Unbalanced minibatch Optimal Transport; applications to Domain Adaptation

Supplementary material

Outline. The supplementary material of this paper is organized as follows:

- In section A, we first review the formalism with definitions and basic property proofs.
- In section B, we demonstrate our statistical and optimization results.
- In section C, we give extra experiments and details for domain adaptation experiments.

A. Minibatch UOT formalism and basic properties

We start with the rigorous formalism of the minibatch UOT transport plan.

A.1. Minibatch UOT plan formalism

Definition 1. *We denote by* Opt_h *the set of all optimal transport plans for* $h = \mathrm{OT}_\phi^{\tau,\varepsilon}$*, cost matrix* C *and a marginal* $\boldsymbol{u}$*. Let* $\mathbf{u}_m, \mathbf{u}_m \in (\mathbb{R}^m)^2$ *be discrete positive uniform vectors. For each pair of index* m*-tuples* $I = (i_1, \dots, i_m)$ *and* $J = (j_1, \dots, j_m)$ *from* $[\![1, n]\!]^m$*, consider* $C' := C_{I,J}$ *the* $m \times m$ *matrix with entries* $C'_{k\ell} = C_{i_k, j_\ell}$ *and denote by* $\Pi^m_{I,J}$ *an arbitrary element of* Opt_h*. It can be lifted to an* $n \times n$ *matrix where all entries are zero except those indexed in* $I \times J$*:*

$$\Pi_{I,J} = Q_I^\top \Pi^m_{I,J} Q_J \tag{1}$$

where Q_I *and* Q_J *are* $m \times n$ *matrices defined entrywise as*

$$(Q_I)_{ki} = \delta_{i_k, i}, 1 \leq k \leq m, 1 \leq i \leq n \tag{2}$$

$$(Q_J)_{\ell j} = \delta_{j_\ell, j}, 1 \leq \ell \leq m, 1 \leq j \leq n. \tag{3}$$

Each row of these matrices is a Dirac vector, hence they satisfy $Q_I \mathbf{1}_n = \mathbf{1}_m$ *and* $Q_J \mathbf{1}_n = \mathbf{1}_m$*.*

We also define the *averaged minibatch transport matrix* which takes into account all possible minibatch couples.

Definition 2 (Averaged minibatch transport matrix)**.** *Consider* $h = \mathrm{OT}_\phi^{\tau,\varepsilon}$*. Given data* n*-tuples* $\boldsymbol{X}, \boldsymbol{Y}$*, consider for each pair of* m*-tuples* I*,* J*, the uniform vector of size* $m, \mathbf{u}_m$*, and let* $\Pi_{I,J}$ *be defined as in Definition 1.*

The averaged minibatch transport matrix and its incomplete variant are :

$$\overline{\Pi}^m(\boldsymbol{X}, \boldsymbol{Y}) := \frac{(n-m)!^2}{n!^2} \sum_{I \in \mathcal{P}^m} \sum_{J \in \mathcal{P}^m} \Pi_{I,J}., \tag{4}$$

$$\widetilde{\Pi}^m_k(\boldsymbol{X}, \boldsymbol{Y}) := k^{-1} \sum_{(I,J) \in D_k} \Pi_{I,J}, \tag{5}$$

where D_k *is a set of cardinality* k *whose elements are drawn at random from the uniform distribution on* $\Gamma := \mathcal{P}_m \times \mathcal{P}_m$*.*

The average in the above definition is always finite so we do not need to concern ourselves with the measurability of selection of optimal transport plans. The same will be true whenever an average of optimal transport plans will be taken in the rest of this paper, since all results concerning such averages will be nonasymptotic. We will therefore avoid further mentioning this issue, for the sake of brevity. Unfortunately, on contrary to the balanced case, the minibatch UOT transport plan do not define OT transport plan as they do not respect the marginals, so in general the averaged minibatch UOT is not an OT transport plan. Note that the Sinkhorn divergence involves three terms, which explains why we can not define an associated averaged minibatch transport matrix.

A.2. Basic properties

Proposition 1 (Positivity, symmetry and bias)**.** *The minibatch UOT are positive and symmetric losses. However, they are not definites,* i.e., $\bar{h}^m(\boldsymbol{X}, \boldsymbol{X}) > 0$ *for non trivial* $\boldsymbol{X}$ *and* $1 < m < n$.

Proof. The first two properties are inherited from the classical UOT cost. Consider a uniform probability vector and random 3-data tuple $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ with distinct vectors $(\boldsymbol{x}_i)_{1 \leq i \leq 3}$. As $\bar{h}^m$ is an average of positive terms, it is equal to 0 if and only if each of its term is 0. But consider the minibatch term $I_1 = (i_1, i_2)$ and $I_2 = (i_1, i_3)$, then obviously $h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}(I_1), \boldsymbol{X}(I_2))) \neq 0$ as $\boldsymbol{x}_2 \neq \boldsymbol{x}_3$, where $\boldsymbol{X}(I_1)$ denotes the data minibatch corresponding to indices in I_1. □

We now give the proof for our claim "A simple combinatorial argument assures that the sum of $\boldsymbol{u}_m$ over all m-tuples I gives $\boldsymbol{u}_n$."

Proposition 2 (Averaged distributions)**.** *Let* $\boldsymbol{u}_m$ *be a uniform vector of size* m*. The average over* m*-tuples* $I \in \mathcal{P}^m$ *for a given index of* $\boldsymbol{u}_m$ *is equal to* $\frac{m_{\boldsymbol{a}}}{n}$*, i.e.,* $\forall i \in [\![1, n]\!], \sum_{I \in \mathcal{P}^m} (\boldsymbol{u}_m)_i = (\boldsymbol{u}_n)_i = \frac{m_{\boldsymbol{a}}}{n}$.

Proof. We recall that $\mathcal{P}^m$ denotes the set of all m-tuples without repeated elements. Let us check we recover the initial weights $(\boldsymbol{u}_n)_i = \frac{m_{\boldsymbol{a}}}{n}$. Observe that $\sum_{i=1}^n a_i = m_{\boldsymbol{a}}$ and that for each $1 \leq i \leq n$

$$\begin{aligned}
\sharp\{I \in \mathcal{P}^m : i \in I\} &= \sharp\{I \in \mathcal{P}^m : n \in I\} \\
&= \sharp\{I = (i_1, \ldots, i_m) \in \mathcal{P}^m : i_1 = n\} + \ldots + \sharp\{I = (i_1, \ldots, i_m) \in \mathcal{P}^m : i_m = n\} \\
&= m \cdot \sharp\{I = (i_1, \ldots, i_m) \in \mathcal{P}^m : i_m = n\}. \qquad (6)
\end{aligned}$$

Since $\sharp\{I = (i_1, \ldots, i_m) \in \mathcal{P}^m : i_m = n\}$ is the number of $(m-1)$-tuples without repeated indices of $[\![1, n-1]\!]$, $(n-1)!/(n-m)!$, it follows that

$$\begin{aligned}
\frac{(n-m)!}{n!} \cdot \sum_{I \in \mathcal{P}^m} \frac{m_{\boldsymbol{a}}}{m} 1_I(i) &= \frac{(n-m)!}{n!} \sum_{I \in \mathcal{P}^m, i \in I} \frac{m_{\boldsymbol{a}}}{m} = \frac{(n-m)!}{n!} \frac{m_{\boldsymbol{a}}}{m} \cdot \sharp\{I \in \mathcal{P}^m : i \in I\} \qquad (7) \\
&= \frac{(n-m)!}{n!} \frac{m_{\boldsymbol{a}}}{m} m \cdot \frac{(n-1)!}{(n-m)!} = \frac{m_{\boldsymbol{a}}}{n} \qquad (8)
\end{aligned}$$

□

B. Proof main results

In this section we prove the UOT properties and the minibatch statistical and optimization theorems. We start with UOT properties as they are necessary to derive the minibatch results.

B.1. Unbalanced Optimal Transport properties

We recall the definition of Csiszàr divergences. Consider a convex, positive, lower-semicontinuous function such that $\phi(1) = 0$. Define its recession constant as $\phi'_\infty = \lim_{x \to +\infty} \phi(\boldsymbol{x})/\boldsymbol{x}$. The Csiszàr divergence between positively weighted vectors $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^d$ reads

$$D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{y}_i \neq 0} \boldsymbol{y}_i \phi\Big(\frac{\boldsymbol{x}_i}{\boldsymbol{y}_i}\Big) + \phi'_\infty \sum_{\boldsymbol{y}_i = 0} \boldsymbol{x}_i.$$

It allows to generalize OT programs. We retrieve common penalties such as Total Variation and Kullback-Leibler divergence by respectively taking $\phi(\boldsymbol{x}) = |\boldsymbol{x} - 1|$ and $\phi(\boldsymbol{x}) = (\boldsymbol{x} \log \boldsymbol{x} - \boldsymbol{x} + 1)$. We provide a generalized definition of all OT programs as

$$\mathrm{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{a}, \boldsymbol{b}, C) = \min_{\Pi \in \mathbb{R}_+^{n \times n}} \mathcal{F}(\Pi, C) = \min_{\Pi \in \mathbb{R}_+^{n \times n}} \langle C, \Pi \rangle + \tau D_\phi(\Pi \mathbf{1}_n | \boldsymbol{a}) + \tau D_\phi(\Pi^T \mathbf{1}_n | \boldsymbol{b}) + \varepsilon \mathtt{KL}(\Pi | \boldsymbol{a} \otimes \boldsymbol{b}).$$

Where $\mathcal{F}$ denotes the UOT energy.

B.1.1. ROBUSTNESS

We start by showing the robustness properties lemma 1 that we split in two different lemmas. Lemma 1.1 shows that the UOT cost is robust to an outlier while lemma 1.2 shows that OT is not robust to an outlier.

Lemma 1.1. *Take (μ, ν) two probability measures with compact support, and z outside of ν's support. Recall the Gaussian-Hellinger distance (Liero et al., 2017) between two positive measures as*

$$\mathrm{GH}_\tau(\mu,\nu) = \inf_{\pi \geq 0} \int C(x,y) d\pi(x,y) + \tau KL(\pi_1|\mu) + \tau KL(\pi_2|\nu).$$

For $\zeta \in [0,1]$, write $\tilde{\mu} = \zeta\mu + (1-\zeta)\delta_z$ a measure perturbed by a Dirac outlier. Write $m(z) = \int C(z,y)d\nu(y)$ One has

$$\mathrm{GH}_\tau(\tilde{\mu},\nu) \leq \zeta\, \mathrm{GH}_\tau(\mu,\nu) + 2\tau(1-\zeta)(1-e^{-m(z)/2\tau}) \tag{9}$$

In particular, with the notation $\mathrm{OT}^{\tau,0}_{KL}$ it reads

$$\mathrm{OT}^{\tau,0}_{KL}(\tilde{\mu},\nu,C) \leq \zeta\, \mathrm{OT}^{\tau,0}_{KL}(\mu,\nu,C) + 2\tau(1-\zeta)(1-e^{-m(\boldsymbol{z})/2\tau})$$

Proof. Recall that the OT program reads

Write π the optimal plan for $\mathrm{OT}^{\tau,0}_{\phi}(\mu,\nu)$. We consider a suboptimal plan for $\mathrm{OT}^{\tau,0}_{\phi}(\tilde{\mu},\nu)$ of the form

$$\tilde{\pi} = \zeta\pi + (1-\zeta)\kappa\delta_z \otimes \nu,$$

where κ is mass parameter which will be optimized after. Note that the marginals of the plan $\tilde{\pi}$ are $\tilde{\pi}_1 = \zeta\pi_1 + (1-\zeta)\kappa\delta_z$ and $\tilde{\pi}_2 = \zeta\pi_2 + (1-\zeta)\kappa\nu$. Note that $\mathtt{KL}$ is jointly convex, thus one has

$$\begin{aligned}
\mathtt{KL}(\tilde{\pi}_1|\tilde{\mu}) &\leq \zeta\mathtt{KL}(\pi_1|\mu) + (1-\zeta)\mathtt{KL}(\kappa\delta_z|\delta_z),\\
\mathtt{KL}(\tilde{\pi}_2|\tilde{\mu}) &\leq \zeta\mathtt{KL}(\pi_2|\nu) + (1-\zeta)\mathtt{KL}(\kappa\nu|\nu).
\end{aligned}$$

Thus a convex inequality yields

$$\begin{aligned}
\mathrm{OT}^{\tau,0}_{\phi}(\tilde{\mu},\nu) \leq \zeta\Big[\int ||x-y|| d\pi(x,y) + \tau\mathtt{KL}(\pi_1|\mu) + \tau\mathtt{KL}(\pi_2|\nu)\Big]\\
+ (1-\zeta)\Big[\kappa m(z) + \tau\mathtt{KL}(\kappa\delta_z|\delta_z) + \tau\mathtt{KL}(\kappa\nu|\nu)\Big].
\end{aligned}$$

We optimize now the upper bound w.r.t. κ. Both $\mathtt{KL}$ terms are equal to $\phi(\kappa) = \kappa\log\kappa - \kappa + 1$, thus differentiating w.r.t. κ yields

$$m(z) + 2\tau\log\kappa = 0 \Rightarrow \kappa = e^{-m(z)/2\tau}.$$

Reusing this expression of κ in the upper bound yields Equation (9). □

Lemma 1.2. *Take (μ, ν) two probability measures with compact support, and z outside of ν's support. Define the Wasserstein distance between two probabilities as*

$$\mathrm{W}(\mu,\nu) = \sup_{f(x)+g(y)\leq C(x,y)} \int f(x)d\mu(x) + \int g(y)d\nu(y). \tag{10}$$

For $\zeta \in [0,1]$, write $\tilde{\mu} = \zeta\mu + (1-\zeta)\delta_z$ a measure perturbed by a Dirac outlier. Write (f,g) the optimal dual potentials of $\mathrm{W}(\mu,\nu)$, and y^* a point in ν's support. One has

$$\mathrm{W}(\tilde{\mu},\nu) \geq \zeta\, \mathrm{W}(\mu,\nu) + (1-\zeta)\Big(||z-y^*||^2 - g(y^*) + \int g d\nu\Big) \tag{11}$$

In particular, with the notation $\mathrm{OT}^{\infty,0}_{KL}$ it reads

$$\mathrm{OT}^{\infty,0}_{\phi}(\tilde{\mu},\nu) \geq \zeta\, \mathrm{OT}^{\infty,0}_{\phi}(\mu,\nu) + (1-\zeta)\Big(C(\boldsymbol{z},y^*) - g(y^*) + \int g d\beta\Big)$$

Proof. We consider a suboptimal pair $(\tilde{f}, \tilde{g})$ of potentials for $\mathrm{OT}_\phi^{\infty,0}(\tilde{\mu}, \nu)$. On the support of (μ, ν) we take the optimal potentials pair (f, g) for (μ, ν), i.e. $\tilde{f} = f$ and $\tilde{g} = g$. We need to extend $\tilde{f}$ at z. To do so we take the c-transform of g, i.e.

$$\tilde{f}(z) = \inf_{y \in spt(\nu)} ||z - y||^2 - g(y) = ||z - y^*||^2 - g(y^*),$$

where the infimum is attained at some y^* since ν has compact support. the pair $(\tilde{f}, \tilde{g})$ is suboptimal, thus

$$\begin{aligned}
\mathrm{OT}_\phi^{\infty,0}(\tilde{\mu}, \nu) &\geq \int \tilde{f}(x) d\tilde{\mu}(x) + \int \tilde{g}(y) d\nu(y) \\
&\geq \zeta \int f(x) d\mu(x) + (1-\zeta)\tilde{f}(z) + \int \tilde{g}(y) \\
&\geq \zeta\, \mathrm{OT}_\phi^{\infty,0}(\mu, \nu) + (1-\zeta)\Big[||z - y^*||^2 - g(y^*) + \int g(y) d\nu(y)\Big]
\end{aligned}$$

Hence the resulted given by Equation (11). □

B.1.2. UOT PROPERTIES

Now let us present results which will be useful for concentration bounds. A key element is to have a bounded plan and a finite UOT cost in order to derive a hoeffding type bound. We start this section by proving lemma 2. We split it in two, lemma 2.1 proves that the UOT cost is finite and provides an upper bound while lemma 2.2 proves that the UOT plan exists and belongs to a compact set.

Lemma 2.1 (Upper bounds)**.** *Let $(\boldsymbol{a}, \boldsymbol{b})$ be two positive vectors and assume that $\langle \boldsymbol{a}\boldsymbol{b}^\top, C\rangle < +\infty$, then the UOT cost is finite. Furthermore, we have the following bound for $h = \mathrm{OT}_\phi^{\tau,\varepsilon}$, one has $|h(\boldsymbol{a}, \boldsymbol{b}, C)| \leqslant M^h_{\boldsymbol{a},\boldsymbol{b}}$, where*

$$M^h_{\boldsymbol{a},\boldsymbol{b}} = M m_{\boldsymbol{a}} m_{\boldsymbol{b}} + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{b}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{a}}). \tag{12}$$

Regarding $h = S_\phi^{\tau,\varepsilon}$, one has $|h(\boldsymbol{a}, \boldsymbol{b}, C)| \leqslant M^S_{\boldsymbol{a},\boldsymbol{b}}$, where

$$M^S_{\boldsymbol{a},\boldsymbol{b}} = 2M m_{\boldsymbol{a}} m_{\boldsymbol{b}} + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{b}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{a}}) + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{a}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{b}}) + \frac{\varepsilon}{2}(m_{\boldsymbol{a}} - m_{\boldsymbol{b}})^2. \tag{13}$$

Proof. As $\langle \boldsymbol{a}\boldsymbol{b}^\top, C\rangle < +\infty$ is finite, one can bound the ground cost C as $0 \leqslant C_{i,j} \leqslant M$. Consider the OT kernel $h \in \{\mathrm{OT}_\phi^{\tau,\varepsilon}\}$ for any $\varepsilon \geq 0$. Let us consider the transport plan $\Pi = \boldsymbol{a}\boldsymbol{b}^\top = (a_i b_j)$ (with respect to the cost matrix C). Because all terms are positive, we have:

$$\begin{aligned}
|h| &\leqslant \langle \boldsymbol{a}\boldsymbol{b}^\top, C\rangle + \varepsilon \mathtt{KL}(\boldsymbol{a}\boldsymbol{b}^\top | \boldsymbol{a}\boldsymbol{b}^\top) + \tau D_\phi\left((\boldsymbol{a}\boldsymbol{b}^\top)\mathbf{1}_n | \boldsymbol{a}\right) + \tau D_\phi\left((\boldsymbol{b}\boldsymbol{a}^\top)\mathbf{1}_n | \boldsymbol{b}\right) \\
&\leqslant M \sum_{i,j} a_i b_j + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{b}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{a}}) \\
&\leqslant M m_{\boldsymbol{a}} m_{\boldsymbol{b}} + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{b}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{a}})
\end{aligned} \tag{14}$$

Defining $M^h_{\boldsymbol{a},\boldsymbol{b}}$ as the last upper bound finishes the proof. The case $h = S_\phi^{\tau,\varepsilon}$, is the sum of three terms of the form $\mathrm{OT}_\phi^{\tau,\varepsilon}$ Thus the sum $M^h_{\boldsymbol{a},\boldsymbol{b}} + \frac{1}{2}M^h_{\boldsymbol{a},\boldsymbol{a}} + \frac{1}{2}M^h_{\boldsymbol{b},\boldsymbol{b}}$ is an upper bound of S_ε. □

We now bound the UOT plan.

Lemma 2.2 (locally compact optimal transport plan)**.** *Assume that $\langle \boldsymbol{a}\boldsymbol{b}^\top, C\rangle < +\infty$. Consider regularized or unregularized UOT with entropy ϕ and penalty D_ϕ such that one has $\phi'_\infty > 0$. Then there exists an open neigbhourhood U around C, and a compact set K, such that the set of optimal transport plan for any $\tilde{C} \in U$ is in K,* i.e., $\mathrm{Opt}_h(\tilde{C}) \subset K$. *Furthermore, if all costs are uniformly bounded such that $0 \leq C \leq M < \infty$, then the compact K can be taken global, i.e. independent of C.*

Proof. We identify the mass of a positive measure with its L1 norm, i.e. $m_{\boldsymbol{a}} = \sum \boldsymbol{a}_i = ||\boldsymbol{a}||_1$. We first consider the case where $0 \leq C \leq M < \infty$. The OT cost is finite because the plan $\pi = \boldsymbol{a}\boldsymbol{b}^\top$ is suboptimal and yields $\mathrm{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{a}, \boldsymbol{b}, C) \leq M m_{\boldsymbol{a}} m_{\boldsymbol{b}} + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{b}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{a}}) < +\infty$.

Take a sequence Π_t approaching the infimum. Note that thanks to the Jensen inequality, one has $D_\phi(\boldsymbol{x}, \boldsymbol{y}) \geq m_{\boldsymbol{y}} \phi(m_{\boldsymbol{x}}/m_{\boldsymbol{y}})$ (see (Liero et al., 2017)). Write $m_\Pi = \sum \Pi_{t,ij}$. One has for any t

$$\begin{aligned}
&\langle \Pi_t, C \rangle + \tau D_\phi(\Pi_{t,1}\mathbf{1}_n | \boldsymbol{a}) + \tau D_\phi(\Pi_{t,2}^\top \mathbf{1}_n | \boldsymbol{b}) + \epsilon \mathrm{KL}(\Pi_t | \boldsymbol{a}\boldsymbol{b}^\top) \\
&\qquad \geq m_\Pi \left[\min C_{ij} + \tau \frac{m_{\boldsymbol{a}}}{m_\Pi} \phi\big(\frac{m_\Pi}{m_{\boldsymbol{a}}}\big) + \tau \frac{m_{\boldsymbol{b}}}{m_\Pi} \phi\big(\frac{m_\Pi}{m_{\boldsymbol{b}}}\big) + \varepsilon \frac{m_{\boldsymbol{a}} m_{\boldsymbol{b}}}{m_\Pi} \phi_{KL}\big(\frac{m_\Pi}{m_{\boldsymbol{a}} m_{\boldsymbol{b}}}\big) \right] \\
&\qquad \geq m_\Pi L(m_\Pi).
\end{aligned}$$

If $||\Pi_t||_1 = m_\Pi \to +\infty$, then $L(m_\Pi) \to +\infty$ if $\varepsilon > 0$ and $L(m_\Pi) \to \min C_{ij} + 2\phi'_\infty > 0$ otherwise. In both cases, as $t \to \infty$ and $||\Pi_t||_1 = m_\Pi \to +\infty$, we are supposed to approach the infimum but its lower bound goes to $+\infty$, which contradicts the fact that the optimal OT cost is finite.

More precisely, there exists a large enough value $\tilde{M}$ such that for $m_\Pi > \tilde{M}$, the lower bound is superior to the upper bound $M m_{\boldsymbol{a}} m_{\boldsymbol{b}} + \tau m_{\boldsymbol{a}} \phi(m_{\boldsymbol{b}}) + \tau m_{\boldsymbol{b}} \phi(m_{\boldsymbol{a}})$ and thus necessarily not optimal. Furthermore, $\tilde{M}$ depends on $(m_{\boldsymbol{a}}, m_{\boldsymbol{b}}, M)$ since $0 \leq C \leq M$. Thus, there exists $\tilde{M} > 0$ and some t_0 such that for $t \geq t_0$ any plan approaching the optimum statisfies $||\Pi_t||_1 \leq \tilde{M}$. The sequence $(\Pi_t)_t$ is in a finite dimensional, bounded, and closed set, i.e. a compact set. One can extract a converging subsequence whose limit is a plan attaining the minimum. Thus any optimal plan is necessarily in a compact set.

To generalize to local compactness, we consider $\delta > 0$ and a neighbourhood U of C such that for any $\tilde{C} \in U$ one has $0 \leq \tilde{C} \leq \max C + \delta$. Reusing the above proof yields the existence of $\tilde{M}$ such that for any $\tilde{C} \in U$, any plan approaching the optimum satisfies $||\Pi_t||_1 \leq \tilde{M}$, but this time $\tilde{M}$ depends on $(m_{\boldsymbol{a}}, m_{\boldsymbol{b}}, \max C + \delta)$, which is independent of C in its neighbourhood. □

We recall we denote the set of all optimal transport plan $\mathrm{Opt}_h(\boldsymbol{X}, \boldsymbol{Y}) \subset \mathcal{M}_+(\mathcal{X})$. While the UOT energy takes positive vectors and a ground cost as inputs, we make a slight abuse of notation with $\mathrm{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$. Indeed, the ground cost can be deduced from $\boldsymbol{X}, \boldsymbol{Y}$ and we associate uniform vectors as $\boldsymbol{a}$ and $\boldsymbol{b}$. As each element Π of $\mathrm{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$ is bounded by a constant M, $\mathrm{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$ is a compact space of $\mathcal{M}_+(\mathcal{X})$. We denote the maximal constant M which bounds all elements of $\mathrm{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$ as $\mathfrak{M}_\Pi$. We now prove that the set of optimal transport plan is convex , which will be useful for the optimization section.

Lemma 3 (optimal transport plan convexity)**.** *Consider regularized or unregularized UOT with entropy ϕ and penalty D_ϕ. The set of all optimal transport plan $\mathrm{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$ is a convex set.*

Proof. It is a general property of convex analysis. Take a convex function f and two points $(\boldsymbol{x}, \boldsymbol{y})$ that both attain the minimum over a convex set E. Write $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$ for $t \in [0,1]$. By convexity and suboptimality of $\boldsymbol{z}$ one has $\min_E f \leq f(\boldsymbol{z}) \leq t f(\boldsymbol{x}) + (1-t) f(\boldsymbol{y}) = \min_E f$. Thus z is also optimal, hence the set of minimizers is convex. The losses $\mathrm{OT}_\phi^{\tau,\varepsilon}$ fall under this setting. □

Finally, we provide a final result about UOT cost which is also useful for the optimization properties.

Lemma 4 (UOT is Lipschitz in the cost C)**.** *The map $C \mapsto h(\boldsymbol{u}, \boldsymbol{u}, C)$ is locally Lipshitz. Furthermore, if the costs are uniformly bounded ($0 \leq C \leq M$) then the loss is globally Lipschitz.*

Proof. We recall that $h(\boldsymbol{u}, \boldsymbol{u}, C) = \min_{\Pi \in \mathbb{R}_+^{n \times n}} \mathcal{F}(\Pi, C)$. Let C_1 and C_2 be two ground costs. Let Π_1 and Π_2 be the optimal solutions of $h(\boldsymbol{u}, \boldsymbol{u}, C_1)$ and $h(\boldsymbol{u}, \boldsymbol{u}, C_2)$, *i.e.,* $h(\boldsymbol{u}, \boldsymbol{u}, C_1) = \mathcal{F}(\Pi_1, C_1)$. Then we have:

$$\mathcal{F}(\Pi_1, C_1) - \mathcal{F}(\Pi_1, C_2) \leq h(\boldsymbol{u}, \boldsymbol{u}, C_1) - h(\boldsymbol{u}, \boldsymbol{u}, C_2) \leq \mathcal{F}(\Pi_2, C_1) - \mathcal{F}(\Pi_2, C_2) \tag{15}$$

Thus we have

$$\begin{aligned}
h(\boldsymbol{u}, \boldsymbol{u}, C_1) - h(\boldsymbol{u}, \boldsymbol{u}, C_2) &\leq \mathcal{F}(\Pi_2, C_1) - \mathcal{F}(\Pi_2, C_2) \\
&= \langle \Pi_2, C_1 - C_2 \rangle \qquad (16) \\
&\leq \|\Pi_2\| \|C_1 - C_2\| \qquad (17)
\end{aligned}$$

Where the last inequality uses the Cauchy-Schwarz inequality. Following the same logic we get a bound for minus the left hand term

$$h(\boldsymbol{u}, \boldsymbol{u}, C_2) - h(\boldsymbol{u}, \boldsymbol{u}, C_1) \leq \mathcal{F}(\Pi_1, C_2) - \mathcal{F}(\Pi_1, C_1) \leq \|\Pi_1\| \|C_1 - C_2\| \tag{18}$$

It remains to bound $\|\Pi_i\|$. When we study the local Lipschitz property, without loss of generality, we fix C_1 and take C_2 in a local neighbourhood of C_1. Thus Lemma 2.2, gives that $\|\Pi_i\| \leq \tilde{M}$, where $\tilde{M}$ only depends on $(\phi, \tau, \epsilon, \boldsymbol{a}, \boldsymbol{b}, \max C)$, with $\boldsymbol{a} = \boldsymbol{b} = \boldsymbol{u}$, i.e. it is locally independant of C in its neighbourhood, hence the local Lipschitz property. When $0 \leq C \leq M$, then $\tilde{M}$ is independent of the cost, hence the bound is global and the map is globally Lipschitz. □

B.2. Statistical and optimization proofs

We consider a positive, symmetric, definite and $\mathbf{C}^1$ ground cost and without loss of generality, we consider our ground cost to be squared euclidean. We recall our definitions and hypothesis. As the distributions α and β are compactly supported, there exists a constant $M > 0$ such that for any $1 \leqslant i, j \leqslant n$, $c(\boldsymbol{x}_i, \boldsymbol{y}_j) \leqslant M$ with $M := \text{diam}(\text{Supp}(\alpha) \cup \text{Supp}(\beta))^2$. We also furthermore suppose that the input masses $m_{\boldsymbol{a}}$ and $m_{\boldsymbol{b}}$ of positive vectors are strictly positive and finite, i.e., $0 < m_{\boldsymbol{a}} < \infty$. These hypothesis assures us that the UOT cost is finite and that the UOT plan is bounded.

B.2.1. Proof of Theorem 1

We now give the details of the proof of theorem 1. We separate theorem 1 in two sub theorem 1.1 and theorem 1.2. In the theorem 1.1, we show the deviation bound between $\widetilde{h}_k^m$ and E_h and in theorem 1.2, we show the deviation bound between $\widetilde{\Pi}_k^m$ and $\overline{\Pi}^m$. For theorem 1.1, we rely on two lemmas. The first lemma bounds the deviation between the complete estimator $\bar{h}^m$ and its expectation E_h. We denote the floor function as $\lfloor x \rfloor$ which returns the biggest integer smaller than x.

Lemma 5 (U-statistics concentration bound)**.** *Let $\delta \in (0, 1)$, three integers $k \geq 1$ and $m \leq n$ be fixed, and two compactly supported distributions α, β. Consider two n-tuples $\boldsymbol{X} \sim \alpha^{\otimes n}$ and $\boldsymbol{Y} \sim \beta^{\otimes n}$ and a kernel $h \in \{\text{OT}_\phi^{\tau,\varepsilon}, S_\phi^{\tau,\varepsilon}\}$. We have a concentration bound between $\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})$ and the expectation over minibatches E_h depending on the number of empirical data n*

$$|\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y}) - E_h| \leq M_{\boldsymbol{u},\boldsymbol{u}}^h \sqrt{\frac{\log(2/\delta)}{2\lfloor n/m \rfloor}} \tag{19}$$

with probability at least $1 - \delta$ and where $M_{\boldsymbol{u},\boldsymbol{u}}^h$ is an upper bound defined in lemma 2.1.

Proof. $\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})$ is a two-sample U-statistic of order $2m$ and E_h is its expectation as $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *iid* random variables. $\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})$ is a sum of dependant variables and it is possible to rewrite $\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})$ as a sum of independent random variables. As α, β are compactly supported by hypothesis, the UOT loss is bounded thanks to lemma 2.1. Thus, we can apply the famous Hoeffding lemma to our U-statistic and get the desired bound. The proof can be found in (Hoeffding, 1963) (the two sample U-statistic case is discussed in section 5.b) . □

The second lemma bounds the deviation between the incomplete estimator $\widetilde{h}_k^m$ and the complete estimator $\bar{h}^m$.

Lemma 6 (Deviation bound)**.** *Let $\delta \in (0, 1)$, three integers $k \geq 1$ and $m \leq n$ be fixed, and two compactly supported distributions α, β. Consider two n-tuples $\boldsymbol{X} \sim \alpha^{\otimes n}$ and $\boldsymbol{Y} \sim \beta^{\otimes n}$ and a kernel $h \in \{\text{OT}_\phi^{\tau,\varepsilon}, S_\phi^{\tau,\varepsilon}\}$. We have a deviation bound between $\widetilde{h}_k^m(\boldsymbol{X}, \boldsymbol{Y})$ and $\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})$ depending on the number of batches k*

$$|\widetilde{h}_k^m(\boldsymbol{X}, \boldsymbol{Y}) - \bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})| \leqslant M_{\boldsymbol{u},\boldsymbol{u}}^h \sqrt{\frac{2\log(2/\delta)}{k}} \tag{20}$$

with probability at least $1 - \delta$ and where $M_{\boldsymbol{u},\boldsymbol{u}}^h$ is an upper bound defined in lemma 2.1.

Proof. First note that $\widetilde{h}_k^m(\boldsymbol{X}, \boldsymbol{Y})$ is an subsample quantity of $\bar{h}^m(\boldsymbol{X}, \boldsymbol{Y})$. Let us consider the sequence of random variables $((\mathfrak{b}_l(I, J)_{(I,J)\in\mathcal{P}^m})_{1\leqslant l\leqslant k}$ such that $\mathfrak{b}_l(I, J)$ is equal to 1 if (I, J) has been selected at the $l-$th draw and 0 otherwise. By construction of $\widetilde{h}_k^m$, the aforementioned sequence is an i.i.d sequence of random vectors and the $\mathfrak{b}_l(I, J)$ are Bernoulli

random variables of parameter $1/|\Gamma|$. We then have

$$\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y}) - \bar{h}^m(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{k}\sum_{l=1}^{k}\omega_l \tag{21}$$

where $\omega_l = \sum_{(I,J)\in\mathcal{P}^m}(\mathfrak{b}_l(I,J) - \frac{1}{|\Gamma|})h(I,J)$. Conditioned upon $\boldsymbol{X} = (\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n)$ and $\boldsymbol{Y} = (\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)$, the variables ω_l are independent, centered and bounded by $2M^h_{\boldsymbol{u},\boldsymbol{u}}$ thanks to lemma 2.1. Using Hoeffding's inequality yields

$$\mathbb{P}(|\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y}) - \bar{h}^m(\boldsymbol{X},\boldsymbol{Y})| > \varepsilon) = \mathbb{E}[\mathbb{P}(|\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y}) - \bar{h}^m(\boldsymbol{X},\boldsymbol{Y})| > \varepsilon|X,Y)] \tag{22}$$

$$= \mathbb{E}[\mathbb{P}(|\frac{1}{k}\sum_{l=1}^{k}\omega_l)| > \varepsilon|X,Y)] \tag{23}$$

$$\leqslant \mathbb{E}[2e^{\frac{-k\varepsilon^2}{2(M^h_{\boldsymbol{u},\boldsymbol{u}})^2}}] = 2e^{\frac{-k\varepsilon^2}{2(M^h_{\boldsymbol{u},\boldsymbol{u}})^2}} \tag{24}$$

which concludes the proof. □

We are now ready to prove Theorem 1.1.

Theorem 1.1 (Maximal deviation bound)**.** *Let $\delta \in (0,1)$, three integers $k \geq 1$ and $m \leq n$ be fixed and two compactly supported distributions α, β. Consider two n-tuples $\boldsymbol{X} \sim \alpha^{\otimes n}$ and $\boldsymbol{Y} \sim \beta^{\otimes n}$ and a kernel $h \in \{\mathrm{OT}^{\tau,\varepsilon}_\phi, S^{\tau,\varepsilon}_\phi\}$. We have a maximal deviation bound between $\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y})$ and the expectation over minibatches E_h depending on the number of empirical data n and the number of batches k*

$$|\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y}) - E_h| \leq M^h_{\boldsymbol{u},\boldsymbol{u}}\sqrt{\frac{\log(2/\delta)}{2\lfloor n/m\rfloor}} + M^h_{\boldsymbol{u},\boldsymbol{u}}\sqrt{\frac{2\log(2/\delta)}{k}} \tag{25}$$

with probability at least 1 - δ and where $M^h_{\boldsymbol{u},\boldsymbol{u}}$ is an upper bound defined in lemma 2.1.

Proof. Thanks to lemma 6 and 5 we get

$$|\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y}) - E_h| \leq |\widetilde{h}_k^m(\boldsymbol{X},\boldsymbol{Y}) - \bar{h}^m(\boldsymbol{X},\boldsymbol{Y})| + |\bar{h}^m(\boldsymbol{X},\boldsymbol{Y}) - E_h| \tag{26}$$

$$\leq M^h_{\boldsymbol{u},\boldsymbol{u}}\sqrt{\frac{\log(2/\delta)}{2\lfloor n/m\rfloor}} + M^h_{\boldsymbol{u},\boldsymbol{u}}\sqrt{\frac{2\log(2/\delta)}{k}} \tag{27}$$

with probability at least $1 - (\frac{\delta}{2} + \frac{\delta}{2}) = 1 - \delta$. □

We now give the details of the proof of theorem 1.2. In what follows, we denote by $\Pi_{(i)}$ the i-th row of matrix Π. Let us denote by $\mathbf{1} \in \mathbb{R}^n$ the vector whose entries are all equal to 1.

Theorem 1.2 (Distance to marginals)**.** *Let $\delta \in (0,1)$, two integers $m \leq n$ be fixed. Consider two n-tuples $\boldsymbol{X} \sim \alpha^{\otimes n}$ and $\boldsymbol{Y} \sim \beta^{\otimes n}$ and the kernel $h = \mathrm{OT}^{\tau,\varepsilon}_\phi$. For all integer $k \geq 1$, all $1 \leqslant i \leqslant n$, with probability at least $1-\delta$ on the draw of $\boldsymbol{X}, \boldsymbol{Y}$ and D_k we have*

$$|\widetilde{\Pi}_k^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1} - \overline{\Pi}^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1}| \leqslant \mathfrak{M}^\infty_\Pi\sqrt{\frac{2\log(2/\delta)}{k}}, \tag{28}$$

where $\mathfrak{M}^\infty_\Pi$ denotes an upper bound of all minibatch UOT plan.

Proof. Let us consider the sequence of random variables $((\mathfrak{b}_p(I,J)_{(I,J)\in\Gamma})_{1\leqslant p\leqslant k}$ such that $\mathfrak{b}_p(I,J)$ is equal to 1 if (I,J) has been selected at the $p-$th draw and 0 otherwise. By construction of $\widetilde{\Pi}_k^m(\boldsymbol{X},\boldsymbol{Y})$, the aforementioned sequence is an i.i.d sequence of random vectors and the $\mathfrak{b}_p(I,J)$ are bernoulli random variables of parameter $1/|\Gamma|$. We then have

$$\widetilde{\Pi}_k^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1} = \frac{1}{k}\sum_{p=1}^{k}\omega_p \tag{29}$$

where $\omega_p = \sum_{(I,J)\in\Gamma}\sum_{j=1}^{n}(\Pi_{I,J})_{i,j}\mathfrak{b}_p(I,J)$. Conditioned upon $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$, the random vectors ω_p are independent, and thanks to lemma 2.2, they are bounded by a constant $\mathfrak{M}_\Pi$ which is the maximum mass of all optimal minibatch unbalanced plan in $\mathrm{Opt}_h(\boldsymbol{X}(I), \boldsymbol{Y}(J))$. We denote the maximum upper bound $\mathfrak{M}_\Pi$ of all minibatch UOT plan as $\mathfrak{M}_\Pi^\infty$. Moreover, one can observe that $\mathbb{E}[\widetilde{\Pi}_k^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1}] = \overline{\Pi}^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1}$. Using Hoeffding's inequality yields

$$\mathbb{P}(|\widetilde{\Pi}_k^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1} - \overline{\Pi}^m(\boldsymbol{X},\boldsymbol{Y})_{(i)}\mathbf{1})| > \varepsilon) = \mathbb{E}[\mathbb{P}(|\frac{1}{k}\sum_{p=1}^{k}\omega_p - \mathbb{E}[\frac{1}{k}\sum_{p=1}^{k}\omega_p])| > \varepsilon|X,Y)] \tag{30}$$

$$\leqslant 2e^{-2\frac{k\varepsilon^2}{(\mathfrak{M}_\Pi^\infty)^2}} \tag{31}$$

which concludes the proof. □

Note that the unbalanced Sinkhorn divergence $S_\phi^{\tau,\varepsilon}$ involves three terms of the form $\mathrm{OT}_\phi^{\tau,\varepsilon}$, hence three transport plans, which explains why we do not attempt to define an associated averaged minibatch transport matrix.

B.2.2. Proof of Theorem 2

To prove the exchange of gradients and expectations over minibatches we rely on Clarke differential. We need to use this non smooth analysis tool as unregularized UOT is not differentiable. It is not differentiable because the set of optimal solutions might not be a singleton. Clarke differential are generalized gradients for locally Lipschitz function and non necessarily convex. A similar strategy was developped in (Fatras et al., 2021). The key element of this section is to rewrite the original UOT problem $\mathrm{OT}_\phi^{\tau,\varepsilon}$ as:

$$\mathrm{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{a},\boldsymbol{b},C) = \min_{\Pi\in\mathbb{R}_+^{n\times n}} \langle C,\Pi\rangle + \varepsilon\mathtt{KL}(\Pi|\boldsymbol{a}\otimes\boldsymbol{b}) + \tau D_\phi(\Pi\mathbf{1}_n|\boldsymbol{a}) + \tau D_\phi(\Pi^\top\mathbf{1}_n|\boldsymbol{b}) \tag{32}$$

$$= \min_{\Pi\in\mathrm{Opt}_{\mathrm{OT}_\phi^{\tau,\varepsilon}}} \langle C,\Pi\rangle + \varepsilon\mathtt{KL}(\Pi|\boldsymbol{a}\otimes\boldsymbol{b}) + \tau D_\phi(\Pi\mathbf{1}_n|\boldsymbol{a}) + \tau D_\phi(\Pi^\top\mathbf{1}_n|\boldsymbol{b}), \tag{33}$$

Where $\mathrm{Opt}_{\mathrm{OT}_\phi^{\tau,\varepsilon}}(\boldsymbol{X},\boldsymbol{Y})$ is a compact set of the set of measures $\mathcal{M}_+(\mathcal{X})$. The compact set is a key element for using Danskin like theorem (Proposition B.25 (Bertsekas, 1997)).

We start by recalling a basic proposition for Clarke regular function:

Proposition 3. *A $\mathbf{C}^1$ or convex map is Clarke regular.*

Proof. see Proposition 2.3.6 (Clarke, 1990) □

We first give a lemma which gives the Clarke regularity of the UOT cost with respect to a parametrized random vector.

Lemma 7. *Let $\boldsymbol{u}$ be a uniform probability vector. Let $\boldsymbol{X}$ be a $\mathbb{R}^{dm}$-valued random variable, and $\{\boldsymbol{Y}_\theta\}$ a family of $\mathbb{R}^{dm}$-valued random variables defined on the same probability space, indexed by $\theta\in\Theta$, where $\Theta\subset\mathbb{R}^q$ is open. Assume that $\theta\mapsto\boldsymbol{Y}_\theta$ is $\mathbf{C}^1$. Consider a $\mathbf{C}^1$ cost C and let $h\in\{\mathrm{OT}_\phi^{\tau,\varepsilon}, S_\phi^{\tau,\varepsilon}\}$. Then the function $\theta\mapsto -h(\boldsymbol{u},\boldsymbol{u},C(\boldsymbol{X},\boldsymbol{Y}_\theta))$ is Clarke regular. Furthermore, for $h=\mathrm{OT}_\phi^{\tau,\varepsilon}$ and for all $1\leq i\leq q$ we have:*

$$\partial_{\theta_i} h(\boldsymbol{u},\boldsymbol{u},C(\boldsymbol{X},\boldsymbol{Y}_\theta)) = \overline{co}\{-\langle\Pi\cdot D\rangle\cdot(\nabla_{\theta_i}Y) : \Pi\in\mathrm{Opt}_h(\boldsymbol{X},\boldsymbol{Y}), \\ D\in\mathbb{R}^{m,m}, D_{j,k} = \nabla_Y C_{j,k}(\boldsymbol{X},\boldsymbol{Y}_\theta)\} \tag{34}$$

where ∂_{θ_i} is the Clarke subdifferential with respect to θ_i, $\nabla_Y C_{j,k}$ is the differential of the cell $C_{j,k}$ of the cost matrix with respect to Y, $\mathrm{Opt}_h(\boldsymbol{X},\boldsymbol{Y}_\theta)$ is the set of optimal transport plan and $\overline{co}$ denotes the closed convex hull. Note that when $\varepsilon>0$ the set $\mathrm{Opt}_h(\boldsymbol{X},\boldsymbol{Y}_\theta)$ is reduced to a singleton, and the notation $\overline{co}$ is superfluous.

Proof. We start with the regularity of $\theta\mapsto -\mathrm{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{u},\boldsymbol{u},C(\boldsymbol{X},\boldsymbol{Y}_\theta))$. To prove the Clarke regularity of this map, we rely on a chain rule argument. Consider the function $Y\mapsto C_{j,k}(\boldsymbol{X},\boldsymbol{Y})$, it is Clarke regular because it is $\mathbf{C}^1$. Since $\theta\mapsto\boldsymbol{Y}_\theta$ is $\mathbf{C}^1$, it follows by the chain rule that $\theta\mapsto C_{j,k}(\boldsymbol{X},\boldsymbol{Y}_\theta)$ is $\mathbf{C}^1$ and thus Clarke regular. The Unbalanced OT cost $\mathrm{OT}_\phi^{\tau,\varepsilon}$ is a minimization of an energy which is linear in C, and it is thus concave in C, hence $-\mathrm{OT}_\phi^{\tau,\varepsilon}$ is Clarke

regular by convexity. Therefore from Theorem 2.3.9(i) and Proposition 2.3.1 (for $s = 1$) in (Clarke, 1990) it follows that $\theta \mapsto -\operatorname{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))$ is Clarke regular.

We now furnish the gradients associated to $\theta \mapsto -\operatorname{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))$. By chain rule, the gradient of $\theta \mapsto C_{j,k}(\boldsymbol{X}, \boldsymbol{Y}_\theta)$ reads

$$\nabla_{\theta_i} C_{j,k}(\boldsymbol{X}, \boldsymbol{Y}_\theta) = \nabla_Y C_{j,k}(\boldsymbol{X}, \boldsymbol{Y}_\theta) \cdot \nabla_{\theta_i} Y_\theta.$$

We now deal with the gradient of the map $C \mapsto \operatorname{OT}_\phi^{\tau,\varepsilon}(\boldsymbol{u}, \boldsymbol{u}, C)$ by verifying the assumptions of Danskin's theorem (Clarke, 1975, Theorem 2.1). We use in particular the remark below (Clarke, 1975, Theorem 2.1) which states that the hypothesis on the map are verified if the map is *u.s.c* in both variables (Π, C) and convex in C. We recall that $\operatorname{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$ is a compact and a convex set, thanks to lemma 3 and lemma 4. Furthermore, the energy associated to $h = \operatorname{OT}_\phi^{\tau,\varepsilon}$ is concave in the cost C and *l.s.c* in (Π, C) (Liero et al., 2017, Lemma 3.9). From (Clarke, 1975, Theorem 2.1) it follows that the subderivatives of the convex function $C \mapsto -h(\boldsymbol{u}, \boldsymbol{u}, C)$ are equal to $\operatorname{Opt}_h(\boldsymbol{X}, \boldsymbol{Y})$, due to the energy's linearity in C. Thus combining the formulas of the Danskin theorem with the Chain rule yields Equation (34). When $\varepsilon > 0$ the set $\operatorname{Opt}_h(\boldsymbol{X}, \boldsymbol{Y}_\theta)$ is reduced to a singleton, and the notation $\overline{\text{co}}$ is superfluous.

We now give the proof for the regularity of the map $\theta \mapsto S_\phi^{\tau,\varepsilon}(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))$ with $\varepsilon > 0$ as when $\varepsilon = 0$, we get the unregularized UOT treated in the above paragraph. We recall that $S_\phi^{\tau,\varepsilon}$ is the summation of three terms of the form $\operatorname{OT}_\phi^{\tau,\varepsilon}$. For $\varepsilon > 0$ and each term of the sum, the set of optimal plans $\operatorname{Opt}_{\operatorname{OT}_\phi^{\tau,\varepsilon}}(\boldsymbol{X}, \boldsymbol{Y})$ is reduced to a unique element and the differential (34) is also a singleton, thus $\operatorname{OT}_\phi^{\tau,\varepsilon}$ is differentiable. Then $S_\phi^{\tau,\varepsilon}$ is differentiable as a difference of differentiable functions. Furthermore $S_\phi^{\tau,\varepsilon}$ is also Clarke regular as a difference of differentiable functions. □

We finally prove theorem 2.

Theorem 2. *Let $\boldsymbol{u}$ be uniform probability vectors and let $\boldsymbol{X}, \boldsymbol{Y}, C$ be as in lemma 7, $h \in \{\operatorname{OT}_\phi^{\tau,\varepsilon}, S_\phi^{\tau,\varepsilon}\}$, and assume in addition that the random variables $\boldsymbol{X}, \{Y_\theta\}_{\theta \in \Theta}$ are compactly supported. If for all $\theta \in \Theta$ there exists an open neighbourhood U, $\theta \in U \subset \Theta$, and a random variable $K_U : \Omega \to \mathbb{R}$ with finite expected value, such that*

$$\|C(\boldsymbol{X}(\omega), \boldsymbol{Y}_{\theta_1}(\omega)) - C(\boldsymbol{X}(\omega), \boldsymbol{Y}_{\theta_2}(\omega))\| \leq K_U(\omega)\|\theta_1 - \theta_2\| \tag{35}$$

then we have

$$\partial_\theta \, \mathbb{E}\left[h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))\right] = \mathbb{E}\left[\partial_\theta h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))\right]. \tag{36}$$

with both expectation being finite. Furthermore the function $\theta \mapsto -\mathbb{E}\left[h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))\right]$ is also Clarke regular.

Proof. Suppose that $U \subset \Theta$ is open and K_U is a function for which (35) is satisfied. As data lie in compacts the ground cost C, which is $\mathbf{C}^1$, is in a compact K_C and as the map $C \mapsto h(\boldsymbol{u}, \boldsymbol{u}, C)$ is locally Lipshitz by lemma 4, there exists a uniform constant which makes the map $C \mapsto h(\boldsymbol{u}, \boldsymbol{u}, C)$ globally Lipshitz on the compact K_C. Thus, a similar bound to (35) is also satisfied for the function $h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}(\omega), \boldsymbol{Y}_\theta(\omega)))$. Thanks to lemma 7, $-h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))$ is Clarke regular, the interchange (36) and regularity of $\theta \mapsto -\mathbb{E}[h(\boldsymbol{u}, \boldsymbol{u}, C(\boldsymbol{X}, \boldsymbol{Y}_\theta))]$ will follow from Theorem 2.7.2 and Remark 2.3.5 (Clarke, 1990), once we establish that the expectation on the left hand side is finite. This is direct as we suppose we have compactly supported distributions and C is a $\mathbf{C}^1$ cost. Indeed consider the function which is equal to $M^h_{\boldsymbol{u},\boldsymbol{u}}$ on the distributions's support and which is set to 0 everywhere else. Taking the expectation on this function is finite as $M^h_{\boldsymbol{u},\boldsymbol{u}}$ is finite. □

C. Domain adaptation and partial domain adaptation experiments

In this section we provide architecture and training procedure details for the domain adaptation experiments. We also discuss the reported scores procedure. Finally we discuss the training behaviour for both JUMBOT and DEEPJDOT.

C.1. Domain adaptation

In this subsection, we detail the setup of our domain adaptation experiments.

Setup. First note that for all datasets, JUMBOT uses a stratified sampling on source minibatches as done in DEEPJDOT (Damodaran et al., 2018). Stratified sampling means that each class has the same number of samples in the minibatches. This is a realistic setting as labels are available in the source dataset.

Method	A-C	A-P	A-R	C-A	C-P	C-R	P-A	P-C	P-R	R-A	R-C	R-P	avg
RESNET-50	34.9	50.0	58.0	37.4	41.9	46.2	38.5	31.2	60.4	53.9	41.2	59.9	46.1
DANN	46.2	65.2	73.0	54.0	61.0	65.2	52.0	43.6	72.0	64.7	52.3	79.2	60.7
CDAN-E	52.8	71.4	76.1	59.7	70.6	71.5	59.8	50.8	77.7	71.4	58.1	**83.5**	67.0
ALDA	53.7	70.1	76.4	60.2	72.6	71.5	56.8	51.9	77.1	70.2	56.3	82.1	66.6
ROT	47.2	70.8	77.6	61.3	69.9	72.0	55.4	41.4	77.6	69.9	50.4	81.5	64.6
DEEPJDOT	53.4	71.7	77.2	62.8	70.2	71.4	60.2	50.2	77.1	67.7	56.5	80.7	66.6
JUMBOT	**55.3**	**75.5**	**80.8**	**65.5**	**74.4**	**74.9**	**65.4**	**52.7**	**79.3**	**74.2**	**59.9**	83.4	**70.1**

Table 1. Office-Home experiments with maximum classification (ResNet50)

For Digits datasets, we used the 9 CNN layers architecture and the 1 dense layer classification proposed in (Damodaran et al., 2018). We trained our neural network on the source domain during 10 epochs before applying JUMBOT. We used Adam optimier with a learning rate of $2e^{-4}$ with a minibatch size of 500. Regarding competitors, we use the official implementations with the considered architecture and training procedure.

For office-home and VisDA, we employed ResNet-50 as generator. ResNet-50 is pretrained on ImageNet and our discriminator consists of two fully connected layers with dropout, which is the same as previous works (Ganin et al., 2016; Long et al., 2018; Chen et al., 2020). As we train the classifier and discriminator from scratch, we set their learning rates to be 10 times that of the generator. We train the model with Stochastic Gradient Descent optimizer with a momentum of 0.9. We schedule the learning rate with the strategy in (Ganin et al., 2016), it is adjusted by $\chi_p = \frac{\chi_0}{(1+\mu q)^\nu}$, where q is the training progress linearly changing from 0 to 1, $\chi_0 = 0.01$, $\mu = 10$, $\nu = 0.75$.

We compare JUMBOT against recent domain adaptation papers, namely DANN(Ganin et al., 2016), CDAN-E (Long et al., 2018), ALDA (Chen et al., 2020), DEEPJDOT (Damodaran et al., 2018) and ROT (Balaji et al., 2020) on all considered datasets. We reproduced their scores and on contrary of these papers we do not report the best classification on the test along the iterations but at the end of training, which explains why there might be a difference between reported results and reproduced results. We sincerely believe that the evaluation shall only be done at the end of training as labels are not available in the target domain. But we also report the maximum accuracy along epochs for the Office-Home DA task in table 1 and it shows that our method is above all of the competitors by a safe margin of 3%.

For Office-Home, we made 10000 iterations with a batch size of 65 and for VisDA, we made 10000 iterations with a batch size of 72. For fair comparison we used our minibatch size and number of iterations to evaluate competitors. The hyperparameters used in our experiments are as follows $\eta_1 = 0.1, \eta_2 = 0.1,\ \eta_3 = 1, \tau = 1, \varepsilon = 0.1$ for the digits and for office-home datasets $\eta_1 = 0.01, \eta_2 = 0.5,\ \eta_3 = 1, \tau = 0.5, \varepsilon = 0.01$. For VisDA, $\eta_1 = 0.005, \eta_2 = 1,\ \eta_3 = 1, \varepsilon = 0.01$ and τ was set to 0.3.

C.2. Partial DA

For Partial Domain Adaptation, we considered a neural network architecture and a training procedure similar as in the domain adapation experiments which also corresponds to the setting in (Jian et al., 2020). Our hyperparameters are set as follows : $\tau = 0.06, \eta_1 = 0.003, \eta_2 = 0.75$ and finally η_3 was set to 10. Regarding training procedure, we made 5000 iterations with a batch size of 65 and for optimization procedure, we used the same as in (Jian et al., 2020). We do not use the ten crop technic to evaluate our model on the test set as we were not able to reproduce the results from ENT and PADA. Furthermore, we do not know if the reported results ENT and PADA were evulated at the end of optimization or during training, but our reported scores are above their scores by at least 5% on average.

C.3. Overfitting

In this subsection, we discuss the training behaviour of DEEPJDOT and our method JUMBOT on the DA task MNIST $\mapsto$ M-MNIST. In figure 1, one can see that DEEPJDOT starts overfitting from epoch 30 on each class. There are some classes which are more affected by overfitting than others. The accuracy on each class is reduced of several points. This behaviour is not shared with our method JUMBOT. Indeed it is more stable, it does not show any sign of overfitting and it has a higher accuracy. This shows the relevance of using our method JUMBOT.

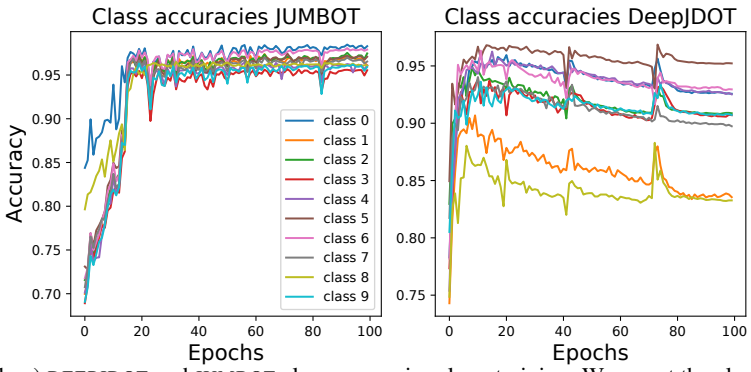

Figure 1. (Best viewed in colors) DEEPJDOT and JUMBOT class accuracies along training. We report the class accuracies along training of DEEPJDOT and JUMBOT on the DA task MNIST $\mapsto$ M-MNIST for optimal hyper-parameters. Each color represents a different class.

References

Balaji, Y., Chellappa, R., and Feizi, S. Robust optimal transport with applications in generative modeling and domain adaptation. In *Advances in Neural Information Processing Systems*, 2020. 10

Bertsekas, D. P. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997. 8

Chen, M., Zhao, S., Liu, H., and Cai, D. Adversarial-learned loss for domain adaptation. *arXiv*, abs/2001.01046, 2020. 10

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990. 8, 9

Clarke, H. F. Generalized gradients and applications. *Transactions of The American Mathematical Society*, pp. 247–247, 1975. 9

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation. In *ECCV 2018 - 15th European Conference on Computer Vision*. Springer, 2018. 9, 10

Damodaran, B. B., Flamary, R., Seguy, V., and Courty, N. An Entropic Optimal Transport Loss for Learning Deep Neural Networks under Label Noise in Remote Sensing Images. In *Computer Vision and Image Understanding*, 2019.

Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. Minibatch optimal transport distances; analysis and applications. *CoRR*, 2021. 8

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 10

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 1963. 6

Jian, L., Yunbo, W., Dapeng, H., Ran, H., and Jiashi, F. A balanced and uncertainty-aware approach for partial domain adaptation. In *European Conference on Computer Vision (ECCV)*, August 2020. 10

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, Dec 2017. ISSN 1432-1297. doi: 10.1007/s00222-017-0759-8. 3, 5, 9

Long, M., Cao, Z., Wang, J., and Jordan, M. I. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1645–1655, 2018. 10

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.