

Unbalanced CO-optimal Transport

Quang Huy Tran^{1,2}, Hicham Janati³, Nicolas Courty¹, Rémi Flamary²,
Ievgen Redko⁴, Pinar Demetci^{5,6}, Ritambhara Singh^{5,6}

¹Université Bretagne Sud, IRISA

²CMAP, Ecole Polytechnique, IP Paris

³LTCI, Télécom Paris, IP Paris

⁴Univ. Lyon, UJM-Saint-Etienne, CNRS, UMR 5516

⁵Center for Computational Molecular Biology, Brown University

⁶Department 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 (Fratras 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; Fratas 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 $[1, n] := \{1, \dots, n\}$. 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 $\text{KL}(\mu|\nu) \stackrel{\text{def}}{=} \int \frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu} d\nu - \int d\mu + \int d\nu$ 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 $(\pi_{\#1}^s, \pi_{\#2}^s)$ close to (μ_1^s, μ_2^s) according to some appropriate divergence D such that the cost $\int c(x_1, x_2) d\pi^s(x_1, x_2) + D(\pi_{\#1}^s | \mu_1^s) + D(\pi_{\#2}^s | \mu_2^s)$ 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 $\iint c((x_1^s, x_1^f), (x_2^s, x_2^f)) d\pi^s(x_1^s, x_2^s) d\pi^f(x_1^f, x_2^f)$ 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, μ^s) and (\mathcal{X}^f, μ^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:*

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

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

	OT	UOT	GW	UGW	COOT	UCOOT
Across spaces	\times	\times	\checkmark	\checkmark	\checkmark	\checkmark
Sample alignment	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark
Feature alignment	\times	\times	\times	\times	\checkmark	\checkmark
Robustness to outliers	\times (Fratras et al. 2021)	\checkmark (Fratras et al. 2021)	\times (Prop. 2)	\checkmark (Thm. 1)	\times (Prop. 2)	\checkmark (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 $D_k(\pi_{\#k}^s \otimes \pi_{\#k}^f | \mu_k^s \otimes \mu_k^f)$, one can consider $D_k(\pi_{\#k}^s | \mu_k^s) + D_k(\pi_{\#k}^f | \mu_k^f)$, or $D_s(\pi_{\#1}^s \otimes \pi_{\#2}^s | \mu_1^s \otimes \mu_2^s)$, 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 theoretical 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_1^s, μ_2^s) and (μ_1^f, μ_2^f) 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 $D_1 = D_2 = \text{KL}$. We write $\text{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2)$ 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 $\llbracket 1, n \rrbracket$ (resp. $\llbracket 1, d \rrbracket$), 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 $D_1, D_2 \in \{\iota_{=}, \text{KL}\}$, 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 $(\mathcal{X}_1^s, \mathcal{X}_2^s)$ 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 $\text{UCOOT}_\lambda(\mathbb{X}_1, \mathbb{X}_2) = 0$ 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 $\mathbb{X}_k = ((\mathcal{X}_k^s, \mu_k^s), (\mathcal{X}_k^f, \mu_k^f), \xi_k)$, 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 $\tilde{\mu}^a = \alpha_a \mu^a + (1 - \alpha_a) \varepsilon^a$, where $\alpha_a \in [0, 1]$. We assume that ξ_1 is defined on $(\mathcal{X}_1^s \cup \mathcal{O}^s) \times (\mathcal{X}_1^f \cup \mathcal{O}^f)$ and that ξ_1, ξ_2 are continuous on their supports. We denote the contaminated sample-feature space by $\tilde{\mathbb{X}}_1 = ((\mathcal{X}_1^s \cup \mathcal{O}^s, \tilde{\mu}_1^s), (\mathcal{X}_1^f \cup \mathcal{O}^f, \tilde{\mu}_1^f), \xi_1)$. Finally, we define some useful minimal and maximal costs:

$$\Delta_0 \stackrel{\text{def}}{=} \min_{\substack{x_1^s \in \mathcal{O}^s, x_1^f \in \mathcal{O}^f \\ x_2^s \in \mathcal{X}_2^s, x_2^f \in \mathcal{X}_2^f}} |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p$$

and

$$\Delta_\infty \stackrel{\text{def}}{=} \max_{\substack{x_1^s \in \mathcal{X}_1^s \cup \mathcal{O}^s, x_1^f \in \mathcal{X}_1^f \cup \mathcal{O}^f \\ x_2^s \in \mathcal{X}_2^s, x_2^f \in \mathcal{X}_2^f}} |\xi_1(x_1^s, x_1^f) - \xi_2(x_2^s, x_2^f)|^p$$

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.

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	$\pi^f = \mathbf{I}_{d_1} = \mathbf{I}_{d_2}$	$\pi^f = \pi^s$	–	–	–
Scalar function	$\xi(i, j) = \mathbf{A}_{ij}$	$\xi(i, j) = \text{dist}(\mathbf{A}_i, \mathbf{A}_j)$	$\xi(i, j) = \mathbf{A}_{ij}$	$\xi(\mathbf{a}, j) = \mathbf{a}_j$	$\xi(i, j) = \mathbf{A}_{ij}$
Divergence	$l =$	$l =$	$l =$	$l =$	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 $\mu_d = \mathbf{1}_d/d$. Here, \mathbf{I}_d is the identity matrix in \mathbb{R}^d .

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.

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

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

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 $\delta \stackrel{\text{def}}{=} 2(\lambda_1 + \lambda_2)(1 - \alpha_s\alpha_f)$ and $K = M + \frac{1}{M} \text{UCOOT}(\mathbb{X}_1, \mathbb{X}_2) + \delta$, where $M = m(\pi^s) = m(\pi^f)$ is the transported mass between clean data. Then:

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

The proof of Theorem 1 is provided in the Appendix and inspired from (Fratras 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 $\tau \mathbf{1}_{15}$, 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.

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.

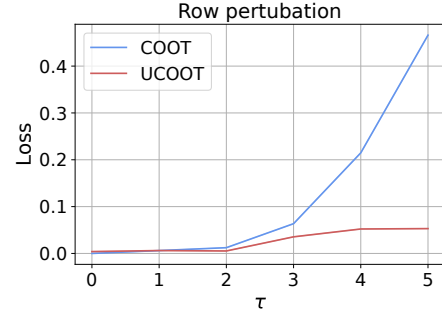


Figure 1: Sensitivity of COOT and UCOOT under the presence of outliers.

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: $u_k \stackrel{\text{def}}{=} \frac{1}{n_k d_k} \mathbf{1}_{n_k} \otimes \mathbf{1}_{d_k}$, 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:

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

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-

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^f using Sinkhorn or NNPR
 Rescale $\pi^f = \sqrt{\frac{m(\pi^s)}{m(\pi^f)}} \pi^f$
until convergence

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 2 (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 2 (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 3 answers this question positively. For $\sigma > 0$, we compute the average accuracy (defined by the percentage of mass within the block-diagonal struc-

ture) 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.

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 represented 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.

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

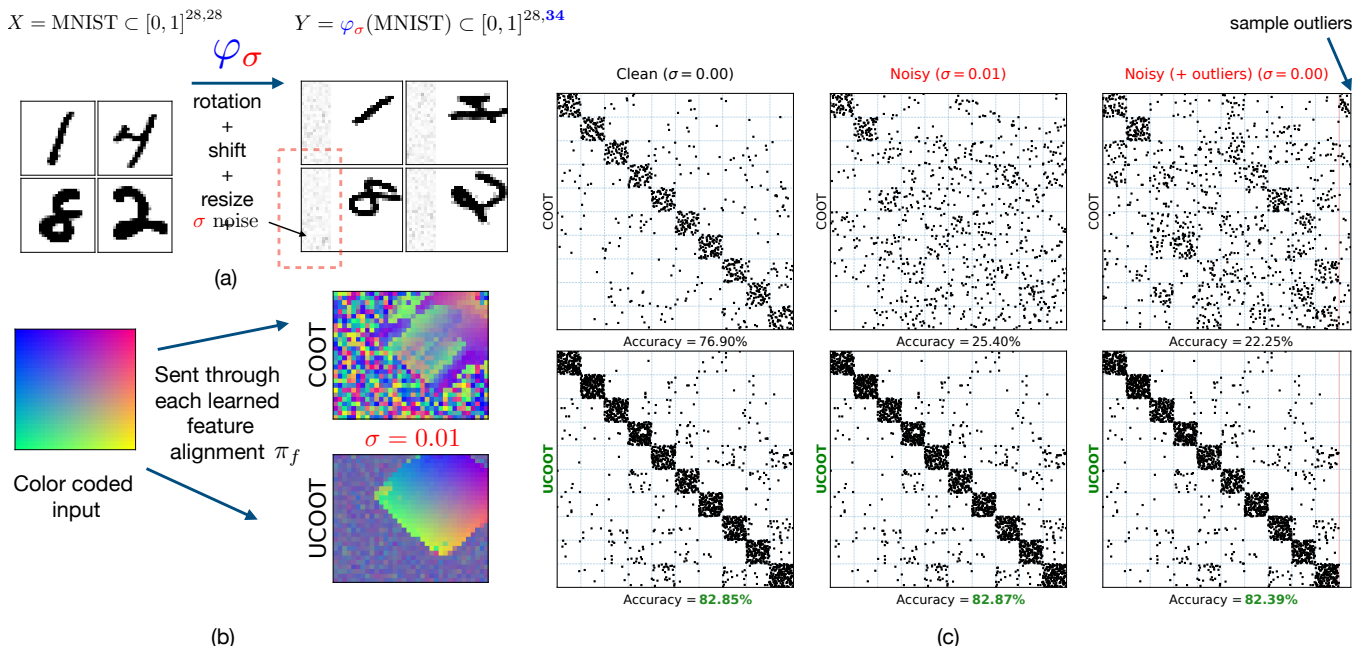


Figure 2: Example illustrating the feature alignment π_f learned by UCOOT and its robustness to outliers. (a) Visualization of 4 random samples from both datasets. The added Gaussian noise only affects the first 10 columns of the images and is different across images. (b) The barycentric mapping (see 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.

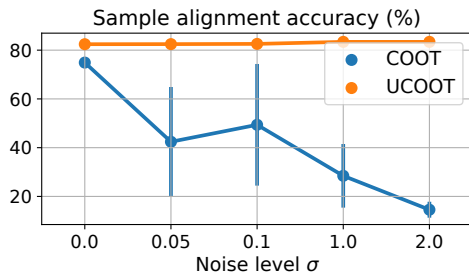


Figure 3: Robustness of UCOOT vs. COOT on MNIST example. The accuracy is averaged on 20 trials.

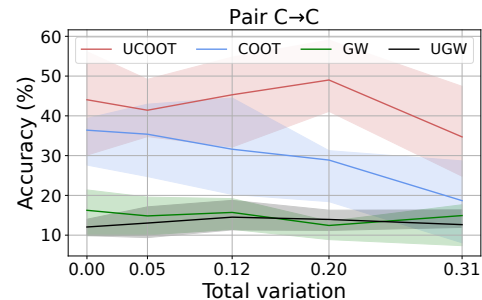


Figure 4: Robustness to class proportion change for increasing TV on the class marginals.

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

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.

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 to 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

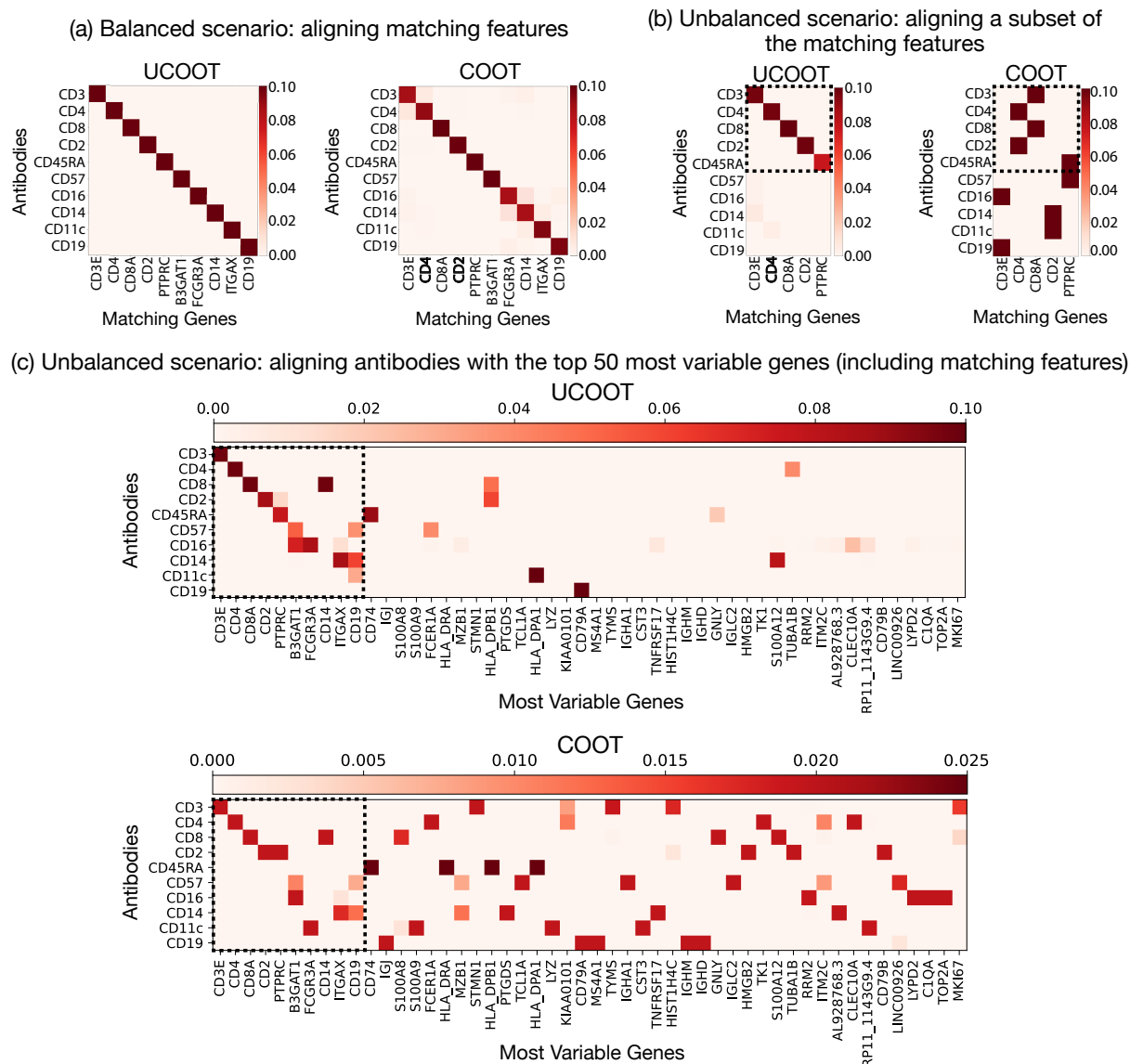


Figure 5: Feature alignments on the single-cell multi-omics dataset of COOT and UCOOT between antibodies (surface proteins) and their matching genes (that encode them). (a) The features are sorted such that the correct alignment would yield a diagonal matrix. (b) Only five of the correct gene matches 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.

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.

Acknowledgements

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 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.
- 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, K.; Hong, Y.; and Wan, L. 2021. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamon. *Bioinformatics*. Btab594.
- 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*.
- Fatras, K.; Séjourné, T.; Courty, N.; and Flamary, R. 2021. Unbalanced minibatch Optimal Transport; applications to Domain Adaptation. *International Conference on Machine Learning*.
- 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.
- 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.
- 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.
- 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 37th 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.
- 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.
- 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.
- 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*.