p-transportation distance (Sturm 2006), which leads to the Sturm-Entropic-Transport distance (Ponti and Mondino 2020), or with the GW distance, which gives rise to the unbalanced GW (UGW) distance (Séjourné, Vialard, and Peyré 2021). Also motivated by the unbalanced OT, (Zhang et al. 2021) proposed a relaxation of the bidirectional Gromov-Monge distance called unbalanced bidirectional Gromov-Monge divergence.

Contributions. In this work, we introduce an unbalanced extension of COOT called "Unbalanced CO-Optimal transport" (UCOOT). UCOOT – defined for both discrete and continuous data – is a general framework that encompasses all the OT variants displayed in Table 1. Our main contribution is to show that UCOOT is provably robust to both samples and features outliers, while its balanced counterpart can be made arbitrarily large with strong enough perturbations. To the best of our knowledge, this is the first time such a general robustness result is established for OT across different spaces. Our theoretical findings are showcased in unsupervised heterogeneous domain adaptation and single-cell multi-omic data alignment, demonstrating a very competitive performance.

Notations. For any integer $n \geq 1$, we write $[|1, n|] := \{1, ..., n\}$. Given a Polish space $\mathcal{X}$, we denote $\mathcal{M}^+(\mathcal{X})$ the set of nonnegative and finite Borel measures over $\mathcal{X}$. For any $\mu \in \mathcal{M}^+(\mathcal{X})$, we denote its mass by $m(\mu) := \mu(\mathcal{X})$. Unless specified otherwise, we always consider fully supported measures, i.e. $\text{supp}(\mu) = \mathcal{X}$, for any measure $\mu \in \mathcal{M}^+(\mathcal{X})$. The product measure of two measures μ and ν is defined as: $d(\mu \otimes \nu)(x, y) := d\mu(x)d\nu(y)$. Given $\pi \in \mathcal{M}^+(\mathcal{X} \times \mathcal{Y})$, we denote $(\pi_{\#1}, \pi_{\#2})$ its marginal distributions i.e. $d\pi_{\#1} = \int_{\mathcal{Y}} d\pi$ and $d\pi_{\#2} = \int_{\mathcal{X}} d\pi$. For $\mu, \nu \in \mathcal{M}^+(\mathcal{X})$, the Kullback-Leibler divergence is defined by $\text{KL}(\mu|\nu) \overset{\text{def}}{=} \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} \text{d}\nu - \int \text{d}\mu + \int \text{d}\nu$ if $\mu \ll \nu$ and set to $+\infty$ otherwise. Finally, the indicator divergence $\iota_=(\mu|\nu)$ is equal to 0 if $\mu = \nu$ and $+\infty$ otherwise.

Unbalanced CO-Optimal Transport (UCOOT)

The ultimate goal behind the CO-Optimal Transport (COOT) framework is the simultaneous alignment of samples *and* features to allow for comparisons across spaces of different dimensions. In this section, we discuss OT formulations including OT, UOT, GW, UGW and COOT, then introduce the proposed UCOOT and show how the aforementioned distances fall into our framework.

From sample alignment to sample-feature alignment. Let $(\mathcal{X}_1^s, \mu_1^s)$ and $(\mathcal{X}_2^s, \mu_2^s)$ be a pair of compact measure spaces such that $\mathcal{X}_1^s$ and $\mathcal{X}_2^s$ belong to some common metric space $(\mathcal{E}, d)$. Classical (unbalanced) optimal transport infers one alignment (or joint distribution) $\pi^s \in \mathcal{M}^+(\mathcal{X}_1^s \times \mathcal{X}_2^s)$ with marginals $(\pi^s_{\#1}, \pi^s_{\#2})$ close to (μ_1^s, μ_2^s) according to some appropriate divergence D such that the cost $\int c(x_1, x_2) \text{d}\pi^{(s)}(x_1, x_2) + D(\pi^s_{\#1} \| \mu_1^s) + D(\pi^s_{\#2} \| \mu_2^s)$ is minimal. For instance, in balanced (resp. unbalanced) OT, D corresponds to the indicator divergence (resp. KL divergence or TV). To define a generalized OT beyond one single alignment, we must first introduce a new pair of measure spaces $(\mathcal{X}_1^f, \mu_1^f)$ and $(\mathcal{X}_2^f, \mu_2^f)$. Intuitively, the two transport plans that must be inferred: π^s across *samples* and π^f across *features*, must minimize a cost of the form $\iint c((x_1^s, x_1^f), (x_2^s, x_2^f)) \text{d}\pi^s(x_1^s, x_2^s) \text{d}\pi^f(x_1^f, x_2^f)$ where $c((x_1^s, x_1^f), (x_2^s, x_2^f))$ is the *joint* cost of aligning the sample-feature pairs (x_1^s, x_1^f) and (x_2^s, x_2^f). However, unlike OT, there is no underlying ambient metric space in which comparisons between these pairs are straightforward. Thus, we consider a simplified cost of the form: $c((x_1^s, x_1^f), (x_2^s, x_2^f)) = |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p$, for $p \geq 1$ and some scalar functions ξ_1, ξ_2 that define the sample-feature interactions. A similar definition was adopted by (Chowdhury et al. 2021) to extend COOT to the continuous setting in the context of hypergraphs. Formally, our general formulation takes pairs of *sample-feature spaces* defined as follows.

Definition 1 (Sample-feature space)**.** *Let $(\mathcal{X}^s, \mu^s)$ and $(\mathcal{X}^f, \mu^f)$ be compact measure spaces, where $\mu^f \in \mathcal{M}^+(\mathcal{X}^f)$ and $\mu^s \in \mathcal{M}^+(\mathcal{X}^s)$. Let ξ be a scalar integrable function in $L^p(\mathcal{X}^s \times \mathcal{X}^f, \mu^s \otimes \mu^f)$. We call the triplet $\mathbb{X} = ((\mathcal{X}^s, \mu^s), (\mathcal{X}^f, \mu^f), \xi)$ a sample-feature space and ξ is called an interaction.*

Definition 2 (Generalized COOT)**.** *Given two divergences D_1 and D_2, we define the generalized COOT of order p between $\mathbb{X}_1 = ((\mathcal{X}_1^s, \mu_1^s), (\mathcal{X}_1^f, \mu_1^f), \xi_1)$ and $\mathbb{X}_2 = ((\mathcal{X}_2^s, \mu_2^s), (\mathcal{X}_2^f, \mu_2^f), \xi_2)$ by:*

$$\inf_{\substack{\pi^s \in \mathcal{M}^+(\mathcal{X}_1^s \times \mathcal{X}_2^s) \\ \pi^f \in \mathcal{M}^+(\mathcal{X}_1^f \times \mathcal{X}_2^f) \\ m(\pi^s) = m(\pi^f)}} \underbrace{\iint |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p \text{d}\pi^s \text{d}\pi^f}_{\textit{transport cost of sample-feature pairs}} + \underbrace{\sum_{k=1}^{2} \lambda_k D_k(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu_k^s \otimes \mu_k^f)}_{\textit{mass destruction / creation penalty}}, \tag{1}$$

for $\lambda_1, \lambda_2 > 0$ and $p \geq 1$.

	OT	UOT	GW	UGW	COOT	UCOOT
Across spaces	✗	✗	✓	✓	✓	✓
Sample alignment	✓	✓	✓	✓	✓	✓
Feature alignment	✗	✗	✗	✗	✓	✓
Robustness to outliers	✗(Fatras et al. 2021)	✓(Fatras et al. 2021)	✗(Prop. 2)	✓(Thm. 1)	✗(Prop. 2)	✓(Thm. 1)

Table 1. Properties of different OT formulations generalized by UCOOT. The proposed UCOOT is not only able to learn informative feature alignments, but also robust to outliers.

As the multiplicative nature between π^s and π^f leads to an invariance by the scaling map $\alpha \mapsto (\alpha\pi^s, \frac{1}{\alpha}\pi^f)$, for $\alpha > 0$, we further impose the equal mass constraint $m(\pi^s) = m(\pi^f)$.

It is worth mentioning that the formulation 1 is not the only way to relax the marginal constraints. For example, instead of $D_k(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu^s_k \otimes \mu^f_k)$, one can consider $D_k(\pi^s_{\#k}|\mu^s_k) + D_k(\pi^f_{\#k}|\mu^f_k)$, or $D_s(\pi^s_{\#1} \otimes \pi^s_{\#2}|\mu^s_1 \otimes \mu^s_2)$, for some divergence D_s. However, amongst these choices, ours is the only one which can be recast as a variation of the unbalanced OT problem. This allows us to leverage the known techniques in unbalanced OT to justify the theoritical and practical properties, namely Proposition 1 and Theorem 1 below.

Note that the problem above is very general and can – with some additional constraints – recover exact OT, UOT, GW, UGW, COOT (see Table 2). In particular, if the measures (μ^s_1, μ^s_2) and (μ^f_1, μ^f_2) are probability measures, then setting $D_1 = D_2 = \iota_=$ leads to the COOT problem first introduced in the discrete case in (Redko et al. 2020) and recently generalized to the continuous setting in (Chowdhury et al. 2021). In this work, we relax the hard constraints and consider a more flexible formulation with the KL divergence:

Definition 3 (UCOOT)**.** *We define Unbalanced COOT (*UCOOT*) as in* (1) *with* $D_1 = D_2 =$ KL*. We write* $\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2)$ *to indicate the UCOOT between two sample-feature spaces* $\mathbb{X}_1$ *and* $\mathbb{X}_2$*, for a given pair of hyperparameters* $\lambda = (\lambda_1, \lambda_2)$*.*

While various properties of the divergences D_k have been extensively studied in the context of unbalanced OT by several authors (Chizat 2017; Frogner et al. 2015), the concept of sample-feature interaction requires more clarification. Let us consider some simple examples. In the discrete case, we consider n observations of d features represented by matrix $\mathbf{A} \in \mathbb{R}^{n,d}$. In this case, the space $\mathcal{X}^s$ (resp. $\mathcal{X}^f$) is not explicitly known but can be characterized by the finite set $[|1, n|]$ (resp. $[|1, d|]$), up to an isomorphism. Assuming that all samples (resp. features) are equally important, the discrete empirical measures can be given by uniform weights $\mu^s = \frac{\mathbb{1}_n}{n}$ (resp. $\mu^f = \frac{\mathbb{1}_d}{d}$). The most natural sample-feature interaction ξ is simply the index function $\xi(i, j) = \mathbf{A}_{ij}$. In the continuous case, we assume that data stream from a continuous random variable $\mathbf{a} \sim \mu_s \in \mathcal{P}(\mathbb{R}^d)$ for which an interaction function can be $\xi(\mathbf{a}, j) = \mathbf{a}_j$.

Proposition 1. *For any* $D_1, D_2 \in \{\iota_=, \mathrm{KL}\}$*, Problem 2 (in Equation 1) admits a minimizer.*

Remark 1. The existence of minimizer shown in Proposition 1 can be extended to a larger family of Csiszár divergences (Csiszár 1963). A general proof is given in the Appendix.

UCOOT and perfect alignment. Suppose that $\mathbb{X}_1$ and $\mathbb{X}_2$ are two finite sample-feature spaces such that $(\mathcal{X}^s_1, \mathcal{X}^s_2)$ and $(\mathcal{X}^f_1, \mathcal{X}^f_2)$ have the same cardinality and are equipped with the uniform measures $\mu^s_1 = \mu^s_2$, $\mu^f_1 = \mu^f_2$. Then $\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$ if and only if there exist perfect alignments between rows (samples) and between columns (features) of the interaction matrices ξ_1 and ξ_2.

Robustness of COOT and UCOOT

When discussing the concept of robustness, outliers are often considered as samples not following the underlying distribution of the data. In our general context of sample-feature alignments, we consider a *pair* $(x^s, x^f) \in \mathcal{X}^s \times \mathcal{X}^f$ to be an outlier if the magnitude of $|\xi(x^s, x^f)|$ is abnormally larger than other interactions between $\mathcal{X}^s$ and $\mathcal{X}^f$. As a result, such outliers lead to abnormally large transportation costs $|\xi_1 - \xi_2|$. To study the robustness of COOT and UCOOT, we consider an outlier scenario where the marginal data distributions are contaminated by some additive noise distribution.

Assumption 1. *Consider two sample-feature spaces* $\mathbb{X}_k = ((\mathcal{X}^s_k, \mu^s_k), (\mathcal{X}^f_k, \mu^f_k), \xi_k)$*, for* $k = 1, 2$*. Let* ε^s *(resp.* ε^f*) be a probability measure with compact support* $\mathcal{O}^s$ *(resp.* $\mathcal{O}^f$*). For* $a \in \{s, f\}$*, define the noisy distribution* $\widetilde{\mu}^a = \alpha_a \mu^a + (1 - \alpha_a)\varepsilon^a$*, where* $\alpha_a \in [0, 1]$*. We assume that* ξ_1 *is defined on* $(\mathcal{X}^s_1 \cup \mathcal{O}^s) \times (\mathcal{X}^f_1 \cup \mathcal{O}^f)$ *and that* ξ_1, ξ_2 *are continuous on their supports. We denote the contaminated sample-feature space by* $\widetilde{\mathbb{X}_1} = ((\mathcal{X}^s_1 \cup \mathcal{O}^s, \widetilde{\mu}^s_1), (\mathcal{X}^f_1 \cup \mathcal{O}^f, \widetilde{\mu}^f_1), \xi_1)$*. Finally, we define some useful minimal and maximal costs:*

$$\begin{cases} \Delta_0 \overset{\text{def}}{=} \min_{\substack{x^s_1 \in \mathcal{O}^s, x^f_1 \in \mathcal{O}^f \\ x^s_2 \in \mathcal{X}^s_2, x^f_2 \in \mathcal{X}^f_2}} |\xi_1(x^s_1, x^f_1) - \xi_2(x^s_2, x^f_2)|^p \\ \Delta_\infty \overset{\text{def}}{=} \max_{\substack{x^s_1 \in \mathcal{X}^s_1 \cup \mathcal{O}^s, x^f_1 \in \mathcal{X}^f_1 \cup \mathcal{O}^f \\ x^s_2 \in \mathcal{X}^s_2, x^f_2 \in \mathcal{X}^f_2}} |\xi_1(x^s_1, x^f_1) - \xi_2(x^s_2, x^f_2)|^p \end{cases}.$$

Here, Δ_0 *accounts for the minimal deviation of the cost between the outliers and target support, while* Δ_∞ *is the maximal deviation between the contaminated source and the target.*

The exact marginal constraints of COOT enforce conservation of mass. Thus, outliers *must* be transported no matter how large their transportation costs are. This intuition is captured by the following result.

Requirements	OT	GW	COOT	semi-d. COOT	UCOOT
Shape of inputs	$d_1 = d_2$	$n_1 = d_1, n_2 = d_2$	–	–	–
Coupling constraint	$\pi^f = \mathrm{I}_{d_1} = \mathrm{I}_{d_2}$	$\pi^f = \pi^s$	–	–	–
Scalar function	$\xi(i,j) = \mathbf{A}_{ij}$	$\xi(i,j) = \mathrm{dist}(\mathbf{A}_{i.}, \mathbf{A}_{j.})$	$\xi(i,j) = \mathbf{A}_{ij}$	$\xi(\mathbf{a}, j) = \mathbf{a}_j$	$\xi(i,j) = \mathbf{A}_{ij}$
Divergence	$\iota_=$	$\iota_=$	$\iota_=$	$\iota_=$	KL

Table 2. Conditions under which different OT formulations fall within the generalized framework of Definition 2. "semi-d" refers to "semi-discrete" setting, where μ_s is a continuous probability and $\mu_d = \mathbb{1}_d/d$. Here, I_d is the identity matrix in $\mathbb{R}^d$.

Proposition 2. *(COOT is sensitive to outliers) Consider* $\widetilde{\mathbb{X}_1}, \mathbb{X}_2$ *as defined in Assumption 1. Then:*

$$\mathrm{COOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \geq (1-\alpha_s)(1-\alpha_f)\Delta_0. \qquad (2)$$

Whenever the outlier proportion $(1-\alpha_s)(1-\alpha_f)$ is positive, COOT increases with the distance between the supports of the outliers and those of the clean data. Thus, the right hand side of (2) can be made arbitrarily large by taking outliers far from the supports of the clean data.

We can now state our main theoretical contribution. Relaxing the marginal constraints leads to a loss that saturates as outliers get further from the data:

Theorem 1. *(UCOOT is robust to outliers) Consider two sample-feature spaces* $\widetilde{\mathbb{X}_1}, \mathbb{X}_2$ *as defined in Assumption 1. Let* $\delta \overset{\text{def}}{=} 2(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f)$ *and* $K = M + \frac{1}{M}\,\mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + \delta$*, where* $M = m(\pi^s) = m(\pi^f)$ *is the transported mass between clean data. Then:*

$$\mathrm{UCOOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \leq \alpha_s\alpha_f\,\mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + \delta M\left[1 - \exp\left(-\frac{\Delta_\infty(1+M)+K}{\delta M}\right)\right].$$

The proof of Theorem 1 is provided in the Appendix and inspired from (Fatras et al. 2021), but in a much more general setting: (1) it covers both sample and feature outliers and (2) considers a noise distribution instead of a Dirac. Note that the inequality (2) indicates that outliers can make COOT arbitrary large, while UCOOT is upper bounded and discards the mass of outliers with high transportation cost.

This is well illustrated in Figure 1, where we simulate outliers by adding a perturbation to a row of the interaction matrix. More precisely, we first generate a matrix $\mathbf{A} \in \mathbb{R}^{20,15}$ by $\mathbf{A}_{ij} = \cos(\frac{i}{20}\pi) + \cos(\frac{j}{15}\pi)$. Then, we replace its last row by $\tau\mathbb{1}_{15}$, for $\tau \geq 0$. Figure 1 depicts COOT and UCOOT between $\mathbf{A}$ and its modified version as a function of τ. The higher the value of τ, the more likely that the last row contains the interaction of outliers. Consequently, as τ increases, so does COOT but at a much higher pace, whereas UCOOT remains stable.

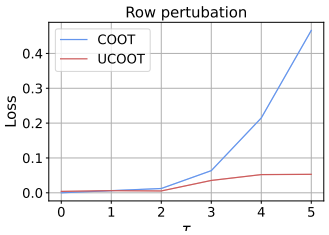

Fig. 1. Sensitivity of COOT and UCOOT under the presence of outliers.

It should be noted that, with minimal adaptation, theorem 1 also holds for the unbalanced GW (UGW) distance. This provides a theoretical explanation of the empirical observation in (Séjourné, Vialard, and Peyré 2021) that unlike GW, the UGW distance is also robust to outliers.

Numerical aspects

Solving COOT-type problems, in general, is not trivial. As highlighted in (Redko et al. 2020), the balanced case corresponds to a convex relaxation of the bilinear assignment problem, which seeks the pair of permutations minimizing the transport cost. Here we argue that relaxing the marginal constraints makes the problem easier in two different aspects: (1) the obtained problem is easier to solve through a sequence of GPU friendly iterations; (2) regularization leads to lower alignment costs and thus better local minima. In this section, we first describe how to compute UCOOT in practice.

Optimization strategy. We consider two tabular datasets $\mathbf{A} \in \mathbb{R}^{n_1,d_1}$ and $\mathbf{B} \in \mathbb{R}^{n_2,d_2}$. Let u_k be the uniform histogram over sample-feature pairs: $u_k \overset{\text{def}}{=} \frac{1}{n_k d_k}\mathbb{1}_{n_k} \otimes \mathbb{1}_{d_k}$, for $k = 1, 2$. For the sake of simplicity, we assume uniform weights over both samples and features. Computing UCOOT can be done using block-coordinate descent (BCD) both with and without entropy regularization. More precisely, given a hyperparameter $\varepsilon \geq 0$, discrete UCOOT can be written as:

$$\min_{\substack{\pi^s,\pi^f \\ m(\pi^s)=m(\pi^f)}} \sum_{i,j,k,l} (\mathbf{A}_{ik} - \mathbf{B}_{jl})^2 \pi^s_{ij}\pi^f_{kl} + \lambda_1\,\mathrm{KL}(\pi^s\mathbb{1}_{n_1} \otimes \pi^f\mathbb{1}_{d_1}|u_1)$$
$$+ \lambda_2\,\mathrm{KL}({\pi^s}^\top\mathbb{1}_{n_2} \otimes {\pi^f}^\top\mathbb{1}_{d_2}|u_2) + \varepsilon\mathrm{KL}(\pi^s \otimes \pi^f|\mu_1^s \otimes \mu_2^s \otimes \mu_1^f \otimes \mu_2^f).$$

Algorithm 1. BCD algorithm to solve UCOOT

Input: $\mathbf{A} \in \mathbb{R}^{n_1,d_1}, \mathbf{B} \in \mathbb{R}^{n_2,d_2}, \lambda_1, \lambda_2, \varepsilon$
Initialize π^s and π^f
repeat
  Update π^s using Sinkhorn or NNPR
  Rescale $\pi^s = \sqrt{\frac{m(\pi^f)}{m(\pi^s)}}\pi^s$
  Update π^s using Sinkhorn or NNPR
  Rescale $\pi^f = \sqrt{\frac{m(\pi^s)}{m(\pi^f)}}\pi^f$
until convergence

The only difference between $\varepsilon = 0$ and $\varepsilon > 0$ lies in the inner-loop algorithm used to update one of transport plans (π^s, π^f) while the other one remains fixed. Note that both cases allow for implementations of a scaling multiplicative algorithm that can be parallelized on GPUs. For $\varepsilon > 0$, updating each transport plan boils down to an entropic UOT prob-

lem, which can be solved efficiently using the unbalanced variant of Sinkhorn's algorithm (Chizat et al. 2018a). The main benefit of entropy regularization is to reduce the number of variables from $(n_1 \times n_2) + (d_1 \times d_2)$ to $n_1 + n_2 + d_1 + d_2$. Moreover, by taking ε sufficiently small, we can recover solutions close to those in the non-entropic case. For $\varepsilon = 0$, the non-regularized UOT problem can be recast as a non-negative penalized regression (NNPR) (Chapel et al. 2021). This problem can be solved using a majorization-minimization algorithm which leads to a multiplicative update on the transport plan. For the sake of reproducibility, we provide the details on the optimization scheme of both algorithms in the Appendix.

Experiments

Illustration and interpretation of UCOOT on MNIST images

We illustrate the robustness of UCOOT and its ability to learn meaningful feature alignments under the presence of both sample and feature outliers in the MNIST dataset. We introduce the feature outliers by applying a handcrafted transformation φ_σ that performs a zero-padding (shift), a 45° rotation, a resize to (28, 34) and adds Gaussian noise $\mathcal{N}(0, \sigma^2)$ entries to the first 10 columns of the image. Figure 3 (a) shows some examples of original and transformed images. We randomly sample 100 images per class (1000 total) from $X = \text{MNIST}$ and $Y = \varphi_\sigma(\text{MNIST})$. Regarding the sample outliers, we add 50 random images with uniform entries in [0, 1] to the target data Y. We then compute the optimal COOT and UCOOT alignments shown in Figure 3 (b) and (c). The flexibility of UCOOT with respect to mass transportation allows it to completely disregard: (1) noisy and uninformative pixels (features), which are all given the same weight as depicted by (b); (2) all the sample outliers of which none are transported as shown by the last blank column of the alignment (c). Moreover, notice how the color-coded input image is transformed according to the transformation φ_σ despite the fact that no spatial information is provided in the OT problem. On the other hand, a very small perturbation ($\sigma = 0.01$) is enough for the sample alignment given by COOT to lose its block-diagonal dominant structure (class information is lost), while the UCOOT alignment remains unscathed. One may wonder whether the performance of UCOOT would still hold for different values of σ. Figure 2 answers this question positively. For $\sigma > 0$, we compute the average accuracy (defined by the percentage of mass within the block-diagonal structure) over 20 different runs. The performance of COOT not only degrades with noisier outliers but is also unstable. By contrast, the accuracy of UCOOT remains almost constant regardless of the level of noise.

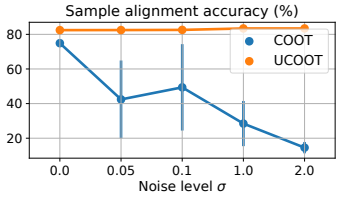

Fig. 2. Robustness of UCOOT vs. COOT on MNIST example. The accuracy is averaged on 20 trials.

CaffeNet → GoogleNet				
Domains	GW	UGW	COOT	UCOOT
C → C	16.25 (± 7.54)	10.85 (± 2.13)	36.40 (± 12.94)	**44.05 (± 19.33)**
C → A	12.95 (± 7.74)	11.60 (± 4.86)	28.30 (± 11.78)	**31.90 (± 7.43)**
C → W	18.95 (± 9.43)	14.15 (± 3.98)	19.55 (± 14.51)	**28.55 (± 6.60)**
A → C	16.40 (± 8.99)	10.25 (± 5.66)	**41.80 (± 14.81)**	39.15 (± 17.98)
A → A	14.75 (± 15.20)	20.20 (± 6.45)	**57.90 (± 16.84)**	42.45 (± 15.47)
A → W	14.55 (± 8.83)	20.65 (± 4.13)	42.10 (± 7.80)	**48.55 (± 13.06)**
W → C	20.65 (± 11.90)	14.20 (± 5.13)	8.60 (± 6.56)	**69.80 (± 14.91)**
W → A	17.00 (± 9.75)	7.10 (± 2.45)	16.65 (± 10.01)	**30.55 (± 10.09)**
W → W	19.30 (± 11.87)	24.40 (± 3.28)	**75.30 (± 3.26)**	51.50 (± 20.51)
Average	16.76 (± 10.14)	14.82 (± 4.23)	36.29 (± 10.95)	**42.94 (± 13.93)**

Table 3. Unsupervised HDA from CaffeNet to GoogleNet.

Heterogeneous Domain Adaptation (HDA)

We now investigate the application of discrete UCOOT in unsupervised Heterogeneous Domain Adaptation (HDA). It is a particularly difficult problem where one wants to predict classes on unlabeled data using labeled data lying in a different space. OT methods across spaces have recently shown good performance on such tasks, in particular using GW distance (Yan et al. 2018) and COOT (Redko et al. 2020).

Datasets and experimental setup. We consider the Caltech-Office dataset (Saenko et al. 2010) containing three domains: Amazon (A) (1123 images), Caltech-256 (C) (958 images) and Webcam (W) (295 images) with 10 overlapping classes amongst them. The image in each domain is representated by the output of the second last layer in the Google Net (Szegedy et al. 2015) and Caffe Net (Jia et al. 2014) neural network architectures, which results in 4096 and 1024-dimensional vectors, respectively (thus $d_s = 4096, d_t = 1024$). We compare 4 OT-based methods: GW, COOT, UGW, and UCOOT. The hyper-parameters for each method are validated on a unique pair of datasets (W→W), then fixed for all other pairs in order to provide truly unsupervised HDA generalization. We follow the same experimental setup as in (Redko et al. 2020). For each pair of domains, we randomly choose 20 samples per class (thus $n_s = n_t = 200$) and perform adaptation from CaffeNet to GoogleNet features, then calculate the accuracy of the generated predictions on the target domain using OT label propagation (Redko et al. 2019). This technique uses the OT plan to estimate the amount of mass transported from each class (since the sources are labeled) to a given target sample. The predicted class corresponds to the one which contains the most mass. We repeat this process 10 times and calculate the average and standard deviation of the performance. In both source and target domains, we assign uniform sample and feature distributions.

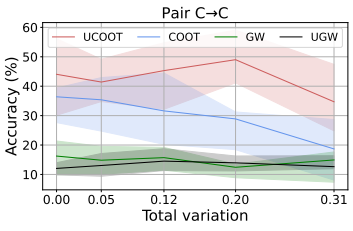

Fig. 4. Robustness to class proportion change for increasing TV on the class marginals.

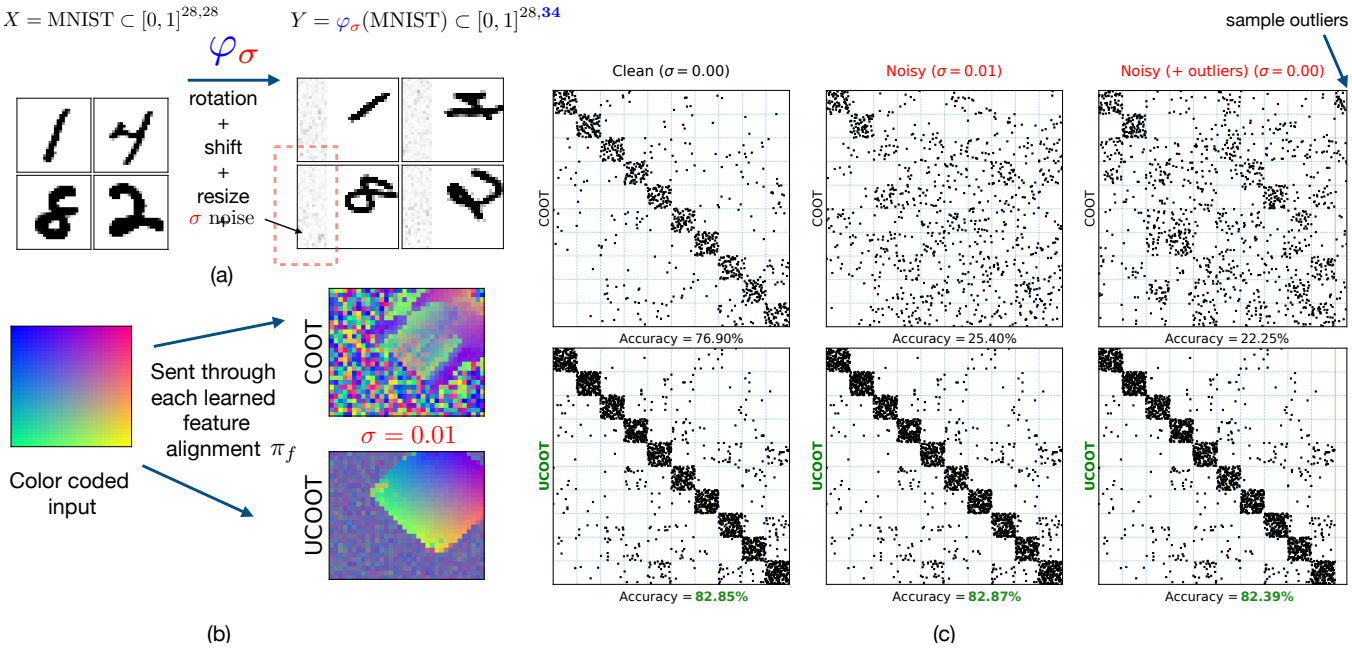

Fig. 3. Example illustrating the feature alignment π_f learned by UCOOT and its robustness to outliers. **(a)** Visualization of 4 random samples from both datasets. The added Gaussian noise only affects the first 10 columns of the images and is different across images. **(b)** The barycentric mapping (see Appendix for details) defined by UCOOT learns the transformation defined by φ_σ while disregarding non-informative features. **(c)** Alignments across samples from X and Y. We contaminated the target Y with 50 sample outliers (images with uniform entries in $[0, 1]$). A very small amount of noise is sufficient to derail COOT. Unlike COOT, UCOOT does not transport any outlier sample. Accuracy is computed as the percentage of mass within the block-diagonal structure.

HDA Results and robustness to target shift. The means and standard deviations of the accuracy on target data are reported in Table 3 for all the methods and all pairs of datasets. We observe that, thanks to its robustness, UCOOT outperforms COOT on 7 out of 9 dataset pairs, with higher average accuracy but also slightly larger variance. This is because of the difficulty of the unsupervised HDA problem and the instability present in all methods. In particular, GW-based approaches perform very poorly. This may be due to the fact that the pre-trained models contain meaningful but a very high-dimensional vectorial representation of the image. Thus, using the Euclidean distance matrices as inputs not only causes information loss but also is less relevant (see for example, (Aggarwal, Hinneburg, and Keim 2001), or Theorem 3.1.1 and Remark 3.1.2 in (Vershynin 2018)). We also illustrate the robustness of UCOOT to a change in class proportions, also known as target shift. More precisely, we simulate a change in proportion only in the source domain by selecting 20ρ samples per class for 4 amongst 10 classes with ρ decreasing from $\rho = 1$ to $\rho = 0.2$. In this configuration, the classes in the source domain are imbalanced and the unlabeled HDA problem becomes more difficult. We report the performance of all the methods as a function of the Total Variation (TV) between the class marginal distributions on one pair of datasets in Figure 4. We can see that UCOOT is quite robust to change in class proportions, while COOT experiences a sharp decrease in accuracy when the class distributions become more imbalanced.

Single-cell multi-omics alignment

Finally, we present a real-world application of UCOOT for the alignment of single-cell measurements. Recent advances in single-cell sequencing technologies allow biologists to measure a variety of cellular features at the single-cell resolution, such as expression levels of genes and epigenomic changes in the genome (Buenrostro et al. 2015; Chen, Lake, and Zhang 2019), or the abundance of surface proteins (Stoeckius et al. 2017). These multiple measurements produce single-cell multi-omics datasets. These datasets measuring different biological phenomena at the single-cell resolution allow scientists to study how the cellular processes are regulated, leading to finer cell variations during development and diseases. However, it is hard to obtain multiple types of measurements from the same individual cells due to experimental limitations. Therefore, many single-cell multi-omics datasets have disparate measurements from different sets of cells. As a result, computational methods are required to align the cells and the features of the different measurements to learn the relationships between them that help with data analysis and integration. Multiple tools (Cao, Hong, and Wan 2021; Hao et al. 2021; Liu et al. 2019), including GW (Cao, Hong, and Wan 2021; Demetci et al. 2022) and UGW (Demetci et al. 2021) based methods, have shown good performance for cell-to-cell alignments. However, aligning both

samples and features is a more challenging and critical task that GW and UGW-based methods cannot address Here we provide an application of UCOOT to simultaneously align the samples and features in a single-cell multi-omics dataset.

For demonstration, we choose a dataset generated by the CITE-seq experiment (Stoeckius et al. 2017), which simultaneously measures gene expression and antibody (or surface protein) abundance in single cells. From this dataset, we use 1000 human peripheral blood cells, which have ten antibodies and 17,014 genes profiled. We selected this specific dataset as we know the ground-truth correspondences on both the samples (i.e., cells) and the features (i.e., genes and their encoded antibodies), thus allowing us to quantify and compare the alignment performance of UCOOT and COOT. As done previously (Cao, Hong, and Wan 2021; Demetci et al. 2022; Liu et al. 2019), we quantify the cell alignment performance by calculating the fraction of cells closer than the true match (FOSCTTM) of each cell in the dataset and averaging it across all cells. This metric quantifies alignment error, so lower values are more desirable. The feature alignments are measured by calculating the accuracy of correct matching. The results are presented after hyperparameter tuning both methods with similar grid size per hyperparameter (see Experimental Details in Appendix).

Balanced Scenario. First, we select and align the same number of samples and features across the two datasets. For this, we subset the gene expression domain with the ten genes that match to the ten antibodies they express. Original data contains the same number of cells across domains since both domains are simultaneously measured in the same single-cells. We observe that both UCOOT and COOT can correctly align features (Figure 5 (a)) and the cells (Appendix Figure S1(a)) across the two measurements. However, UCOOT gives better performance, as demonstrated by a lower FOCSTTM score (0.0062 vs 0.0127) for cells. Both COOT and UCOOT recover the diagonal for matching features (100% accuracy), but UCOOT recovers the exact matching, likely due to its robustness to noise, whereas COOT assigns weights to other features as well.

Unbalanced Scenarios. Next, we perform alignment with an unequal number of features. This setting is more likely to occur for real-world single-cell datasets as different features are measured. In the first simple scenario, we align the ten antibodies with only a subset (five) of their matching genes. As visualized in Figure 5 (b), COOT struggles to find the correct feature alignments (60% accuracy), which would lie in the diagonal of the highlighted square (dashed lines). However, the relaxation of the mass conservation constraint in UCOOT allows it to shrink the mass of antibodies that lack matches in the gene expression domain, leading do higher accuracy (100% accuracy). Next, we align the ten antibodies with the 50 most variable genes in the dataset, including their matching genes. This alignment task is the most realistic scenario, as single-cell multi-omics data consists of high-dimensional datasets with a different number of features for different measurements. Therefore, biologists focus their analyses on the reduced set of most variable features (e.g. genes). It is also the most computationally challenging case among all our experiments on this dataset. Hence, we provide sample-level supervision to both methods by giving a cost penalization matrix based on the correct sample alignments to the sample alignment computation. We see in Figure 5(c) that in comparison to COOT (50% accuracy), UCOOT recovers more of the correct feature alignments (70% accuracy), and yields fewer redundant alignments (for more detail, see Experimental Details in Appendix). Note that UCOOT avoids incorrect mappings by locally shrinking the mass of the features or samples that lack correspondences. This avoids subsequent incorrect downstream analysis of the integration results. This property can also help users to discover rare cell types by observing the extent of mass relaxation per cell or prune uninformative features in the single-cell datasets.

Lastly, we also consider the case of unequal number of samples across the two measurements. This case is common in real world single-cell multi-omics datasets that are not simultaneously measured. Demetci *et al.* (Demetci et al. 2021) have shown that single-cell alignment methods that do not account for this mismatch yield poor alignment results. Therefore, we downsample the number of cells in one of the domains by 25% and perform alignment with the full set of cells in the other domain (Appendix Figure S1(b & c). We compute the FOSCTTM score for all cells that have a true match in the dataset and report the average values. UCOOT continues to yield a low FOSCTTM score (0.0081 compared to 0.0062 in the balanced scenario), while COOT shows a larger drop in performance (0.1342 compared to 0.0127 in the balanced scenario).

Discussion and conclusion

In this work, we present an extension of COOT called unbalanced COOT, where the hard constraint on the marginal distributions is replaced by a soft control via the KL divergence. The resulting problem not only benefits from the flexibility of COOT but also enjoys the provable robustness property under the presence of outliers, which is not the case for COOT. The experimental results confirm our findings, yielding a very competitive performance in the unsupervised HDA task, as well as meaningful feature couplings for the single-cell multi-omics alignment. Also, while UCOOT introduces additional hyper-parameters, domain knowledge can help narrow down the range of feasible values, thus reducing the time and computational cost of the tuning process. Further investigation should be carried out to fully understand and assess the observed efficiency of UCOOT in real-world applications, and also explore the possibilities of UCOOT in more diverse applicative settings, including its use as a loss in deep learning architectures. Lastly, from a theoretical perspective, statistical properties such as sample complexity or stability analysis are needed to better understand the intricate relations between the two sample and feature couplings.

Acknowledgement

The authors thank to Tuan Binh Nguyen for the fruitful discussion and invaluable suggestions. This work is funded by the projects OTTOPIA ANR-20-CHIA-0030, 3IA Côte d'Azur Investments ANR-19-P3IA-0002 of the French National Research Agency (ANR), the 3rd Programme d'Investissements

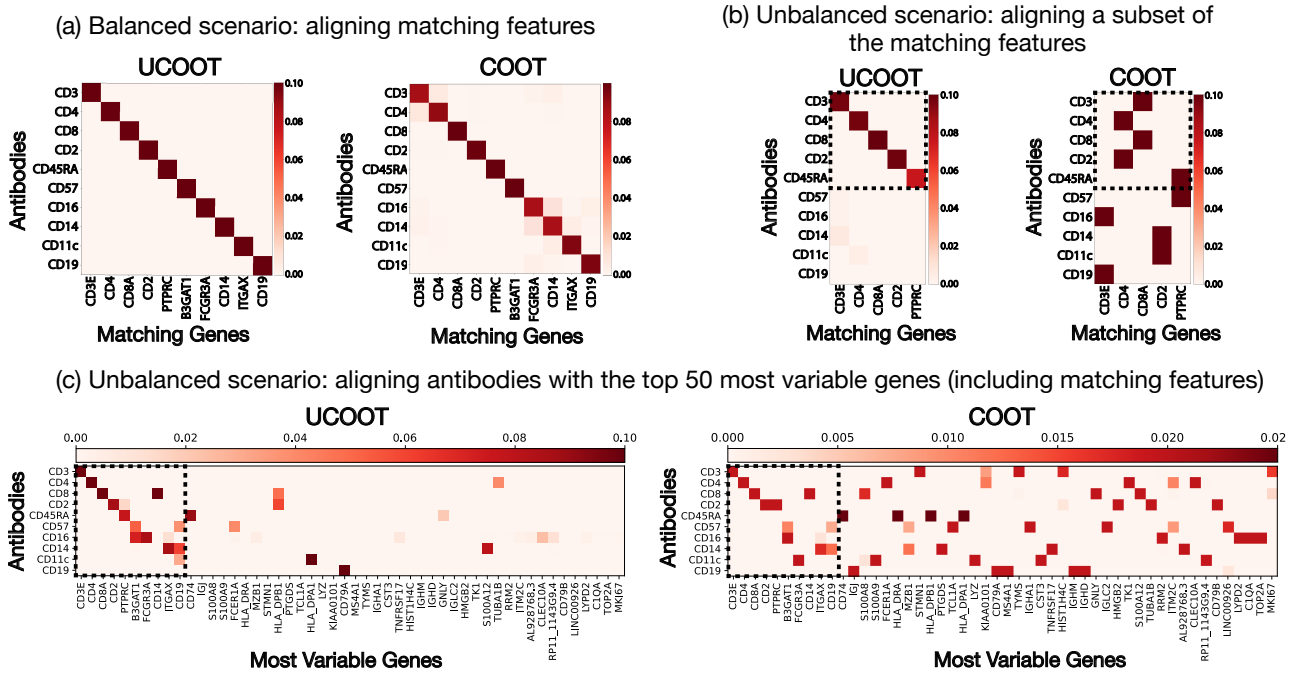

Fig. 5. Feature alignments on the single-cell multi-omics dataset of COOT and UCOOT between antibodies (surface proteins) and their matching genes (that encode them). **(a)** The features are sorted such that the correct alignment would yield a diagonal matrix. **(b)** Only five of the correct gene matches are kept (the last five genes from (a) are excluded). **(c)** Alignments between the ten antibodies and the top 50 most variable genes, including the matching genes. For **(b)** and **(c)**, the diagonal within the dashed square highlights the correct matches. Overall, UCOOT gives better feature alignments.

d'Avenir ANR-18-EUR-0006-02, the Chair "Challenging Technology for Responsible Energy" led by l'X – Ecole Polytechnique and the Fondation de l'Ecole Polytechnique, sponsored by TOTAL, and the Chair "Business Analytics for Future Banking" sponsored by NATIXIS. This research is produced within the framework of Energy4Climate Interdisciplinary Center (E4C) of IP Paris and Ecole des Ponts ParisTech. Pinar Demetci's and Ritambhara Singh's contribution was funded by R35 HG011939.

References

Aggarwal, C. C.; Hinneburg, A.; and Keim, D. A. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory — ICDT 2001*, 420–434. Springer Berlin Heidelberg.

Alvarez-Melis, D.; and Jaakkola, T. S. 2018. Gromov-Wasserstein Alignment of Word Embedding Spaces. *Empirical Methods in Natural Language Processing (EMNLP)*, 1881–1890.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein Generative Adversarial Networks. *International Conference on Machine Learning*, 70: 214–223.

Balaji, Y.; Chellappa, R.; and Feizi, S. 2020. Robust Optimal Transport with Applications in Generative Modeling and Domain Adaptation. *Advances in Neural Information Processing Systems*.

Bazeille, T.; Richard, H.; Janati, H.; and Thirion, B. 2019. Local optimal transport for functional brain template estimation. In *International Conference on Information Processing in Medical Imaging*, 237–248. Springer.

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

Ben Guebila, M.; Lopes-Ramos, C. M.; Weighill, D.; Sonawane, A.; Burkholz, R.; Shamsaei, B.; Platig, J.; Glass, K.; Kuijjer, M.; and Quackenbush, J. 2021. GRAND: a database of gene regulatory network models across human conditions. *Nucleic Acids Research*, 50(D1): D610–D621.

Bernhard, O. K.; Cunningham, A. L.; and Sheil, M. M. 2004. Analysis of proteins copurifying with the cd4/lck complex using one-dimensional polyacrylamide gel electrophoresis and mass spectrometry: comparison with affinity-tag based protein detection and evaluation of different solubilization methods. *Journal of the American Society for Mass Spectrometry*, 15(4): 558–567.

Billingsley, P. 1999. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics, 2 edition.

Buenrostro, J. D.; Wu, B.; Chang, H. Y.; and Greenleaf, W. J. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology*, 109(1): 21.29.1–21.29.9.

Burago, D.; Burago, Y.; and Ivanov, S. 2001. *A Course in Metric Geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society.

Cao, J.; Mo, L.; Zhang, Y.; Jia, K.; Shen, C.; and Tan, M. 2019. Multi-marginal Wasserstein GAN. *Advances in Neural Information Processing Systems*, 1774–1784.

Cao, K.; Hong, Y.; and Wan, L. 2021. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*. Btab594.

Carter, R. H.; Doody, G. M.; Bolen, J. B.; and Fearon, D. T. 1997. Membrane IgM-induced tyrosine phosphorylation of CD19 requires a CD19 domain that mediates association with components of the B cell antigen receptor complex. *The Journal of Immunology*, 158(7): 3062–3069.

Chapel, L.; Flamary, R.; Wu, H.; Févotte, C.; and Gasso, G. 2021. Unbalanced Optimal Transport through Non-negative Penalized Linear Regression. In *Neural Information Processing Systems (NeurIPS)*.

Chen, S.; Lake, B. B.; and Zhang, K. 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12): 1452–1457.

Chizat, L. 2017. *Unbalanced Optimal Transport: Models, Numerical Methods, Applications*. Ph.D. thesis, PSL Research University.

Chizat, L.; Peyré, G.; Schmitzer, B.; and Vialard, F.-X. 2018a. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87: 2563–2609.

Chizat, L.; Peyré, G.; Schmitzer, B.; and Vialard, F.-X. 2018b. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11): 3090–3123.

Chowdhury, S.; Needham, T.; Semrad, E.; Wang, B.; and Zhou, Y. 2021. Hypergraph Co-Optimal Transport: Metric and Categorical Properties. *arXiv preprint arXiv:1810.09646*.

Courty, N.; Flamary, R.; Tuia, D.; and Rakotomamonjy, A. 2016. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 1853–1865.

Csiszár, I. 1963. Eine informationstheoretische Ungleichung und ihre anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 8: 85–108.

Demetci, P.; Santorella, R.; Sandstede, B.; Noble, W. S.; and Singh, R. 2022. SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. *Journal of computational biology*, 29(1): 3–18.

Demetci, P.; Santorella, R.; Sandstede, B.; and Singh, R. 2021. Unsupervised integration of single-cell multi-omics datasets with disparities in cell-type representation. *bioRxiv*.

Fang, L.; Li, Y.; Ma, L.; Xu, Q.; Tan, F.; and Chen, G. 2020. GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Research*, 49(D1): D97–D103.

Fatras, K.; Séjourné, T.; Courty, N.; and Flamary, R. 2021. Unbalanced minibatch Optimal Transport; applications to Domain Adaptation. *International Conference on Machine Learning*.

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

Frogner, C.; Zhang, C.; Mobahi, H.; Araya, M.; and Poggio, T. A. 2015. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems*, 2053–2061.

Go, C. D.; Knight, J. D. R.; Rajasekharan, A.; Rathod, B.; Hesketh, G. G.; Abe, K. T.; Youn, J.-Y.; Samavarchi-Tehrani, P.; Zhang, H.; Zhu, L. Y.; Popiel, E.; Lambert, J.-P.; Coyaud, É.; Cheung, S. W. T.; Rajendran, D.; Wong, C. J.; Antonicka, H.; Pelletier, L.; Palazzo, A. F.; Shoubridge, E. A.; Raught, B.; and Gingras, A.-C. 2021. A proximity-dependent biotinylation map of a human cell. *Nature*, 595(7865): 120–124.

Gromov, M. 1981. Groups of polynomial growth and expanding maps (with an appendix by Jacques Tits). *Publications Mathématiques de l'IHÉS*, 53: 53–78.

Gromov, M. 1999. *Metric Structures for Riemannian and Non-Riemannian Spaces*, volume 152 of *Progress in Mathematics*. Birkhäuser, Boston, US.

Hao, Y.; Hao, S.; Andersen-Nissen, E.; III, W. M. M.; Zheng, S.; Butler, A.; Lee, M. J.; Wilk, A. J.; Darby, C.; Zagar, M.; Hoffman, P.; Stoeckius, M.; Papalexi, E.; Mimitou, E. P.; Jain, J.; Srivastava, A.; Stuart, T.; Fleming, L. B.; Yeung, B.; Rogers, A. J.; McElrath, J. M.; Blish, C. A.; Gottardo, R.; Smibert, P.; and Satija, R. 2021. Integrated analysis of multimodal single-cell data. *Cell*.

Janati, H.; Bazeille, T.; Thirion, B.; Cuturi, M.; and Gramfort, A. 2019. Group level MEG/EEG source imaging via optimal transport: minimum Wasserstein estimates. In *International Conference on Information Processing in Medical Imaging*, 743–754. Springer.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *Proceedings of the 22nd ACM International Conference on Multimedia*, 675–678.

Kalton, N. J.; and Ostrovskii, M. I. 1999. Distances between Banach spaces. *Forum Mathematicum*, 11(1): 17–48.

Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; and Tanabe, M. 2020. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1): D545–D551.

Kantorovich, L. 1942. On the transfer of masses (in Russian). *Doklady Akademii Nauk*, 37: 227–229.

Lee, J.; Bertrand, N. P.; and Rozell, C. J. 2020. Parallel Unbalanced Optimal Transport Regularization for Large Scale Imaging Problems. *arXiv preprint arXiv:1909.00149*.

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

Liu, J.; Huang, Y.; Singh, R.; Vert, J.-P.; and Noble, W. S. 2019. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, 644310.

Ma, Z.; Wei, X.; Hong, X.; Lin, H.; Qiu, Y.; and Gong, Y. 2021. Learning to Count via Unbalanced Optimal Transport.

Proceedings of the AAAI Conference on Artificial Intelligence, 35(3): 2319–2327.

Monge, G. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, 666–704.

Mukherjee, D.; Guha, A.; Solomon, J. M.; Sun, Y.; and Yurochkin, M. 2021. Outlier-Robust Optimal Transport. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 7850–7860. PMLR.

Mémoli, F. 2007. On the use of Gromov-Hausdorff Distances for Shape Comparison. In *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association.

Mémoli, F. 2011. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 1–71.

Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F.; Dolma, S.; Coulombe-Huntington, J.; Chatr-aryamontri, A.; Dolinski, K.; and Tyers, M. 2020. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1): 187–200.

Peyré, G.; Cuturi, M.; and Solomon, J. 2016. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. *International Conference on Machine Learning*, 48.

Pham, K.; Le, K.; Ho, N.; Pham, T.; and Bui, H. 2020. On Unbalanced Optimal Transport: An Analysis of Sinkhorn Algorithm. *Proceedings of the 37 th International Conference on Machine Learning*, 119.

Ponti, N. D.; and Mondino, A. 2020. Entropy-Transport distances between unbalanced metric measure spaces. *arXiv preprint arXiv:2009.10636*.

Redko, I.; Courty, N.; Flamary, R.; and Tuia, D. 2019. Optimal Transport for Multi-source Domain Adaptation under Target Shift. *Proceedings of Machine Learning Research*, 89.

Redko, I.; Vayer, T.; Flamary, R.; and Courty, N. 2020. CO-Optimal Transport. *Advances in Neural Information Processing Systems*, 33.

Rolet, A.; Cuturi, M.; and Peyré, G. 2016. Fast Dictionary Learning with a Smoothed Wasserstein Loss. *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 51: 630–638.

Saenko, K.; Kulis, B.; Fritz, M.; and Darrell, T. 2010. Adapting visual category models to new domains. *Proceedings of the 11th European Conference on Computer Vision*, 213–226.

Scetbon, M.; Cuturi, M.; and Peyré, G. 2021. Low-Rank Sinkhorn Factorization. *Proceedings of the 38th International Conference on Machine Learning*, 139: 9344–9354.

Schiebinger, G.; Shu, J.; Tabaka, M.; Cleary, B.; Subramanian, V.; Solomon, A.; Gould, J.; Liu, S.; Lin, S.; Berube, P.; Lee, L.; Chen, J.; Brumbaugh, J.; Rigollet, P.; Hochedlinger, K.; Jaenisch, R.; Regev, A.; and Lander, E. S. 2019. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4): 928–943.

Schmitzer, B. 2019. Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems. *SIAM Journal on Scientific Computing*, 41: 1443–1481.

Schraven, B.; Samstag, Y.; Altevogt, P.; and Meuer, S. C. 1990. Association of CD2 and CD45 on human T lymphocytes. *Nature*, 345(6270): 71–74.

Singh, R.; Demetci, P.; Bonora, G.; Ramani, V.; Lee, C.; Fang, H.; Duan, Z.; Deng, X.; Shendure, J.; Disteche, C.; et al. 2020. Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–10.

Solomon, J.; Peyré, G.; Kim, V. G.; and Sra, S. 2016. Entropic Metric Alignment for Correspondence Problems. *ACM Transactions on Graphics*, 35(4).

Solomon, J.; Rustamov, R.; Guibas, L.; and Butscher, A. 2014. Wasserstein Propagation for Semi-Supervised Learning. *Proceedings of the 31st International Conference on Machine Learning*, 32: 306–314.

Stoeckius, M.; Hafemeister, C.; Stephenson, W.; Houck-Loomis, B.; Chattopadhyay, P. K.; Swerdlow, H.; Satija, R.; and Smibert, P. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9): 865–868.

Sturm, K.-T. 2006. On the geometry of metric measure spaces. *Acta Mathematica*, 196: 65–131.

Sturm, K.-T. 2012. The space of spaces: curvature bounds and gradient flows on the space of metric measure spaces. *arXiv preprint arXiv:1208.0434*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper with Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

Szklarczyk, D.; Gable, A. L.; Nastou, K. C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N. T.; Legeay, M.; Fang, T.; Bork, P.; Jensen, L. J.; and von Mering, C. 2020. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1): D605–D612.

Séjourné, T.; Vialard, F.-X.; and Peyré, G. 2021. The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation. *Advances in Neural Information Processing Systems*, 34.

Vayer, T.; Chapel, L.; Flamary, R.; Tavenard, R.; and Courty, N. 2019. Optimal Transport for structured data with application on graphs. *International Conference on Machine Learning*, 97.

Vershynin, R. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Villani, C. 2003. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society.

Xu, H.; Luo, D.; and Carin, L. 2019. Scalable Gromov-Wasserstein Learning for Graph Partitioning and Matching. *Advances in Neural Information Processing Systems*, 32.

Xu, H.; Luo, D.; Zha, H.; and Duke, L. C. 2019. Gromov-Wasserstein Learning for Graph Matching and Node Embedding. *Proceedings of the 36th International Conference on Machine Learning*, 6932–6941.

Yan, Y.; Li, W.; Wu, H.; Min, H.; Tan, M.; and Wu, Q. 2018. Semi-Supervised Optimal Transport for Heterogeneous Domain Adaptation. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 2969–2975.

Yang, K. D.; and Uhler, C. 2019. Scalable Unbalanced Optimal Transport using Generative Adversarial Networks. *7th International Conference on Learning Representations*.

Zhang, Z.; Mroueh, Y.; Goldfeld, Z.; and Sriperumbudur, B. K. 2021. Cycle Consistent Probability Divergences Across Different Spaces. *arXiv preprint arXiv:2111.11328*.

Appendix

Additional concepts and notations

Denote $\mathcal{C}_b(\mathcal{X})$ the space of bounded continuous functions on $\mathcal{X}$. Given a Borel measurable map $T : \mathcal{X} \to \mathcal{Y}$ and a measure $\mu \in \mathcal{M}^+(\mathcal{X})$, we define the push-forward (or image) measure $T_{\#}\mu \in \mathcal{M}^+(\mathcal{Y})$ is the one which satisfies: for every $\phi \in \mathcal{C}_b(\mathcal{Y})$, $\int_{\mathcal{X}}(\phi \circ T)\ d\mu = \int_{\mathcal{Y}} \phi\ dT_{\#}\mu$. For example, the marginal distributions of a measure on $\mathcal{X} \times \mathcal{Y}$ are the push-forward measures induced by the canonical projections: $P_{\mathcal{X}}(x, y) = x$ and $P_{\mathcal{Y}}(x, y) = y$. Given two probability measures $(\mu, \nu) \in \mathcal{M}_1^+(\mathcal{X}) \times \mathcal{M}_1^+(\mathcal{Y})$, denote $U(\mu, \nu) = \{\pi \in \mathcal{M}_1^+(\mathcal{X} \times \mathcal{Y}) : \pi_{\#1} = \mu, \pi_{\#2} = \nu\}$ the set of admissible transport plans.

Given an entropy function $\varphi : \mathbb{R}_{>0} \to [0, \infty]$ (i.e. it is convex, positive and lower semi-continuous such that $\varphi(1) = 0$), we define the recession constant $\varphi'_\infty \in \mathbb{R} \cup \{\infty\}$ as $\varphi'_\infty = \lim_{x\to\infty} \frac{\varphi(x)}{x}$. The Csiszár divergence (or φ-divergence) (Csiszár 1963) between two measures μ and ν in a certain space $\mathcal{M}^+(\mathcal{S})$ is defined as $D_\varphi(\mu|\nu) = \int_{\mathcal{S}} \varphi\Big(\frac{d\mu}{d\nu}\Big)d\nu + \varphi'_\infty \int_{\mathcal{S}} d\mu^\perp$, where, by Lebesgue decomposition, we have $\mu = \frac{d\mu}{d\nu}\nu + \mu^\perp$. For example,

- If $\varphi(x) = x \log x - x + 1$, then D_φ is the Kullback-Leibler (KL) divergence.
- If $\varphi(x)$ is equal to 0 if $x = 1$ and $+\infty$ otherwise, then D_φ is the indicator divergence $\iota_=$.

For later convenience, we define the function $|\xi_1 - \xi_2|^p : (\mathcal{X}_1^s \times \mathcal{X}_2^s) \times (\mathcal{X}_1^f \times \mathcal{X}_2^f) \to \mathbb{R}_{\geq 0}$ by

$$|\xi_1 - \xi_2|^p\big((x_1^s, x_2^s), (x_1^f, x_2^f)\big) := |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p,$$

and write the objective function of generalised COOT as

$$F_\lambda(\pi^s, \pi^f) = \iint |\xi_1 - \xi_2|^p \mathrm{d}\pi^s \mathrm{d}\pi^f + \sum_{k=1}^{2} \lambda_k D_k(\pi^s_{\#k} \otimes \pi^f_{\#k} | \mu^s_k \otimes \mu^f_k).$$

The generalized COOT now reads compactly as

$$\inf_{\substack{\pi^s \in \mathcal{M}^+(\mathcal{X}_1^s \times \mathcal{X}_2^s) \\ \pi^f \in \mathcal{M}^+(\mathcal{X}_1^f \times \mathcal{X}_2^f) \\ m(\pi^s) = m(\pi^f)}} F_\lambda(\pi^s, \pi^f) \tag{3}$$

Proofs related to UCOOT

UCOOT and its properties

Claim 1. *When $D_k = \iota_=$ and μ_k^s, μ_k^f are probability measures, for $k = 1, 2$, then we recover COOT from generalized COOT.*

PROOF. Under the above assumptions, the generalized COOT problem becomes

$$\begin{aligned} \inf_{\substack{\pi^s \in \mathcal{M}^+(\mathcal{X}_1^s \times \mathcal{X}_2^s) \\ \pi^f \in \mathcal{M}^+(\mathcal{X}_1^f \times \mathcal{X}_2^f)}} & \iint |\xi_1 - \xi_2|^p \mathrm{d}\pi^s \mathrm{d}\pi^f \\ \text{subject to } & \pi^s_{\#1} \otimes \pi^f_{\#1} = \mu^s_1 \otimes \mu^f_1 \text{ (C1)} \\ & \pi^s_{\#2} \otimes \pi^f_{\#2} = \mu^s_2 \otimes \mu^f_2 \text{ (C2)} \\ & m(\pi^s) = m(\pi^f) \text{ (C3)}\ . \end{aligned}$$

As $m(\pi) = m(\pi_{\#1}) = m(\pi_{\#2})$, for any measure π, and μ_k^s, μ_k^f are probability measures, for $k = 1, 2$, one has $m(\pi^s)m(\pi^f) = 1$, thus $m(\pi^s) = m(\pi^f) = 1$. Now, the constraint C1 implies that $\int_{\mathcal{X}_1^s} \mathrm{d}\pi^s_{\#1}\mathrm{d}\pi^f_{\#1} = \int_{\mathcal{X}_1^s} \mathrm{d}\mu^s_1\mathrm{d}\mu^f_1$. Thus, $\pi^f_{\#1} = \mu^f_1$. Similarly, we have $\pi^s_{\#k} = \mu^s_k$ and $\pi^f_{\#k} = \mu^f_k$, for any $k = 1, 2$. We conclude that $\pi^f \in U(\mu^f_1, \mu^f_2)$ and $\pi^s \in U(\mu^s_1, \mu^s_2)$, and we obtain the COOT problem. ■

Proposition 3. *(Existence of minimizer) Denote $\mathcal{S} := (\mathcal{X}_1^s \times \mathcal{X}_2^s) \times (\mathcal{X}_1^f \times \mathcal{X}_2^f)$. The problem 3 admits a minimizer if at least one of the following conditions hold:*

1. *The entropy functions ϕ_1 and ϕ_2 are superlinear, i.e. $(\phi_1)'_\infty = (\phi_2)'_\infty = \infty$.*
2. *The function $|c_X - c_Y|^p$ has compact sublevels in $\mathcal{S}$ and $\inf_{\mathcal{S}} |\xi_1 - \xi_2|^p + \lambda_1(\phi_1)'_\infty + \lambda_2(\phi_2)'_\infty > 0$.*

PROOF. We adapt the proof of Theorem 3.3 in (Liero, Mielke, and Savaré 2018) and of Proposition 3 in (Séjourné, Vialard, and Peyré 2021). For convenience, we write $\mu_1 = \mu_1^s \otimes \mu_1^f$ and $\mu_2 = \mu_2^s \otimes \mu_2^f$. For each pair (π^s, π^f), denote $\pi = \pi^s \otimes \pi^f$. It can be shown that $\pi_{\#k} := (P_{\mathcal{X}_k^s \times \mathcal{X}_k^f})_\# \pi = (P_{\mathcal{X}_k^s})_\# \pi^s \otimes (P_{\mathcal{X}_k^f})_\# \pi^f = \pi^s_{\#k} \otimes \pi^f_{\#k}$, for $k = 1, 2$. Indeed, for any function $\phi \in \mathcal{C}_b(\mathcal{X}_k^s \times \mathcal{X}_k^f)$, we have

$$\begin{aligned}\int_{\mathcal{X}_k^s \times \mathcal{X}_k^f} \phi \, \mathrm{d}(P_{\mathcal{X}_k^s \mathcal{X}_k^f})_\# \pi &= \int_{\mathcal{S}} (\phi \circ P_{\mathcal{X}_k^s \mathcal{X}_k^f}) \, \mathrm{d}\pi \\ &= \int_{\mathcal{S}} \phi(x_k^s, x_k^f) \, \mathrm{d}\pi^s(x_1^s, x_2^s) \, \mathrm{d}\pi^f(x_1^f, x_2^f) \\ &= \int_{\mathcal{X}_k^s \times \mathcal{X}_k^f} \phi \, \mathrm{d}\pi^s_{\#k} \, \mathrm{d}\pi^f_{\#k}.\end{aligned}$$

Thus, the problem 3 can be rewritten as

$$\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = \inf_{\pi \in E_{uco}} \int_{\mathcal{S}} |\xi_1 - \xi_2|^p \mathrm{d}\pi + \sum_{k=1,2} \lambda_k D_{\phi_k}(\pi_{\#k} | \mu_k),$$

where

$$E_{uco} = \{\pi \in \mathcal{M}^+(\mathcal{S}) | \pi = \pi^s \otimes \pi^f, \pi^s \in \mathcal{M}^+(\mathcal{X}_1^s \times \mathcal{X}_2^s), \pi^f \in \mathcal{M}^+(\mathcal{X}_1^f \times \mathcal{X}_2^f)\}.$$

Define

$$L(\pi) := \int_{\mathcal{S}} |\xi_1 - \xi_2|^p \mathrm{d}\pi + \sum_{k=1,2} \lambda_k D_{\phi_k}(\pi_{\#k} | \mu_k).$$

By Jensen's inequality, we have

$$\begin{aligned}L(\pi) &\geq m(\pi) \inf_{\mathcal{S}} |\xi_1 - \xi_2|^p + \sum_{k=1,2} \lambda_k m(\mu_k) \phi_k\Big(\frac{m(\pi_{\#k})}{m(\mu_k)}\Big) \\ &= m(\pi) \Big[\inf_{\mathcal{S}} |\xi_1 - \xi_2|^p + \sum_{k=1,2} \lambda_k \frac{m(\mu_k)}{m(\pi)} \phi_k\Big(\frac{m(\pi)}{m(\mu_k)}\Big) \Big],\end{aligned}$$

where, in the last equality, we use the relation $m(\pi) = m(\pi_{\#k})$, for $k = 1, 2$. It follows from the assumption that L is coercive, i.e. $L(\pi) \to \infty$ when $m(\pi) \to \infty$.

Clearly $\inf_{E_{uco}} L < \infty$ because $L\big((\mu_1^s \otimes \mu_2^s) \otimes (\mu_1^f \otimes \mu_2^f)\big) < \infty$. Let $(\pi_n)_n \subset E_{uco}$ be a minimizing sequence, i.e. $L(\pi_n) \to \inf_{E_{uco}} L$. Such sequence is necessarily bounded (otherwise, there exists a subsequence $(\pi_{n_k})_{n_k}$ with $m(\pi_{n_k}) \to \infty$ and the coercivity of L implies $L(\pi_{n_k}) \to \infty$, which is absurd). Suppose $m(\pi_n) \leq M$, for some $M > 0$. By Tychonoff's theorem, as $\mathcal{X}_k^s$ and $\mathcal{X}_k^f$ are compact spaces, so is the product space $\mathcal{S}$. Thus, by Banach-Alaoglu theorem, the ball $B_M = \{\pi \in \mathcal{M}^+(\mathcal{S}) : m(\pi) \leq M\}$ is weakly compact in $\mathcal{M}^+(\mathcal{S})$.

Consider the set $\overline{E}_{uco} = E_{uco} \cap B_M$, then clearly $(\pi_n)_n \subset \overline{E}_{uco}$. We will show that there exists a converging subsequence of $(\pi_n)_n$, whose limit is in $\overline{E}_{uco}$, thus $\overline{E}_{uco}$ is weakly compact. Indeed, by definition of E_{uco}, there exist two sequences $(\pi_n^s)_n$ and $(\pi_n^f)_n$ such that $\pi_n = \pi_n^s \otimes \pi_n^f$. We can assume furthermore that $m(\pi_n^s) = m(\pi_n^f) = \sqrt{m(\pi_n)} \leq \sqrt{M}$. As $m(\pi_n^s)$ and $m(\pi_n^f)$ are bounded, by reapplying Banach-Alaoglu theorem, one can extract two converging subsequences (after reindexing) $\pi_n^s \rightharpoonup \pi^s \in \mathcal{M}^+(\mathcal{X}_1^s \times \mathcal{X}_2^s)$ and $\pi_n^f \rightharpoonup \pi^f \in \mathcal{M}^+(\mathcal{X}_1^f \times \mathcal{X}_2^f)$, with $m(\pi^s) = m(\pi^f) \leq \sqrt{M}$. An immediate extension of Theorem 2.8 in (Billingsley 1999) to the convergence of the products of bounded positive measures implies $\pi_n^s \otimes \pi_n^f \rightharpoonup \pi^s \otimes \pi^f \in \overline{E}_{uco}$.

Now, the lower semicontinuity of L implies that $\inf_{E_{uco}} L \geq L(\pi^s \otimes \pi^f)$, thus $L(\pi^s \otimes \pi^f) = \inf_{E_{uco}} L$ and (π^s, π^f) is a solution of the problem 3. ■

Claim 2. *Suppose that $\mathbb{X}_1$ and $\mathbb{X}_2$ are two finite sample-feature spaces such that $(\mathcal{X}_1^s, \mathcal{X}_2^s)$ and $(\mathcal{X}_1^f, \mathcal{X}_2^f)$ have the same cardinality and are equipped with the uniform measures $\mu_1^s = \mu_2^s$, $\mu_1^f = \mu_2^f$. Then $\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$ if and only if there exist perfect alignments between rows (samples) and between columns (features) of the interaction matrices ξ_1 and ξ_2.*

PROOF. Without loss of generality, we can assume that μ_k^s and μ_k^f are discrete uniform probability distributions, for $k = 1, 2$. By Proposition 1 in (Redko et al. 2020), under the assumptions on $\mathbb{X}_1$ and $\mathbb{X}_2$, we have $\mathrm{COOT}(\mathbb{X}_1, \mathbb{X}_2) = 0$ if and only if there exist perfect alignments between rows (samples) and between columns (features) of the interaction matrices ξ_1 and ξ_2. So, it is enough to prove that $\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$ if and only if $\mathrm{COOT}(\mathbb{X}_1, \mathbb{X}_2) = 0$.

Let (π^s, π^f) be a pair of equal-mass couplings such that $\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$. It follows that $\pi^s_{\#k} \otimes \pi^f_{\#k} = \mu^s_k \otimes \mu^f_k$, for $k = 1, 2$. Consequently, $m(\pi^s)m(\pi^f) = m(\mu^s_1)m(\mu^f_1) = 1$, so $m(\pi^s) = m(\pi^f) = 1$. Now, we have $\int_{\mathcal{X}^s_k} \mathrm{d}\pi^s_{\#k}\, \mathrm{d}\pi^f_{\#k} = \int_{\mathcal{X}^s_k} \mathrm{d}\mu^s_k\, \mathrm{d}\mu^f_k$, or equivalently, $\pi^f_{\#k} = \mu^f_k$. Similarly, $\pi^s_{\#k} = \mu^s_k$, meaning that $\pi^s \in U(\mu^s_1, \mu^s_2)$ and $\pi^f \in U(\mu^f_1, \mu^f_2)$. Thus, $\mathrm{COOT}(\mathbb{X}_1, \mathbb{X}_2) = \mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$.

For the other direction, suppose that $\mathrm{COOT}(\mathbb{X}_1, \mathbb{X}_2) = 0$. Let (π^s, π^f) be a pair of couplings such that $\mathrm{COOT}(\mathbb{X}, \mathbb{Y}) = 0$. As $\pi^s \in U(\mu^s_1, \mu^s_2)$ and $\pi^f \in U(\mu^f_1, \mu^f_2)$, one has $\mathrm{COOT}(\mathbb{X}_1, \mathbb{X}_2) = F_\lambda(\pi^s, \pi^f) \geq \mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) \geq 0$, for every $\lambda_1, \lambda_2 > 0$. So, $\mathrm{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$. ■

Robustness of UCOOT and sensitivity of COOT

First, we recall our assumptions.

Assumption 2. *Consider two sample-feature spaces $\mathbb{X}_k = ((\mathcal{X}^s_k, \mu^s_k), (\mathcal{X}^f_k, \mu^f_k), \xi_k)$, for $k = 1, 2$. Let ε^s (resp. ε^f) be a probability measure with compact support $\mathcal{O}^s$ (resp. $\mathcal{O}^f$). For $a \in \{s, f\}$, define the noisy distribution $\widetilde{\mu}^a = \alpha_a \mu^a + (1 - \alpha_a)\varepsilon^a$, where $\alpha_a \in [0, 1]$. We assume that ξ_1 is defined on $(\mathcal{X}^s_1 \cup \mathcal{O}^s) \times (\mathcal{X}^f_1 \cup \mathcal{O}^f)$ and that ξ_1, ξ_2 are continuous on their supports. We denote the contaminated sample-feature space by $\widetilde{\mathbb{X}_1} = ((\mathcal{X}^s_1 \cup \mathcal{O}^s, \widetilde{\mu}^s_1), (\mathcal{X}^f_1 \cup \mathcal{O}^f, \widetilde{\mu}^f_1), \xi_1)$. Finally, we define some useful minimal and maximal costs:*

$$\begin{cases} \Delta_0 \overset{\text{def}}{=} & \min_{\substack{x^s_1 \in \mathcal{O}^s, x^f_1 \in \mathcal{O}^f \\ x^s_2 \in \mathcal{X}^s_2, x^f_2 \in \mathcal{X}^f_2}} \quad |\xi_1(x^s_1, x^f_1) - \xi_2(x^s_2, x^f_2)|^p \\ \Delta_\infty \overset{\text{def}}{=} & \max_{\substack{x^s_1 \in \mathcal{X}^s_1 \cup \mathcal{O}^s, x^f_1 \in \mathcal{X}^f_1 \cup \mathcal{O}^f \\ x^s_2 \in \mathcal{X}^s_2, x^f_2 \in \mathcal{X}^f_2}} \quad |\xi_1(x^s_1, x^f_1) - \xi_2(x^s_2, x^f_2)|^p \end{cases}.$$

For convenience, we write $C \overset{\text{def}}{=} |\xi_1 - \xi_2|^p$ and $\widetilde{\mathcal{S}} := (\mathcal{X}^s_1 \cup \mathcal{O}^s) \times \mathcal{X}^s_2 \times (\mathcal{X}^f_1 \cup \mathcal{O}^f) \times \mathcal{X}^f_2$.

Proposition 4. *(COOT is sensitive to outliers) Consider $\widetilde{\mathbb{X}_1}, \mathbb{X}_2$ as defined in Assumption 2. Then:*

$$\mathrm{COOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \geq (1 - \alpha_s)(1 - \alpha_f)\Delta_0.$$

PROOF. Consider a pair of feasible alignments (π^s, π^f). Since C is non-negative, taking the COOT integral over a smaller set leads to the lower bound:

$$\begin{aligned} \int_{\widetilde{\mathcal{S}}} C \mathrm{d}\pi^s \mathrm{d}\pi^f &\geq \int_{\mathcal{O}^s \times \mathcal{X}^s_2 \times \mathcal{O}^f \times \mathcal{X}^f_2} C \mathrm{d}\pi^s \mathrm{d}\pi^f \\ &\geq \Delta_0 \int_{\mathcal{O}^s \times \mathcal{X}^s_2 \times \mathcal{O}^f \times \mathcal{X}^f_2} \mathrm{d}\pi^s \mathrm{d}\pi^f \\ &= \Delta_0 \int_{\mathcal{O}^s \times \mathcal{O}^f} \mathrm{d}\pi^s_{\#1} \mathrm{d}\pi^f_{\#1} \\ &\geq (1 - \alpha_s)(1 - \alpha_f)\Delta_0, \end{aligned}$$

where the last inequality follows from the marginal constraints. ■

Theorem 2. *(UCOOT is robust to outliers) Consider two sample-feature spaces $\widetilde{\mathbb{X}_1}, \mathbb{X}_2$ as defined in Assumption 2. Let $\delta \overset{\text{def}}{=} 2(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f)$ and $K = M + \frac{1}{M}\mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + \delta$, where $M = m(\pi^s) = m(\pi^f)$ is the transported mass between clean data. Then:*

$$\mathrm{UCOOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \leq \alpha_s\alpha_f\, \mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + \delta M \left[1 - \exp\left(-\frac{\Delta_\infty(1 + M) + K}{\delta M}\right)\right].$$

To get the exponential bound of this theorem, we use the following lemma.

Lemma 1. *Let $\varphi : t \in (0, 1] \mapsto t\log(t) - t + 1$ and $f_{a,b} : t \in (0, 1] \mapsto t \to at + b\varphi(t)$ for some $a, b > 0$. Then:*

$$\min_{t \in (0,1]} f_{a,b}(t) = b(1 - e^{-a/b}) = f_{a,b}(e^{-\frac{a}{b}}).$$

PROOF. Since $f_{a,b}$ is convex, cancelling the gradient is sufficient for optimality. The solution follows immediately. ■

PROOF. The proof uses the same core idea of (Fatras et al. 2021) but is slightly more technical for two reasons: (1) we consider arbitrary outlier distributions instead of simple Diracs; (2) we consider sample-feature outliers which requires more technical derivations.

The idea of proof is as follows. First, we construct sample and feature couplings from the solution of "clean" UCOOT and the reference measures. Then, they are used to upper bound the "noisy" UCOOT. By manipulating this bound, the "clean" UCOOT term will appear. A variable $t \in (0, 1)$ is also introduced in the fabricated couplings. The upper bound becomes a function of t and can be optimized to obtain the final bound.

Now, we prove Theorem 2.

Fabricating sample and feature couplings. Given the equal-mass solution (π^s, π^f) of the UCOOT problem, with $m(\pi^s) = m(\pi^f) = M$, consider, for $t \in (0, 1)$, a pair of sub-optimal transport plans:

$$\begin{aligned}\widetilde{\pi}^s &= \alpha_s \pi^s + t(1-\alpha_s)\varepsilon_s \otimes \mu_2^s \\ \widetilde{\pi}^f &= \alpha_f \pi^f + t(1-\alpha_f)\varepsilon_f \otimes \mu_2^f.\end{aligned}$$

Then, for $a \in \{s, f\}$, it holds:

- $\widetilde{\pi}^a_{\#1} = \alpha_k \pi^a_{\#1} + t(1-\alpha_a)\varepsilon_a$,
- $\widetilde{\pi}^a_{\#2} = \alpha_k \pi^a_{\#2} + t(1-\alpha_a)\mu_2^a$,
- $m(\widetilde{\mu}_1^a) = 1$ and $m(\widetilde{\pi}^a) = \alpha_a M + (1-\alpha_a)t$.

Establishing and manipulating the upper bound. Denote $q \overset{\text{def}}{=} (1-\alpha_s)(1-\alpha_f)$, $s \overset{\text{def}}{=} \alpha_s(1-\alpha_f) + \alpha_f(1-\alpha_s)$ and recall that on $\widetilde{\mathcal{S}}$, the cost C is upper bounded by $\Delta_\infty = \max_{\widetilde{\mathcal{S}}} |\xi_1 - \xi_2|^p$. First we upper bound the transportation cost:

$$\begin{aligned}&\int_{\widetilde{\mathcal{S}}} C \, \mathrm{d}\widetilde{\pi}^s \, \mathrm{d}\widetilde{\pi}^f \\ &= \alpha_s\alpha_f \int_{\widetilde{\mathcal{S}}} C \, \mathrm{d}\pi^s \, d\pi^f + t\sum_{k\neq i}(1-\alpha_i)\alpha_k \int_{\widetilde{\mathcal{S}}} C \, \mathrm{d}\varepsilon_i \, \mathrm{d}\mu_2^i \, \mathrm{d}\pi^k + qt^2 \int_{\widetilde{\mathcal{S}}} C \, \mathrm{d}\varepsilon_s \, \mathrm{d}\mu_2^s \, \mathrm{d}\varepsilon_f \, \mathrm{d}\mu_2^f \\ &\leq \alpha_s\alpha_f \int_{\mathcal{S}} C \, \mathrm{d}\pi^s \, \mathrm{d}\pi^f + \Delta_\infty(Ms + q)t \ ,\end{aligned}$$

since $t^2 \leq t$.

Second, we turn to the KL marginal discrepancies. We would like to extract the KL terms involving only the clean transport plans from the contaminated ones. We first detail both joint KL divergences for the source measure indexed by 1. The same holds for the target measure:

$$\begin{aligned}\mathrm{KL}(\widetilde{\pi}^s_{\#1} \otimes \widetilde{\pi}^f_{\#1} | \widetilde{\mu}^s_1 \otimes \widetilde{\mu}^f_1) &= \sum_{k\neq i} m(\widetilde{\pi}^i)\,\mathrm{KL}(\widetilde{\pi}^k_{\#1}|\widetilde{\mu}^k_1) + \prod_k \big(m(\widetilde{\pi}^k) - 1\big) \\ \mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) &= M\sum_k \mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + (M-1)^2.\end{aligned} \tag{4}$$

Now we upper bound each smaller KL term using the joint convexity of the KL divergence:

$$\begin{aligned}\mathrm{KL}(\widetilde{\pi}^k_{\#1}|\widetilde{\mu}^k_1) &\leq \alpha_k\,\mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + (1-\alpha_k)\,\mathrm{KL}(t\varepsilon_k|\varepsilon_k) \\ &= \alpha_k\,\mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + (1-\alpha_k)\varphi(t),\end{aligned}$$

where $\varphi(t) = t\log t - t + 1$, for $t > 0$. Thus, for $k \neq i$:

$$\begin{aligned}&m(\widetilde{\pi}^i)\,\mathrm{KL}(\widetilde{\pi}^k_{\#1}|\widetilde{\mu}^k_1) \leq m(\widetilde{\pi}^i)\alpha_k\,\mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + m(\widetilde{\pi}^i)(1-\alpha_k)\varphi(t) \\ &= \alpha_i\alpha_k M\,\mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + t(1-\alpha_i)\alpha_k\,\mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + \alpha_i(1-\alpha_k)M\varphi(t) + tq\varphi(t).\end{aligned}$$

Summing over f and s, we obtain:

$$\begin{aligned}&\sum_{k\neq i} m(\widetilde{\pi}^i)\,\mathrm{KL}(\widetilde{\pi}^k_{\#1}|\widetilde{\mu}^k_1) \\ &\leq \alpha_s\alpha_f M\sum_k \mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + t\sum_{k\neq i}(1-\alpha_i)\alpha_k\,\mathrm{KL}(\pi^k_{\#1}|\mu^k_1) + Ms\varphi(t) + 2qt\varphi(t) \\ &\leq (\alpha_s\alpha_f + \frac{ts}{M})\Big(\mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1}|\mu^s_1 \otimes \mu^f_1) - (1-M)^2\Big) + Ms\varphi(t) + 2qt\varphi(t).\end{aligned}$$

where, in the last bound, we used the second equation of (4) and the fact that $\alpha_s(1-\alpha_f) \leq s$ and $\alpha_f(1-\alpha_s) \leq s$. The product of masses of (4) can be written:

$$\begin{aligned}\prod_k \big(m(\widetilde{\pi}^k) - 1\big) &= \prod_k \big(\alpha_k(M-1) + (1-\alpha_k)(t-1)\big) \\ &= \alpha_s\alpha_f(1-M)^2 + s(1-M)(1-t) + q(1-t)^2.\end{aligned}$$

Thus, combining these upper bounds for the source measure:

$$\begin{aligned}\mathrm{KL}(\widetilde{\pi}^s_{\#1} \otimes \widetilde{\pi}^f_{\#1} | \widetilde{\mu}^s_1 \otimes \widetilde{\mu}^f_1) \leq \alpha_s \alpha_f \, \mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) \\ + \frac{ts}{M} \left(\mathrm{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | \mu^s_1 \otimes \mu^f_1) - (1-M)^2 \right) \\ + \left[sM\varphi(t) + 2qt\varphi(t) + s(1-M)(1-t) + q(1-t)^2 \right],\end{aligned}$$

and similarly, for the target measure:

$$\begin{aligned}\mathrm{KL}(\widetilde{\pi}^s_{\#2} \otimes \widetilde{\pi}^f_{\#2} | \mu^s_2 \otimes \mu^f_2) \leq \alpha_s \alpha_f \, \mathrm{KL}(\pi^s_{\#2} \otimes \pi^f_{\#2} | \mu^s_2 \otimes \mu^f_2) \\ + \frac{ts}{M} \left(\mathrm{KL}(\pi^s_{\#2} \otimes \pi^f_{\#2} | \mu^s_2 \otimes \mu^f_2) - (1-M)^2 \right) \\ + \left[sM\varphi(t) + 2qt\varphi(t) + s(1-M)(1-t) + q(1-t)^2 \right].\end{aligned}$$

Then, for every $0 < t \leq 1$, by summing all bounds:

$$\begin{aligned}\mathrm{UCOOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \leq \alpha_s \alpha_f \, \mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + \Delta_\infty (Ms + q)t \\ + \frac{ts}{M} (\mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) - (\lambda_1 + \lambda_2)(1-M)^2) \\ + (\lambda_1 + \lambda_2) \left[sM\varphi(t) + 2qt\varphi(t) + s(1-M)(1-t) + q(1-t)^2 \right].\end{aligned}$$

Minimizing the upper bound with respect to t. To obtain the exponential bound, we would like have an upper bound of the form $at + b\varphi(t)$, so that lemma 1 applies. Knowing that $1 \leq 2(t + \varphi(t))$ for any $t \in [0, 1]$:

Let's first isolate the quantity that is not of this form: We have:

$$\begin{aligned}2qt\varphi(t) + s(1-M) + q(t-1)^2 &= 2qt^2 \log(t) - 2qt^2 + 2qt + s(1-M) + qt^2 - 2qt + q \\ &= 2qt^2 \log(t) - qt^2 + s(1-M) + q \\ &= q\varphi(t^2) + s(1-M) \leq q + s(1-M) \\ &\leq 2(q + s(1-M))(t + \varphi(t)) \\ &= 2(1 - \alpha_s\alpha_f - sM)(t + \varphi(t)).\end{aligned}$$

The new full bound is given by:

$$\mathrm{UCOOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \leq \alpha_s \alpha_t \, \mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + A't + B'\varphi(t),$$

where

$$\begin{aligned}A' &\stackrel{\text{def}}{=} \Delta_\infty(Ms + q) + s(M-1) + \frac{s}{M} \mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) - \frac{s}{M}(\lambda_1 + \lambda_2)(1-M)^2 \\ &\quad + 2(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f - sM) \\ &\leq \Delta_\infty(M+1) + M + \frac{1}{M} \mathrm{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + 2(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f) \stackrel{\text{def}}{=} A \\ B' &\stackrel{\text{def}}{=} 2sM(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f) \leq 2M(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f) \stackrel{\text{def}}{=} B.\end{aligned}$$

In both inequalities, we use the fact that $s \leq 1 - \alpha_s\alpha_f \leq 1$. Using Lemma 1, we obtain

$$\mathrm{UCOOT}(\widetilde{\mathbb{X}_1}, \mathbb{X}_2) \leq \alpha_s \alpha_f \, \mathrm{UCOOT}(\mathbb{X}, \mathbb{Y}) + B \left[1 - \exp\left(-\frac{A}{B} \right) \right].$$

The upper bound of Theorem 2 then follows. ■

Numerical aspects

We claim that, in the discrete setting, by taking ε sufficiently small in the entropic UCOOT problem, we can obtain a solution "close" to the non-entropic case. We formalize this claim and prove it in the following result.

Claim 3. *Let $(\pi^s_\varepsilon, \pi^f_\varepsilon)$ be an equal-mass solution of the problem $\mathrm{UCOOT}_{\lambda,\varepsilon}(\mathbb{X}_1, \mathbb{X}_2)$. Denote $\mu^s = \mu^s_1 \otimes \mu^s_2$ and $\mu^f = \mu^f_1 \otimes \mu^f_2$.*

1. When $\varepsilon \to \infty$, $\pi^s_\varepsilon \rightharpoonup \sqrt{\frac{m(\mu^f)}{m(\mu^s)}} \mu^s$ and $\pi^f_\varepsilon \rightharpoonup \sqrt{\frac{m(\mu^s)}{m(\mu^f)}} \mu^f$.

2. *When $\varepsilon \to 0$, if the spaces $\mathcal{X}_k^s$ and $\mathcal{X}_k^f$ are finite, for $k = 1, 2$, then* $\mathrm{UCOOT}_{\lambda,\varepsilon}(\mathbb{X}_1, \mathbb{X}_2) \to \mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$ *and any cluster point $\widehat{\pi}^s \otimes \widehat{\pi}^f$ of the sequence $(\pi_\varepsilon^s \otimes \pi_\varepsilon^f)_\varepsilon$ will induce an equal-mass solution $(\widehat{\pi}^s, \widehat{\pi}^f)$ of the problem* $\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$. *Furthermore,*

$$\mathrm{KL}(\widehat{\pi}^s \otimes \widehat{\pi}^f | \mu^s \otimes \mu^f) = \min_{(\pi^s, \pi^f)} \mathrm{KL}(\pi^s \otimes \pi^f | \mu^s \otimes \mu^f),$$

where the infimum is taken over all equal-mass solutions of $\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$.

PROOF. Denote $\pi_\varepsilon = \pi_\varepsilon^s \otimes \pi_\varepsilon^f$.

1. When $\varepsilon \to \infty$: the sub-optimality of $\left(\sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s, \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f\right)$ implies

$$\begin{aligned}
\varepsilon\,\mathrm{KL}(\pi_\varepsilon | \mu^s \otimes \mu^f) &\leq F_\lambda(\pi_\varepsilon^s, \pi_\varepsilon^f) + \varepsilon\,\mathrm{KL}(\pi_\varepsilon | \mu^s \otimes \mu^f) \\
&\leq F_\lambda\left(\sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s, \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f\right) + \varepsilon\,\mathrm{KL}(\mu^s \otimes \mu^f | \mu^s \otimes \mu^f) \\
&= \iint |\xi_1 - \xi_2|^p \mathrm{d}\mu^s \mathrm{d}\mu^f.
\end{aligned}$$

Thus,

$$0 \leq \mathrm{KL}(\pi_\varepsilon | \mu^s \otimes \mu^f) \leq \frac{1}{\varepsilon} \iint |\xi_1 - \xi_2|^p \mathrm{d}\mu^s \mathrm{d}\mu^f \to 0,$$

whenever $\varepsilon \to \infty$. We deduce that $\mathrm{KL}(\pi_\varepsilon | \mu^s \otimes \mu^f)$, thus $\pi_\varepsilon \rightharpoonup \mu^s \otimes \mu^f$. The conclusion then follows.

2. Let (π_*^s, π_*^f) be a solution of $\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$. The optimality of $(\pi_\varepsilon^s, \pi_\varepsilon^f)$ implies

$$\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2) \leq \mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2) + \varepsilon\,\mathrm{KL}(\pi_*^s \otimes \pi_*^f | \mu^s \otimes \mu^f).$$

Thus, when $\varepsilon \to 0$, one has $\mathrm{UCOOT}_{\lambda,\varepsilon}(\mathbb{X}_1, \mathbb{X}_2) \to \mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$. Now, for every $\varepsilon > 0$,

$$\begin{aligned}
\langle C, \mu^s \otimes \mu^f \rangle &= F_\lambda\left(\sqrt{\frac{m(\mu^f)}{m(\mu^s)}}\mu^s, \sqrt{\frac{m(\mu^s)}{m(\mu^f)}}\mu^f\right) + \varepsilon\,\mathrm{KL}(\mu^s \otimes \mu^f | \mu^s \otimes \mu^f) \\
&\geq F_\lambda(\pi_\varepsilon^s, \pi_\varepsilon^f) + \varepsilon\,\mathrm{KL}(\pi_\varepsilon^s \otimes \pi_\varepsilon^f | \mu^s \otimes \mu^f) \\
&\geq F_\lambda(\pi_\varepsilon^s, \pi_\varepsilon^f).
\end{aligned}$$

On the other hand, following the same proof in Proposition 3, we can show that if $m(\pi_\varepsilon) \to \infty$, then $F_\lambda(\pi_\varepsilon^s, \pi_\varepsilon^f) \to \infty$, which contradicts the above inequality. So, there exists $M > 0$ such that $m(\pi_\varepsilon) \leq M$, for every $\varepsilon > 0$.

The set $\widetilde{E}_{uco} = \{\pi \in \mathcal{M}^+(\mathcal{S}) : m(\pi) \leq M\} \cap E_{uco}$ is clearly compact, thus from the sequence of minimisers $(\pi_\varepsilon)_\varepsilon \subset \widetilde{E}_{uco}$ (i.e. $\pi_\varepsilon = \pi_\varepsilon^s \otimes \pi_\varepsilon^f$), we can extract a converging subsequence $(\pi_{\varepsilon_n})_{\varepsilon_n}$ such that $\pi_{\varepsilon_n} \to \widehat{\pi} = \widehat{\pi}^s \otimes \widehat{\pi}^f \in \widetilde{E}_{uco}$, with $m(\widehat{\pi}^s) = m(\widehat{\pi}^f)$. The continuity of the divergences implies that, $F_{\lambda,\varepsilon}(\pi_{\varepsilon_n}^s, \pi_{\varepsilon_n}^f) \to F_\lambda(\widehat{\pi}^s, \widehat{\pi}^f)$, when $\varepsilon \to 0$. We deduce that $\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2) = F_\lambda(\widehat{\pi}^s, \widehat{\pi}^f)$, or equivalently $(\widehat{\pi}^s, \widehat{\pi}^f)$ is a solution of $\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$. Moreover, we have

$$\begin{aligned}
0 &\leq F_\lambda(\pi_{\varepsilon_n}^s, \pi_{\varepsilon_n}^f) - F_\lambda(\pi_*^s, \pi_*^f) \\
&\leq \varepsilon_n\Big(\mathrm{KL}(\pi_*^s \otimes \pi_*^f | \mu^s \otimes \mu^f) - \mathrm{KL}(\pi_{\varepsilon_n}^s \otimes \pi_{\varepsilon_n}^f | \mu^s \otimes \mu^f)\Big).
\end{aligned} \tag{5}$$

Dividing by ε_n in 5 and let $\varepsilon_n \to 0$, we have

$$\mathrm{KL}(\widehat{\pi}^s \otimes \widehat{\pi}^f | \mu^s \otimes \mu^f) \leq \mathrm{KL}(\pi_*^s \otimes \pi_*^f | \mu^s \otimes \mu^f).$$

and we deduce that

$$\mathrm{KL}(\widehat{\pi}^s \otimes \widehat{\pi}^f | \mu^s \otimes \mu^f) = \min_{(\pi^s, \pi^f)} \mathrm{KL}(\pi^s \otimes \pi^f | \mu^s \otimes \mu^f),$$

where the infimum is taken over all solutions of $\mathrm{UCOOT}_{\lambda}(\mathbb{X}_1, \mathbb{X}_2)$. ■

Algorithmic details

Optimization procedure

Recall that in discrete form, the UCOOT problem reads

$$\min_{\substack{\pi^s,\pi^f \\ m(\pi^s)=m(\pi^f)}} \sum_{i,j,k,l} (\mathbf{A}_{ik} - \mathbf{B}_{jl})^2 \pi^s_{ij} \pi^f_{kl} + \lambda_1 \operatorname{KL}(\pi^s_{\#1} \otimes \pi^f_{\#1} | u_1) + \lambda_2 \operatorname{KL}(\pi^s_{\#2} \otimes \pi^f_{\#2} | u_2) + \varepsilon \operatorname{KL}(\pi^s \otimes \pi^f | \mu^s_1 \otimes \mu^s_2 \otimes \mu^f_1 \otimes \mu^f_2), \tag{6}$$

where $\pi_{\#1} = (\sum_j \pi_{ij})_i$ and $\pi_{\#2} = (\sum_i \pi_{ij})_j$. Here $\mu_k = \mu^s_k \otimes \mu^f_k$, for $k = 1, 2$. By Proposition 4 in (Séjourné, Vialard, and Peyré 2021), for fixed $\pi^f \in \mathbb{R}^{d_1,d_2}_{\geq 0}$, the minimization in 6 is equivalent to solving the following unbalanced OT problem

$$\min_{\pi \in \mathbb{R}^{n_1,n_2}_{\geq 0}} \langle L_\varepsilon, \pi \rangle + \lambda_1 m_s \operatorname{KL}(\pi_{\#1} | \mu^s_1) + \lambda_2 m_s \operatorname{KL}(\pi_{\#2} | \mu^s_2) + \varepsilon m_s \operatorname{KL}(\pi | \mu^s_1 \otimes \mu^s_2), \tag{7}$$

where $m_s = m(\pi^s)$ and

$$L_\varepsilon := \int |A - B|^2 \mathrm{d}\pi^f + \lambda_1 \langle \log \frac{\pi^f_{\#1}}{\mu^f_1}, \pi^f_{\#1} \rangle + \lambda_2 \langle \log \frac{\pi^f_{\#2}}{\mu^f_2}, \pi^f_{\#2} \rangle + \varepsilon \langle \log \frac{\pi^f}{\mu^f_1 \otimes \mu^f_2}, \pi^f \rangle,$$

and $\int |A - B|^2 \mathrm{d}\pi^f \in \mathbb{R}^{n_1,n_2}$ defined by $\int |A - B|^2 \mathrm{d}\pi^f = A^{\odot 2} \pi^f_{\#1} \oplus B^{\odot 2} \pi^f_{\#2} - 2A\pi^f B^T$. Here, the notations $\otimes$ and $\oplus$ denote the Kronecker product and sum, respectively. For any matrix M, we write $M^{\odot 2} := M \odot M$, where $\odot$ is the element-wise multiplication. The exponential, division and logarithm operations are also element-wise. The scalar product is denoted by $\langle \cdot, \cdot \rangle$.

Now, the problem 7 is of the form

$$\min_{P \geq 0} \langle C, P \rangle + \rho_1 \operatorname{KL}(P_{\#1} | \mu) + \rho_2 \operatorname{KL}(P_{\#2} | \nu) + \varepsilon \operatorname{KL}(P | \mu \otimes \nu),$$

for $\varepsilon, \rho_1, \rho_2 \geq 0$, and can be solved using the scaling algorithm (Chizat et al. 2018b) or non-negative penalized regression (NNPR) (Chapel et al. 2021), depending on the values of parameters. The complete approximation schemes can be found in Algo 2 and Algo 3.

Algorithm 2. Scaling algorithm (Chizat et al. 2018b)

Input: $\mathbf{C} \in \mathbb{R}^{m,n}, \mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}, (\rho_1, \rho_2) \in [0, \infty]^2, \varepsilon > 0$.
Initialize f and g.
repeat
  Update f by: $f = -\frac{\rho_1}{\rho_1 + \varepsilon} \log \sum_j \exp\left(g_j + \log \nu_j - \frac{C_{\cdot,j}}{\varepsilon}\right)$.
  Update g by: $g = -\frac{\rho_2}{\rho_2 + \varepsilon} \log \sum_i \exp\left(f_i + \log \mu_i - \frac{C_{i,\cdot}}{\varepsilon}\right)$.
until convergence
Calculate: $P = (\mu \otimes \nu) \exp\left(f \oplus g - \frac{C}{\varepsilon}\right)$.

Algorithm 3. Non-negative penalized regression (NNPR) (Chapel et al. 2021)

Input: $\mathbf{C} \in \mathbb{R}^{m,n}, \mu \in \mathbb{R}^m_{>0}, \nu \in \mathbb{R}^n_{>0}, (\rho_1, \rho_2) \in [0, \infty)^2, \varepsilon \geq 0$.
Calculate $\lambda = \rho_1 + \rho_2 + \varepsilon$, then $r = \frac{\varepsilon}{\lambda}$ and $\lambda_i = \frac{\rho_i}{\lambda}$, for $i = 1, 2$.
Initialize P.
repeat
  Update P by: $P = \frac{P^{\lambda_1 + \lambda_2}}{P^{\lambda_1}_{\#1} \otimes P^{\lambda_2}_{\#2}} \odot \left(\mu^{\lambda_1 + r} \otimes \nu^{\lambda_2 + r}\right) \odot \exp\left(-\frac{C}{\lambda}\right)$.
until convergence

It should be noted that scaling algorithm allows for $\rho_k = \infty$ (so $\frac{\rho_k}{\rho_k + \varepsilon} = 1$ and we recover the usual Sinkhorn update), but ε must be *strictly* positive. On the other hand, NNPR allows for every $\varepsilon \geq 0$, but both ρ_1 and ρ_2 must be *both finite*.

Complexity UCOOT's complexity has similar complexity to the entropic COOT that had been investigated in the supplementary of (Redko et al. 2020). The latter solves two inner entropic OT problems which implies roughly quadratic complexity, which is also similar to the complexity of solving GW. However, we note that UCOOT can benefit from the recent advance in OT, for example (Schmitzer 2019; Scetbon, Cuturi, and Peyré 2021).

UCOOT helps finding better minima

Interestingly, we find that COOT can achieve better minima when initialized with the UCOOT solutions. Figure S1 illustrates how UCOOT can lead to better alignments by finding better local minima of the COOT transportation cost. For 100 random Gaussian datasets $\mathbf{A}, \mathbf{B}$ with uniformly sampled shapes n_1, n_2, d_1, d_2, we visualize the COOT loss with uniform initialization (y-axis) versus the COOT loss when initializing the COOT BCD with UCOOT's solution. For both $\varepsilon = 0$ and $\varepsilon > 0$, the latter leads to lower COOT costs than the former, on average.

As COOT is a non-convex problem, the choice of initialization plays an important role. Intuitively, by choosing λ_1 and λ_2 sufficiently large, one can use UCOOT to approach COOT. For this reason, the solution of UCOOT can be more informative than the usual uniform initialization, and one can expect to reach better local optimal of COOT.

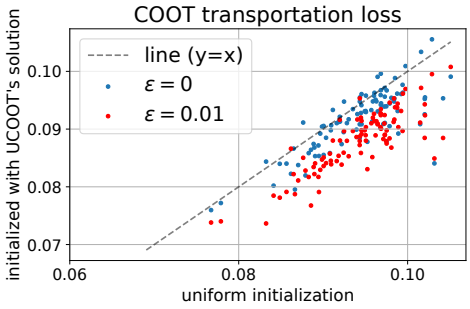

Fig. S1. Scatter plot of the COOT transportation cost with naive (uniform) initialization (y-axis) vs initialization with UCOOT.

Experimental details

More details on barycentric mapping

The barycentric mapping (Ferradans et al. 2014; Courty et al. 2016) is a method to transform the source data to the target domain. Given the source data $X_s \in \mathbb{R}^{n_s \times d_s}$ and target data $X_t \in \mathbb{R}^{n_t \times d_t}$, once the optimal transportation plan $P \in \mathbb{R}^{n_s \times n_t}$ is learned, the transformation of the source to the target domain, can be expressed as: for $i = 1, ..., n_s$,

$$\widehat{x}_i^{(s)} \in \arg\min_{x \in \mathbb{R}^{d_t}} \sum_{j=1}^{n_t} P_{ij}\, c(x, x_j^{(t)}), \tag{8}$$

where the example $x_j^{(t)} \in \mathbb{R}^{d_t}$ corresponds to the j-th row of X_t and the cost $c : \mathbb{R}^{d_t} \times R^{d_t} \to \mathbb{R}$ measures the discrepancy between two examples in $\mathbb{R}^{d_t}$. Typically, c is the squared Euclidean distance, so the problem 8 admits a closed form solution, which is a weighted average of examples in the target domain:

$$\widehat{x}_i^{(s)} = \sum_{j=1}^{n_t} \frac{P_{ij}}{p_i} x_j^{(t)}, \tag{9}$$

where $p_i = \sum_j P_{ij}$, or in matrix notation:

$$\widehat{X}_s = \text{diag}(\frac{1}{P1_{n_t}}) P X_t \in \mathbb{R}^{n_s \times d_t}, \tag{10}$$

where the division is element-wise.

Heterogenous Domain Adaptation (HDA)

More details on label propagation Once the sample coupling P is learned, the label propagation works as follows: suppose the labels contain K different classes, we apply the one-hot encoding to the source label $y^{(s)}$ to obtain $D^{(s)} \in \mathbb{R}^{K \times n_s}$ where $D_{ki}^{(s)} = 1_{\{y_i^{(s)} = k\}}$. The label proportions on the target data are estimated by: $L = D^{(s)} P \in \mathbb{R}^{K \times n_t}$. Then the prediction can be generated by choosing the label with the highest proportion, i.e. $\widehat{y}_j^{(t)} = \arg\max_k L_{kj}$.

Paramater validation We tune the hyperparameters of each method via grid search.

- For COOT, we choose the regularisation on the feature and sample couplings $\varepsilon_f, \varepsilon_s \in \{0, 0.01, 0.1, 0.5\}$.
- For GW, we choose the regularisation parameter $\varepsilon \in \{0, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.
- For UGW and UCOOT, we choose $\lambda_1, \lambda_2 \in \{1, 5, 20, 50\}$ and $\varepsilon \in \{0.01, 0.05, 0.1, 0.5\}$. Furthermore, for UGW and GW, before calculating the Euclidean distance matrix for each domain, the matrix of domain data is normalised by max scaling, so that its coordinates are bounded in $[-1, 1]$. This pre-processing step improves the performance of the for UGW and GW.

For each method, for each combination of tuple of hyperparameters, first, we choose a pair amongst 9 pairs, then repeat 10 times the training procedure, in which the optimal plan is estimated, then used to calculate the accuracy. We choose the tuple of hyperparameters corresponding to the highest average accuracy. This optimal tuple is then applied to all other 8 tasks, where in each task, the training procedure is repeated 10 times and we report the average accuracy.

When there is no regularization In the above hyperparameter tuning process, we only considered $\varepsilon > 0$ for UCOOT and UGW, so that the scaling algorithm (Chizat et al. 2018b) is applicable. As discussed in Section , the NNPR solver can allow us to handle the case $\varepsilon = 0$ (i.e. we can estimate directly UCOOT, rather than via its entropic approximation). In this case, we also tune $\lambda_1, \lambda_2 \in \{1, 50, 20, 50\}$ and follow exactly the same tuning and testing procedure as in the case $\varepsilon > 0$. We report our finding in Table S1. We observe that, in many tasks, the performance remains competitive while enjoying lower variance.

CaffeNet → GoogleNet			
Domains	COOT	UCOOT ($\varepsilon > 0$)	UCOOT ($\varepsilon = 0$)
C → C	36.40 (± 12.94)	**44.05 (± 19.33)**	38.60 (± 9.16)
C → A	28.30 (± 11.78)	**31.90 (± 7.43)**	29.45 (± 9.94)
C → W	19.55 (± 14.51)	28.55 (± 6.60)	**40.85 (± 12.53)**
A → C	**41.80 (± 14.81)**	39.15 (± 17.98)	18.00 (± 9.22)
A → A	**57.90 (± 16.84)**	42.45 (± 15.47)	40.40 (± 8.40)
A → W	42.10 (± 7.80)	48.55 (± 13.06)	**49.15 (± 6.64)**
W → C	8.60 (± 6.56)	**69.80 (± 14.91)**	19.70 (± 5.79)
W → A	16.65 (± 10.01)	**30.55 (± 10.09)**	25.90 (± 5.48)
W → W	**75.30 (± 3.26)**	51.50 (± 20.51)	49.55 (± 6.02)
Average	36.29 (± 10.95)	**42.94 (± 13.93)**	34.62 (± 11.17)

Table S1. Unsupervised HDA from CaffeNet to GoogleNet for $\varepsilon > 0$ and $\varepsilon = 0$. UCOOT ($\varepsilon > 0$) corresponds to the model where ε, λ_1 and λ_2 are tuned, with $\varepsilon > 0$, and UCOOT ($\varepsilon = 0$) means that $\varepsilon = 0$ and only λ_1, λ_2 are tuned.

Sensitivity analysis We report the sensitivity of UCOOT's performance to the hyper-parameters ε, λ_1 and λ_2 for two tasks C→W and A→A in Tables S2, S3 and S4, respectively. In general, the performance depends significantly on the choice of hyperparameters. In Table S2, given fixed values of λ_1 and λ_2, UCOOT performs badly for either too small or large values of ε, indicating that regularization is necessary but should not be too strong. From Table S3, we see that large value of λ_1 degrades the performance, meaning that the marginal constraints on the source distributions should not be too tight. Meanwhile, it seems that large λ_2 is preferable, so the marginal distributions on the target spaces should not be too relaxed.

CaffeNet → GoogleNet							
Domains	$\varepsilon = 0.03$	0.05	0.07	0.1	0.2	0.3	0.4
C → W	27.65 (± 11.34)	37.20 (± 9.35)	34.50 (± 11.07)	34.75 (± 13.04)	17.00 (± 5.92)	18.45 (± 1.11)	11.25 (± 1.66)
A → A	21.95 (± 9.46)	35.30 (± 15.11)	35.65 (± 15.05)	41.15 (± 19.16)	58.45 (± 15.54)	22.30 (± 3.74)	8.90 (± 1.34)

Table S2. Sensitivity of UCOOT to ε in tasks C→W and A→A. We fix $\lambda_2 = 50$ and $\lambda_1 = 1$ and show the accuracy for various value of ε.

CaffeNet → GoogleNet							
Domains	$\lambda_1 = 20$	30	40	50	60	70	80
C → W	35.80 (± 9.33)	34.15 (± 12.98)	37.35 (± 13.82)	27.45 (± 8.33)	32.45 (± 11.62)	30.00 (± 8.04)	30.15 (± 12.89)
A → A	55.20 (± 18.44)	53.35 (± 18.74)	44.15 (± 21.54)	24.30 (± 15.58)	36.10 (± 23.97)	32.35 (± 14.88)	24.80 (± 15.08)

Table S3. Sensitivity of UCOOT to λ_1 in tasks C→W and A→A. We fix $\lambda_2 = 1$ and $\varepsilon = 0.1$ and show the accuracy for various value of λ_1.

CaffeNet → GoogleNet							
Domains	$\lambda_2 = 0.3$	0.5	0.7	1	2	3	4
C → W	34.20 (± 9.83)	34.45 (± 10.80)	34.20 (± 10.50)	34.75 (± 13.04)	29.70 (± 10.55)	37.70 (± 17.96)	32.30 (± 18.81)
A → A	20.75 (± 10.11)	29.00 (± 15.79)	29.25 (± 20.66)	41.15 (± 19.16)	32.65 (± 8.80)	42.10 (± 20.71)	49.95 (± 15.75)

Table S4. Sensitivity of UCOOT to λ_2 in tasks C→W and A→A. We fix $\lambda_1 = 50$ and $\varepsilon = 0.1$ and show the accuracy for various value of λ_2.

Additional results We also perform the adaptation from GoogleNet to CaffeNet. The results can be found in the tables S5. We draw the same conclusions as in the adaptation from CaffeNet to GoogleNet.

GoogleNet → CaffeNet				
Domains	GW	UGW	COOT	UCOOT
C → C	19.45 (± 10.88)	17.50 (± 4.88)	46.20 (± 14.94)	**46.50 (± 5.81)**
C → A	9.35 (± 7.73)	10.50 (± 7.06)	33.25 (± 17.56)	**34.45 (± 4.89)**
C → W	19.15 (± 10.59)	11.95 (± 7.49)	14.95 (± 12.44)	**33.60 (± 10.07)**
A → C	7.90 (± 4.92)	11.70 (± 5.57)	28.80 (± 12.02)	**40.55 (± 6.50)**
A → A	19.75 (± 9.51)	18.40 (± 11.71)	**59.30 (± 20.77)**	58.95 (± 10.37)
A → W	14.55 (± 14.62)	10.05 (± 4.70)	9.75 (± 7.75)	**65.20 (± 9.80)**
W → C	14.05 (± 5.97)	21.95 (± 4.33)	13.70 (± 7.01)	**33.45 (± 6.67)**
W → A	22.85 (± 11.87)	20.90 (± 5.98)	**47.70 (± 5.53)**	44.45 (± 6.02)
W → W	24.10 (± 15.78)	27.95 (± 8.34)	**72.55 (± 4.82)**	68.80 (± 10.24)
Average	16.79 (± 10.21)	16.77 (± 6.67)	36.24 (± 11.43)	**47.33 (± 7.82)**

Table S5. Unsupervised HDA from GoogleNet to CaffeNet.

Multi-omic dataset alignment

Data Preprocessing For the single-cell multi-omics experiments, we use the "PBMC" dataset from Stoeckius *et al* (Stoeckius et al. 2017), accessed on Gene Expression Omnibus (GEO) with the accession code: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE100866GSE100866. This dataset contains a mix of 7,985 mouse and human peripheral blood mononuclear cells (PBMC) and profiles ten antibodies, 17,014 human genes, and 12,915 mouse genes. To pick the human cells, we follow the description in (Stoeckius et al. 2017), and select the cells that have at least 500, and more than 90% of all unique molecular identifiers (UMIs) mapped to the human genes (rather than the mouse genes). From the resulting ~ 4500 cells, we pick the first 1000 to use in our experiments. We use the CLR-normalized antibody count data provided in GEO and apply log normalization to the gene expression data using Seurat package in R to remove biases in sequencing across cells (Hao et al. 2021). Prior to alignment, we follow the existing single-cell alignment methods (Demetci et al. 2022, 2021; Liu et al. 2019; Singh et al. 2020), and also apply L2 normalization to both modalities. The top 50 most variable genes (Figure 5(c)) are selected using the `FindVariableFeatures()` function from Seurat (Hao et al. 2021).

Hyperparameter tuning Hyperparameters were tuned using grid search. For both COOT and UCOOT, we considered the following range for the entropic regularization coefficients $\epsilon_f, \epsilon_s \in \{1e{-}5, 5e{-}5, 1e{-}5, 5e{-}4, ..., 0.1, 0.5\}$. For the mass relaxation coefficients λ_1, λ_2 in UCOOT, the following range was considered $\lambda_1, \lambda_2 \in \{1e{-}3, 5e{-}3, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$. Each combination of hyperparameters were run on three randomly chosen subsets of the dataset that included 30% of the samples and the hyperparameter combinations that on average yielded the highest feature matches and lowest FOSCTTM were picked for the experiments on the full dataset. Below, we list the hyperparameter combinations used for the final alignment results reported in this paper:

- **Balanced scenario of aligning matching features (Figure 5(a)):**
  $\lambda_1 = 1, \lambda_2 = 0.1, \epsilon_1 = 1e-4, \epsilon_2 = 1e-4$
- **Unbalanced scenario of aligning a subset of the matching features (Figure 5(b)):**
  $\lambda_1 = 1, \lambda_2 = 1e-2, \epsilon_1 = 1e-4, \epsilon_2 = 1e-4$
- **Unbalanced scenario of aligning antibodies with the top 50 most variable genes (Figure 5(c)):** $\lambda_1 = 10, \lambda_2 = 5e-5, \epsilon_1 = 1e-4, \epsilon_2 = 0.5$
- **Balanced scenario of aligning the same number of cells (Figure S2(a)):**
  $\lambda_1 = 1, \lambda_2 = 0.1, \epsilon_1 = 1e-4, \epsilon_2 = 1e-4$
- **Unbalanced scenario 1 of aligning different number of cells (Figure S2(b)):**
  $\lambda_1 = 0.01, \lambda_2 = 0.1, \epsilon_1 = 5e-3, \epsilon_2 = 1e-4$
- **Unbalanced scenario 2 of aligning different number of cells (Figure S2(c)):**
  $\lambda_1 = 0.01, \lambda_2 = 0.1, \epsilon_1 = 5e-3, \epsilon_2 = 1e-4$

Further investigation of the feature alignments In the unbalanced experiment, where we align the most variable genes and the antibodies, we expect a well-performing alignment method to correctly match antibodies with the genes that express them. This would be the strongest biological connection between a protein (i.e. an antibody, in this case) and a gene. However, other biological connections can also exist, such as between an antibody and a gene that regulates the expression of that antibody, a gene that codes for a protein the antibody physically interacts with, or a gene that codes for a protein that is active in the same biological pahtway as the antibody of interest. To investigate whether there are such matches recovered outside of the ten genes we label as "matching genes", we refer to two gene regulatory network databases that contain data on human PBMCs, GRNdb (Fang et al. 2020) and GRAND (Ben Guebila et al. 2021) (for the first kind of relationship), two protein–protein interaction databases, BioGRID (Oughtred et al. 2020), and STRING (Szklarczyk et al. 2020) (for the second kind of relationship), and KEGG (Kanehisa et al. 2020), a database of biological pathways (for the last kind of relationship).

Of the 46 correspondences yielded by COOT, and 16 by UCOOT, outside of the 'correspondences with 'matching genes", only a few show up on these databases:

- **CD19 antibody correspondences:** Both BIOGRID (Oughtred et al. 2020) and STRING (Szklarczyk et al. 2020) databases return a physical interaction with CD79A protein (encoded by the *CD79A* gene), which is experimentally validated by affinity capture-Western (Carter et al. 1997). The correspondence with *CD79A* is yielded by both UCOOT and COOT. Additionally, according to KEGG (Kanehisa et al. 2020), CD19 participates in the B-cell receptor (BCR) signaling pathway along with IGH, which is formed by multiple segments joining together, including IGHD and IGHM [1]. COOT yields correspondences with the genes that code for these.

- **CD57 antibody correspondences:** There is an experimentally validated physical interaction with ITM2C (encoded by the *ITM2C* gene), which shows up on BIOGRID. This interaction has been validated using proximity labeling mass spectrometry (Go et al. 2021). *ITM2C* is among the correspondences yielded by COOT.

- **CD2 antibody correspondences:** According to the BIOGRID database, a physical interaction between CD2 and PTPRC has been proposed via an *in vitro* study *et al* (Schraven et al. 1990). *PTPRC* shows up among the correspondences yielded by both UCOOT and COOT for the CD2 antibody.

- **CD4 antibody correspondences:** According to BIOGRID, CD4 has been shown to physically interact with TUBB using affinity capture mass spectrometry by Bernhard *et al* (Bernhard, Cunningham, and Sheil 2004). TUBB is a component of the tubulin protein, which made out of $\beta-$tubulin (TUBB) and $\alpha-$tubulin (TUBA). UCOOT yields a correspondence between CD4 and TUBA1B (gene that codes for a component of the $\alpha-$tubulin) [2].

Outside of these, no other correspondences returned biological relevance based on our database and literature search, which leads us to conclude COOT yields more redundant correspondences that UCOOT.

Sample alignment experiments Below in Figure S2, we visualize the aligned samples (first two principal components of the two domains together upon barycentric projection) and report the alignment performance as measured by the "average fraction of samples closer than true match (FOSCTTM)" and "label transfer accuracy" metrics. We borrow these metrics from previously published single-cell multi-omic data alignment methods (Liu et al. 2019; Cao et al. 2019; Demetci et al. 2022; Cao, Hong, and Wan 2021; Demetci et al. 2021). For label transfer accuracy, we follow the previously published methods (Cao et al. 2019; Cao, Hong, and Wan 2021; Demetci et al. 2022, 2021) and train a $k-$NN classifier (for $k = 5$) on the cell-type labels of the measurement domain with the full set of cells, and apply it to predict the cell-type labels of the downsampled domain. We report the prediction accuracy. For the balanced scenario, we train the classifier on the antibody domain to predict the labels in the gene expression domain upon transportation. For the average FOSCTTM metric used in unbalanced scenarios, we use the cells that remain to have a correspondence after subsampling to calculate the FOSCTTM scores. We note that lower average FOSCTTM and higher label transfer accuracy results indicate better alignments.

[1] https://www.genecards.org/cgi-bin/carddisp.pl?gene=IGHD, and https://www.genecards.org/cgi-bin/carddisp.pl?gene=IGH&keywords=IGH

[2] https://www.nature.com/scitable/content/microtubules-the-basics-14673338/

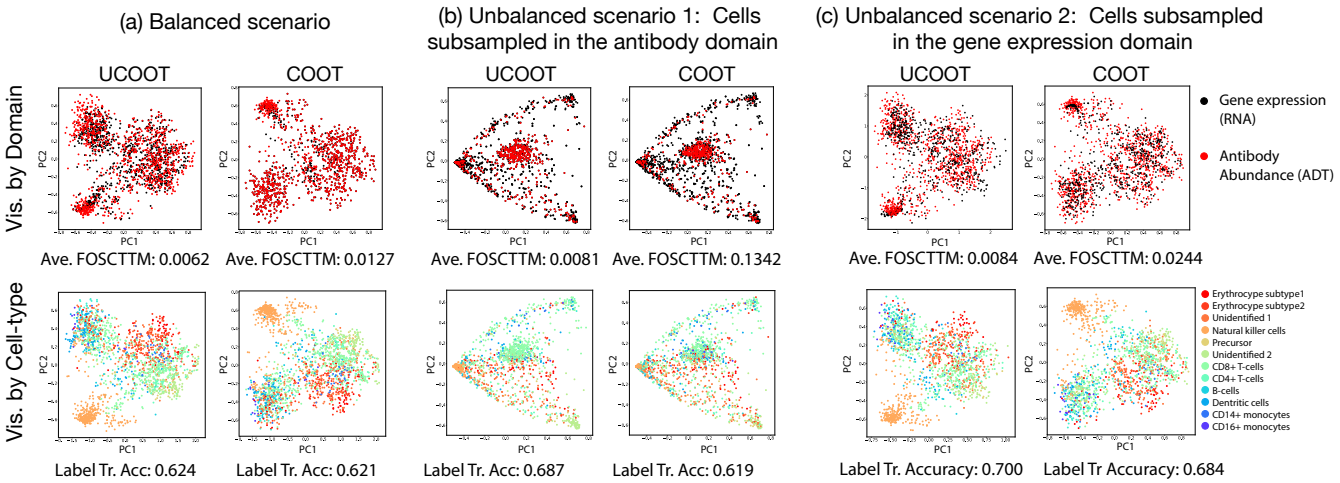

Fig. S2. Visualization of the sample alignments with UCOOT and COOT after barycentric projection (First two principal components). The top row visualizes results with samples colored based on measurement modality (black points show the gene expression domain samples, and red points show the antibody domain samples). The bottom row visualizes alignments with samples colored based on cell-type labels. **(a)** presents sample alignments in the balanced scenario, where we align the same number of cells (1000) in each measurement modality with the matching features (same scenario as Fig 6 (a), but presenting sample alignments). **(b)** In this unbalanced scenario, we randomly downsample the cells in antibody domain by 25%. **(c)** In this second unbalanced scenario, we randomly downsample the cells in the gene expression domain by 25% and align with the full set of samples in the antibody domain. For all alignments, we quantify alignment quality using average FOSCTTM ("Ave. FOSCTTM") and label transfer accuracy ("Label Tr. Acc."), and report them under the plots. We calculate both metrics prior to applying dimensionality reduction with principal component analysis (PCA). PCA is only applied for visualization purposes. Note that the overall increase in label transfer accuracy between **a-c** is likely due to the removal of groups of heterogenous cell types during downsampling.