On the potential benefits of entropic regularization for smoothing Wasserstein estimators

Jérémie Bigot[1,2], Paul Freulon *[1,2], Boris P. Hejblum[1,3,4], and Arthur Leclaire[1,2]

[1]Université de Bordeaux, Bordeaux, 33000, France.
[2]Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.
[3]Bordeaux Population Health Research Center Inserm U1219, Inria SISTM, 33000 Bordeaux, France.
[4]Vaccine Research Institute (VRI), 94010 Créteil, France.

Abstract

This paper is focused on the study of entropic regularization in optimal transport as a smoothing method for Wasserstein estimators, through the prism of the classical tradeoff between approximation and estimation errors in statistics. Wasserstein estimators are defined as solutions of variational problems whose objective function involves the use of an optimal transport cost between probability measures. Such estimators can be regularized by replacing the optimal transport cost by its regularized version using an entropy penalty on the transport plan. The use of such a regularization has a potentially significant smoothing effect on the resulting estimators. In this work, we investigate its potential benefits on the approximation and estimation properties of regularized Wasserstein estimators. Our main contribution is to discuss how entropic regularization may reach, at a lowest computational cost, statistical performances that are comparable to those of un-regularized Wasserstein estimators in statistical learning problems involving distributional data analysis. To this end, we present new theoretical results on the convergence of regularized Wasserstein estimators. We also study their numerical performances using simulated and real data in the supervised learning problem of proportions estimation in mixture models using optimal transport.

1 Introduction

Wasserstein estimators are defined as solutions of variational problems whose objective function involves the use of an optimal transport (OT) cost between probability measures. Such estimators typically arise in statistical problems involving the minimization of a Wasserstein distance (or more generally an OT cost) between the empirical measure of the data and a distribution belonging to a parametric model (see Bernton et al. [2019]), and this class of estimators has found important applications in generative adversarial models for image processing (see e.g. Arjovsky et al. [2017]). Wasserstein estimators also represent an important class of inference methods in the field of statistical optimal transport for distributional data analysis where the observations at hand can be modeled as a set of histograms (see e.g. Bigot [2020], Panaretos and Zemel [2018], Petersen et al. [2022] for recent reviews).

Despite the appealing geometric properties of Wasserstein distances for comparing probability distributions, the computational burden required to evaluate an optimal transport cost is an important limitation for its application in data analysis. The seminal paper Cuturi [2013] has opened a breach in the computational complexity of optimal transport by the addition of an entropic regularizing term in the OT Kantorovich's formulation. In the last years, the benefit of this regularization has been to allow the use of OT based methods in statistics and machine learning with

*correspondence: paul.freulon@math.u-bordeaux.fr a time complexity that scales quadratically in the number of data using the Sinkhorn algorithm. This represents a significant improvement over the computational cost of un-regularized OT that scales cubically in the number of observations using linear programming. However, regularized OT has been mainly used so far as a fast numerical method to approximate un-regularized OT.

In this paper, we advocate the use of entropic regularization in computational OT as a smoothing method for un-regularized Wasserstein estimators. These estimators are obtained by replacing the standard OT cost in a variational problem by its entropy regularized version. The use of such a regularization has a beneficial smoothing effect on the resulting estimators as shown in Bigot et al. [2018] for the specific problem of computing a smooth Wasserstein barycenter from a set of discrete probability measures. In this paper, we investigate the impact of this smoothing effect of regularized Wasserstein estimators through the prism of the tradeoff between approximation and estimation errors in statistics which is reminiscent of the classical bias versus variance tradeoff). Our main contribution is to discuss how entropic regularization yields estimators that may reach, at a lowest computational cost, statistical performances that are comparable to those of un-regularized Wasserstein estimators in statistical learning problems involving distributional data analysis. To this end, we present new theoretical results on the convergence of regularized Wasserstein estimators. We also study their numerical performances using simulated and real data in the supervised learning problem of proportions estimation in mixture models using optimal transport.

1.1 Proportions estimation in mixture models using optimal transport

The motivation of this work comes from the active research field of automated analysis of flow cytometry measurements, see Aghaeepour et al. [2013]. Flow cytometry is a high-throughput biotechnology used to characterize a large amount of m cells from a biological sample (with $m \geq 10^5$) that produces a data set $X_1, \ldots, X_m$ where each observation $X_i \in \mathbb{R}^d$ corresponds to a vector of d biomarkers of each single cell. Automated approaches in flow cytometry aim at clustering the data in order to estimate cellular population proportions in the biological sample. In Freulon et al. [2020], we have considered that such a data set can be represented with a discrete probability distribution $\frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}$ with support in $\mathbb{R}^d$, and we have introduced a new supervised algorithm based on regularized OT to estimate the different cell population proportions from a biological sample characterized with flow cytometry measurements. This approach optimally re-weights class proportions in a mixture model between a source data set (with known segmentation into cell sub-populations) to fit a target data set with unknown segmentation.

To be more precise, let us denote by $Y_1, ..., Y_n$, the dataset from the target sample, and by $X_1, ..., X_m$ the observations from the source biological sample. Thanks to the knowledge of a clustering of the source dataset into K classes $C_1, \ldots, C_K$, the empirical measure $\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}$ can be decomposed as the following mixture of probability measures,

$$\hat{\mu} = \sum_{k=1}^{K} \frac{m_k}{m} \left(\sum_{i:X_i \in C_k} \frac{1}{m_k} \delta_{X_i} \right) = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_k, \text{ where } \hat{\pi}_k = \frac{m_k}{m}, \tag{1.1}$$

and each component $\hat{\mu}_k = \sum_{i:X_i \in C_k} \frac{1}{m_k} \delta_{X_i}$ corresponds to a known sub-population of cells with $m_k = \#C_k$. Then, the method proposed in Freulon et al. [2020] aims at modifying the weights $(\hat{\pi}_k)_{1 \leq k \leq K}$ in such a way that the re-weighted source measure minimizes a regularized OT cost with respect to the target measure $\frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}$. Then, the resulting weights yield an estimation of the proportions of sub-population of cells in the target sample. However, despite the efficiency of the method for the analysis of flow cytometry data, the work in Freulon et al. [2020] opens questions on the influence of the regularization, and we set to answer some of them in this work.

Let us now formalize the problem of proportions estimation in mixture models using regularized OT. We denote by $\mu = \sum_{k=1}^{K} \pi_k \mu_k$ a probability measure that can be decomposed as a mixture of K probability measures $\mu_1, \ldots, \mu_K$. For $\theta \in \Sigma_K$, where $\Sigma_K = \{(\theta_1, \ldots, \theta_K) \in \mathbb{R}_+^K \; : \; \sum_{k=1}^{K} \theta_k = 1\}$

is the K-dimensional simplex, we define μ_θ as the re-weighted version of μ that is defined as

$$\mu_\theta = \sum_{k=1}^{K} \theta_k \mu_k. \tag{1.2}$$

Let ν be another probability measure. Proportions estimation in mixture models using OT is defined as the problem of finding $\theta^* \in \Sigma_K$ that minimizes an OT cost between μ_θ and ν. Denoting $W_0(\mu, \nu)$ the un-regularized OT cost between μ and ν (we shall focus on the squared Wasserstein metric associated to the quadratic cost), the optimal vector of class proportions that we are targeting is:

$$\theta^* \in \arg\min_{\theta \in \Sigma_K} W_0(\mu_\theta, \nu).$$

In practice, one only has access to independent samples from μ and ν denoted by $X_1, ..., X_m$ (with a know clustering) and $Y_1, ..., Y_n$ respectively. Therefore, estimators of θ^* will be obtained from the empirical versions of μ_θ and ν denoted by

$$\hat{\mu}_\theta = \sum_{k=1}^{K} \theta_k \hat{\mu}_k \quad \text{and} \quad \hat{\nu} = \frac{1}{n} \sum_{j=1}^{n} \delta_{Y_j}.$$

The computational cost to numerically evaluate $W_0(\mu, \nu)$ can be prohibitive, which led Freulon et al. [2020] to consider its regularized version denoted by $W_\lambda(\mu, \nu)$ where $\lambda > 0$ represents the amount of entropic penalty that is put on the transport plan in the primal formulation of OT. Here, this regularized version of the OT cost is computed using the Sinkhorn algorithm, an iterative procedure whose convergence properties are now well understood (see Chizat et al. [2020] for a recent overview). However, after ℓ iterations of the Sinkhorn algorithm, it should be noted that one only has an approximation of the regularized OT cost that we will denote by $W_\lambda^{(\ell)}(\mu, \nu)$. In this work, we focuses on the study of the convergence rate of the following estimator towards the optimal vector of class proportions θ^*:

$$\hat{\theta}_\lambda^{(\ell)} = \arg\min_{\theta \in \Sigma_K} W_\lambda^{(\ell)}(\hat{\mu}_\theta, \hat{\nu}). \tag{1.3}$$

This takes into account both the effect of entropic regularization and the influence of the number of iterations of the Sinkhorn algorithm. Our theoretical results shed some light on how the parameters λ and ℓ influence the performance of the estimator $\hat{\theta}_\lambda^{(\ell)}$. We demonstrate the practical efficiency of our method and the impact of the regularization parameter λ on simulated and real data (flow cytometry measurements). We also analyze the performance of a related estimator defined by replacing the regularized OT cost $W_\lambda(\mu, \nu)$ in (1.3) by the so-called Sinkhorn divergence

$$S_\lambda(\mu, \nu) = W_\lambda(\mu, \nu) - \frac{1}{2} \left(W_\lambda(\mu, \mu) + W_\lambda(\nu, \nu) \right),$$

introduced by Feydy et al. [2019] to remove a bias effect induced by the use of $W_\lambda(\mu, \nu)$ that has been empirically observed in machine learning problems involving the use of OT.

1.2 Related works based on regularized optimal transport

Aside from the computational benefits of entropic regularization mentioned previously, recent developments have studied the statistical properties of a regularized OT cost computed from empirical measures. Indeed, in most cases, μ and ν are not available, and one has only access to their empirical versions $\hat{\mu}$ and $\hat{\nu}$ respectively built from $X_1, ..., X_n$ sampled from μ and $Y_1, ..., Y_n$ sampled from ν. In this setting, it is natural to investigate the convergence rate of the *plug-in estimator* $W_0(\hat{\mu}, \hat{\nu})$ towards $W_0(\mu, \nu)$. This question is addressed in Fournier and Guillin [2015] where the authors proved that the resulting estimation error decays to zero at the rate $n^{-2/d}$ when using the quadratic cost in high dimension d. Due to its attractive computational efficiency, it is obviously interesting to examine the statistical efficiency of the regularized Wasserstein *plug-in estimator* naturally defined as $W_\lambda(\hat{\mu}, \hat{\nu})$. This issue as well as the approximation error induced by the regularization parameter is studied in Genevay et al. [2019]. These questions are thoroughly pursued in Chizat et al. [2020] as well as the effect of substituting $W_\lambda(\mu, \nu)$ by its debiased counterpart $S_\lambda(\mu, \nu)$. Putting the computational issues aside, the OT loss functions W_0, W_λ and S_λ also constitute efficient tools for statistical estimation. For instance, a framework of parametric estimation where regularized OT acts as a loss function in learning problems is considered in Ballu et al. [2020]. Regularized Wasserstein losses are also considered in Genevay et al. [2018], Sanjabi et al. [2018], Liu et al. [2019] for the design of generative models in image processing. In a more applied context, the use if regularized OT is investigated in Huizing et al. [2021], Freulon et al. [2020] to tackle estimation problems in biostatistics. The influence of the regularization parameter λ for the purpose of computing smooth Wasserstein barycenters is also analyzed in Bigot et al. [2018].

1.3 Organization of the paper

In Section 2 we recall the mathematical aspects of regularized OT needed to derive our results, and we detail the problem of optimal class proportions estimation in mixture models using OT. In Section 3, we introduce the various parametric Wasserstein estimators used to estimate the optimal class proportions. We also give the main results of this paper on a theoretical comparison of the convergence rates of regularized and un-regularized Wasserstein estimators. The influence of the number of iterations of the Sinkhorn algorithm on these convergence rates is also discussed. Section 4 is focused on numerical experiments that highlight the potential benefits of regularized Wasserstein estimators over un-regularized ones for appropriate choices of the entropic regularization parameter. Section 5 contains a conclusion and some perspectives. In Appendix A, we detail the main arguments to obtain the convergence rates of regularized and un-regularized Wasserstein estimators. Finally, auxiliary results to derive the proofs of convergence are gathered in technical Appendices at the end of paper.

2 Background on optimal transport and the problem of class proportions estimation

In this section, we introduce the notion of entropy regularized OT, and we present some of its mathematical properties needed to derive our results. Then, we describe the main application of this work on class proportions estimation in mixture models using OT. Finally, we discuss some identifiability issues in such models.

2.1 The OT problem and its regularized counterpart

We begin by setting some notations. In the whole paper, we will work in the space $\mathbb{R}^d$ equipped with the quadratic cost $c(x, y) = \|x - y\|^2$, where $\|x\| = \sqrt{\sum x_i^2}$ is the Euclidean norm. Let $\mathcal{X}$ and $\mathcal{Y}$ be two subsets of $\mathbb{R}^d$ that *are assumed to be compact* and included in $B(0, R) = \{x \in \mathbb{R}^d \; : \; \|x\| \leq R\}$ throughout the paper. We denote by $\mathcal{M}_+^1(\mathcal{X})$ and $\mathcal{M}_+^1(\mathcal{Y})$ the sets of probability measures on $\mathcal{X}$ and $\mathcal{Y}$ respectively. For $Y_1, \ldots, Y_n \sim \nu$, we denote by $\hat{\nu}$ the empirical counterpart of ν defined by $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$. The notation $\lesssim$ means inequality up to a multiplicative universal constant. For $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, we let $\Pi(\mu, \nu)$ be the set of probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. The problem of *entropic optimal transportation* between $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ is then defined as follows.

Definition 2.1 (Primal formulation). *For any $(\mu, \nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$, the Kantorovich formulation of the regularized optimal transport between μ and ν is the following convex minimization* *problem*

$$W_\lambda(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2 d\pi(x,y) + \lambda KL(\pi|\mu\otimes\nu), \tag{2.1}$$

where $\|x-y\|$ is the Euclidean distance between x and y, $\lambda \geq 0$ is a regularization parameter, and KL stands for the Kullback-Leibler divergence, between π and a positive measure ξ on $\mathcal{X}\times\mathcal{Y}$, up to the additive term $\int_{\mathcal{X}\times\mathcal{Y}} d\xi(x,y)$, namely

$$KL(\pi|\xi) = \int_{\mathcal{X}\times\mathcal{Y}} \log\Big(\frac{d\pi}{d\xi}(x,y)\Big) d\pi(x,y),$$

in the case π absolutely continuous w.r.t. ξ, otherwise $KL(\pi|\xi) = +\infty$. For $\lambda = 0$, the quantity $W_0(\mu,\nu)$ is the *standard (un-regularized) OT cost*, and for $\lambda > 0$, we refer to $W_\lambda(\mu,\nu)$ as the *regularized OT cost* between μ and ν. Note that the continuity of c and the compactness of $\mathcal{X}$ and $\mathcal{Y}$ imply that $W_\lambda(\mu,\nu)$ is finite for any value of $\lambda \geq 0$. Let us now introduce the dual and semi-dual formulations (see e.g. Santambrogio [2015], Genevay et al. [2016]) of the minimization problem (2.1).

Theorem 2.1 (Dual formulation)**.** *Strong duality holds for the primal problem* (2.1) *in the sense that*

$$W_\lambda(\mu,\nu) = \sup_{\substack{\varphi\in L^\infty(\mathcal{X}),\\ \psi\in L^\infty(\mathcal{Y})}} \int \varphi(x)d\mu(x) + \int \psi(y)d\nu(y) - \int m_\lambda(\varphi(x)+\psi(y)-\|x-y\|^2)d\mu(x)d\nu(y) \tag{2.2}$$

where $L^\infty(\mathcal{X})$ denotes the space of essentially bounded functions quotiented by a.e. equality, and

$$m_\lambda(t) = \begin{cases} +\infty \mathbb{1}_{\{t\geqslant 0\}} & \textit{if} \quad \lambda = 0 \\ \lambda(e^{\frac{t}{\lambda}} - 1) & \textit{if} \quad \lambda > 0 \end{cases}$$

A solution (φ,ψ) of the dual problem (2.2) is called a pair of Kantorovich potentials. Besides, since $\mathcal{X},\mathcal{Y}$ are compact and c is continuous, it follows that the dual problem admits a solution $(\varphi,\psi) \in \mathscr{C}_b(\mathcal{X})\times\mathscr{C}_b(\mathcal{Y})$. Moreover, when $\lambda > 0$, there exists solutions φ,ψ to the dual problem (2.2) which are uniquely defined almost everywhere, up to an additive constant. To introduce the semi-dual formulation, we define, for the quadratic cost $c(x,y) = \|x-y\|^2$, the regularized c-transform as in Feydy et al. [2019]: for $\lambda > 0$, we set

$$\forall x \in \mathbb{R}^d, \quad \psi_\nu^{c,\lambda}(x) = -\lambda \log \int e^{-\frac{\|x-y\|^2-\psi(y)}{\lambda}} d\nu(y), \tag{2.3}$$

and for $\lambda = 0$, the c-transform simply reads

$$\forall x \in \mathbb{R}^d, \quad \psi^c(x) = \min_{y\in\mathcal{Y}} \{\|x-y\|^2 - \psi(y)\}. \tag{2.4}$$

We also define the analogous operators for the y-variable (and for simplicity, we use the same notation for c-transforms of x-functions or y-functions). Notice that the operation used in (2.3) can be understood as a smoothed minimum that depends on ν. Therefore, when $\lambda > 0$ we will stick to the notation $\psi_\nu^{c,\lambda}$ to keep in mind the possible dependence on ν of the regularized c-transform. Notice also that, even if $\psi_\nu^{c,\lambda}$ will be integrated only on $\mathcal{X}$, the formulae allow to extrapolate the c-transforms on the whole space $\mathbb{R}^d$. In the sequel of this paper we extrapolate the c-transform on $B(0,R)$ to manipulate functions defined on a convex subset of $\mathbb{R}^d$ without imposing the convexity of $\mathcal{X}$ and $\mathcal{Y}$.

Definition 2.2 (Semi-dual formulation)**.** *The dual problem* (2.2) *is equivalent to the following semi-dual problem in the sense that*

$$W_\lambda(\mu,\nu) = \sup_{\psi\in L^\infty(\mathcal{Y})} \int \psi_\nu^{c,\lambda}(x)d\mu(x) + \int \psi(y)d\nu(y). \tag{2.5}$$

A solution ψ of the semi-dual problem is called a *Kantorovich potential*. In other words, ψ is a Kantorovich potential if and only if $(\psi_\nu^{c,\lambda}, \psi)$ is a pair of Kantorovich potentials. By symmetry, we can also formulate a semi-dual problem on the dual variable φ. For discrete probability distributions, the iterative Sinkhorn algorithm, as defined below, (see e.g. Cuturi [2013]) allows to approximate the regularized OT cost $W_\lambda(\mu,\nu)$ as follows.

Definition 2.3. *For $\mu = \sum_{i=1}^I \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^J \nu_j \delta_{y_j}$ two discrete distributions on $\mathbb{R}^d$, the approximation of the regularized OT cost returned by the Sinkhorn approximation after ℓ iterations equals*

$$W_\lambda^{(\ell)}(\mu,\nu) = \sum_{i=1}^I \mu_i \varphi_i^{(\ell)} + \sum_{j=1}^J \nu_j \psi_j^{(\ell)}. \tag{2.6}$$

The variables $\varphi^{(\ell)}$ and $\psi^{(\ell)}$ being the dual variables returned after ℓ iterations of the Sinkhorn algorithm. Starting from $\psi^0 = 0_J \in \mathbb{R}^J$, the Sinkhorn ℓ^{th} iteration is defined by the update of the dual variables:

$$\begin{aligned} \varphi_i^{(\ell)} &= -\lambda \log\left(\textstyle\sum_{j=1}^J \exp\left(-\frac{\|x_i-y_j\|^2-\psi_j^{(\ell-1)}}{\lambda}\right)\nu_j\right) \\ \psi_j^{(\ell)} &= -\lambda \log\left(\textstyle\sum_{i=1}^I \exp\left(-\frac{\|x_i-y_j\|^2-\varphi_i^{(\ell)}}{\lambda}\right)\mu_i\right). \end{aligned} \tag{2.7}$$

Then, the de-biased version of the regularized OT cost, also known as the Sinkhorn divergence, is defined as follows.

Definition 2.4 (Sinkhorn divergence, from Feydy et al. [2019]). *Let $\lambda > 0$. For any $(\mu,\nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$, the Sinkhorn divergence between μ and ν is defined as*

$$S_\lambda(\mu,\nu) = W_\lambda(\mu,\nu) - \frac{1}{2}\left(W_\lambda(\mu,\mu) + W_\lambda(\nu,\nu)\right). \tag{2.8}$$

Since we only consider here the quadratic cost $c(x,y) = \|x-y\|^2$, we will be able to use important properties of the Sinkhorn divergence S_λ established by Chizat et al. [2020] in order to derive our results on the convergence of Wasserstein estimators.

2.2 An alternative dual problem

We now introduce an alternative dual formulation of regularized OT that is specific to the quadratic cost. This alternative dual problem is restricted to a class of Kantorovich potentials that are concave and Lipschitz functions, which proves useful to derive some of the convergence rates given in Section 3. The relation between those dual problems has already been explicited for unregularized OT (for example in Chizat et al. [2020]), and we extend it to the regularized case. Let $\lambda \geq 0$. By expanding the squared Euclidean cost, we have for any $\pi \in \Pi(\mu,\nu)$,

$$\begin{aligned} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2 d\pi(x,y) + \lambda \mathrm{KL}(\pi|\mu\otimes\nu) &= \int_{\mathcal{X}} \|x\|^2 d\mu(x) + \int_{\mathcal{Y}} \|y\|^2 d\nu(y) \qquad (2.9) \\ &\quad - 2\int_{\mathcal{X}\times\mathcal{Y}} \langle x,y\rangle d\pi(x,y) + \lambda \mathrm{KL}(\pi|\mu\otimes\nu). \qquad (2.10) \end{aligned}$$

The above decomposition leads us to consider the new regularized transportation problem

$$W_\lambda^s(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} s(x,y)d\pi(x,y) + \lambda \mathrm{KL}(\pi|\mu\otimes\nu), \tag{2.11}$$

with $s(x,y) = -2\langle x,y\rangle$. First, we remark that the standard regularized Wasserstein distance $W_\lambda(\mu,\nu)$ and the alternative regularized Wasserstein distance $W_\lambda^s(\mu,\nu)$ are related through the relation

$$W_\lambda(\mu,\nu) = \int_{\mathcal{X}} \|x\|^2 d\mu(x) + \int_{\mathcal{Y}} \|y\|^2 d\nu(y) + W_\lambda^s(\mu,\nu). \tag{2.12}$$

A dual formulation associated to the problem (2.11) is given by the next proposition.

Proposition 2.1. *The dual problem associated to (2.11) writes as*

$$W_\lambda^s(\mu,\nu) = \sup_{\substack{\varphi\in L^\infty(\mathcal{X})\\ \psi\in L^\infty(\mathcal{Y})}} \int \varphi(x)d\mu(x) + \int \psi(y)d\nu(y) - \int m_\lambda(\varphi(x)+\psi(y)+2\langle x,y\rangle)d\mu(x)d\nu(y), \quad (2.13)$$

where m_λ is defined in Theorem 2.1.

Proof. The key argument is to remark that (2.11) is a regularized optimal transportation problem with cost function $s(x,y) = -2\langle x,y\rangle$. Hence, as $\mathcal{X}$ and $\mathcal{Y}$ are assumed to be compact and s continuous, it follows that strong duality holds (see e.g. Genevay et al. [2016], Bercu and Bigot [2021]) in the sense of equation (2.13). □

Fort the cost function $s(x,y) = -2\langle x,y\rangle$, we can also define a s-transform and a semi-dual problem as follows. For the cost $s(x,y) = -2\langle x,y\rangle$ and for $\varphi \in L^\infty(\mathcal{X})$ the s-transform is defined as

$$\forall y \in \mathbb{R}^d, \quad \varphi_\mu^{s,\lambda}(y) = \begin{cases} -\lambda\log\left(\int \exp\left(\frac{\varphi(x)+2\langle x,y\rangle}{\lambda}\right)d\mu(x)\right), & \text{for } \lambda > 0, \\ -\max\limits_{x\in\mathcal{X}}(\varphi(x)+2\langle x,y\rangle) & \text{for } \lambda = 0. \end{cases} \quad (2.14)$$

By the above s-transform in the dual problem (2.13) we obtain the following semi-dual formulation

$$\sup_{\varphi\in L^\infty(\mathcal{X})} \int \varphi(x)d\mu(x) + \int \varphi_\mu^{s,\lambda}(y)d\nu(y). \quad (2.15)$$

We conclude this section by studying some properties of this s-transform.

Proposition 2.2. *For $\lambda \geq 0$, the s-transform $\varphi_\mu^{s,\lambda}$ is concave and R-Lipschitz on $B(0,R)$.*

Proof. We start with the concavity of $\varphi_\mu^{s,\lambda}$. For $\lambda = 0$, it follows from the fact that a maximum of convex functions is convex. Now, for $\lambda > 0$, $y_1, y_2 \in \mathcal{Y}$ and $t \in (0,1)$, we have

$$\begin{aligned} &\int_{\mathcal{X}} \exp\left(\frac{\varphi(x)+\langle x, ty_1+(1-t)y_2\rangle}{\lambda}\right)d\mu(x) \\ &\quad= \int_{\mathcal{X}} \exp\left(t\frac{\varphi(x)+2\langle x,y_1\rangle}{\lambda}\right)\exp\left((1-t)\frac{\varphi(x)+2\langle x,y_2\rangle}{\lambda}\right)d\mu(x) \\ &\quad\leq \left(\int_{\mathcal{X}} \exp\left(\frac{\varphi(x)+2\langle x,y_1\rangle}{\lambda}\right)d\mu(x)\right)^t \left(\int_{\mathcal{X}} \exp\left(\frac{\varphi(x)+2\langle x,y_2\rangle}{\lambda}\right)d\mu(x)\right)^{1-t}, \end{aligned}$$

thanks to Hölder inequality with exponents $p = 1/t$ and $p' = 1/(1-t)$. Applying $-\lambda\log$ on both sides gives directly

$$\varphi_\mu^{s,\lambda}(ty_1+(1-t)y_2) \geq t\varphi_\mu^{s,\lambda}(y_1) + (1-t)\varphi_\mu^{s,\lambda}(y_2). \quad (2.16)$$

Now, we will see as in Feydy et al. [2019] that the regularized s-transform inherits the Lipschitz constant of the cost. For $y_1, y_2 \in \mathcal{Y}$ and $x \in \mathcal{X}$, we have $|\langle x, y_1-y_2\rangle| \leq R\|y_1-y_2\|$ thanks to Cauchy-Schwarz inequality, and thus

$$\varphi(x)+2\langle x,y_1\rangle \leq R\|y_1-y_2\| + \varphi(x)+2\langle x,y_2\rangle.$$

Taking $\lambda\log\int_{\mathcal{X}}\exp(\frac{\cdot}{\lambda})d\mu(x)$ for $\lambda > 0$ (resp. $\max_{x\in\mathcal{X}}$ for $\lambda = 0$) on both sides gives

$$\varphi_\mu^{s,\lambda}(y_2) \leq R\|y_1-y_2\| + \varphi_\mu^{s,\lambda}(y_1)\ .$$

By symmetry, we get $|\varphi_\mu^{s,\lambda}(y_1) - \varphi_\mu^{s,\lambda}(y_2)| \leq R\|y_1-y_2\|$. □

2.3 Class proportions estimation in mixture models

Let $\mu = \sum_{k=1}^{K} \pi_k \mu_k$ be a probability measure that can be decomposed as a mixture of K probability measures $\mu_1, \ldots, \mu_K$ in $\mathcal{M}_+^1(\mathcal{X})$. For $\theta \in \Sigma_K = \{(\theta_1, \ldots, \theta_K) \in \mathbb{R}_+^K \ : \ \sum_{k=1}^{K} \theta_k = 1\}$, the re-weighted version of μ is defined as

$$\mu_\theta = \sum_{k=1}^{K} \theta_k \mu_k. \tag{2.17}$$

Let ν be another probability measure in $\mathcal{M}_+^1(\mathcal{Y})$ referred to as the target distribution. The problem of class proportions estimation consists in estimating an optimal weighting vector

$$\theta^* \in \arg\min_{\theta \in \Sigma_K} W_0(\mu_\theta, \nu) \tag{2.18}$$

from empirical versions of the $\mu_1, \ldots, \mu_K$ and ν. In what follows, we discuss some properties of the optimisation problem (2.18). First, this minimization problem is motivated by the implicit assumption that representing the target measure ν as a mixture of K probability measures is relevant. To illustrate this point, we first state a result showing that one can recover the true class proportions in the ideal setting where the target distribution ν is also a mixture of $\mu_1, \ldots, \mu_K$.

Lemma 2.1. *Suppose that μ_θ and ν are mixtures of probability measures with the same components $\mu_1, ..., \mu_K$ but with different class proportions, respectively denoted by $\theta \in \Sigma_K$ and by $\tau \in \Sigma_K$. If the model $\left\{\mu_\theta = \sum_{k=1}^{K} \theta_k \mu_k \mid \theta \in \Sigma_K\right\}$ is identifiable (in the sense that the mapping $\theta \mapsto \mu_\theta$ is injective), then the solution of optimization problem (2.18) is unique and one has that $\theta^* = \tau$.*

Proof. The non-negativity property of W_0 ensures that for all $\theta \in \Sigma_K$, $W_0(\mu_\theta, \nu) \geq 0$. Next, for $\theta \in \Sigma_K$,

$$\begin{aligned} W_0(\mu_\theta, \nu) = 0 &\Leftrightarrow \sum_{k=1}^{K} \theta_k \mu_k = \sum_{k=1}^{K} \tau_k \mu_k \\ &\Leftrightarrow \theta = \tau, \end{aligned}$$

where the last equivalence comes from the assumption that the model $\{\mu_\theta \mid \theta \in \Sigma_K\}$ is identifiable. From this result, we deduce $\arg\min_{\theta \in \Sigma_K} W_0(\mu_\theta, \nu) = \{\tau\}$. □

Notice that the injectivity of $\theta \mapsto \mu_\theta$ relates to the affine independence of $\{\mu_1, \ldots, \mu_K\}$. It is satisfied for example when the measures $\mu_1, \ldots, \mu_K$ have disjoint supports. If all the scenarios are not as friendly as the one considered in Lemma 2.1, in numerous applications (for instance when the data can be clustered into sub-populations), it is relevant to approximate the distribution ν as a mixture. The next result is about the smoothness of the minimization problem (2.18).

Lemma 2.2. *Suppose that μ_θ is defined as in (1.2). Then, the function $F : \left\{ \begin{array}{ccc} \Sigma_K & \to & \mathbb{R}_+ \\ \theta & \mapsto & W_0(\mu_\theta, \nu) \end{array} \right.$ is continuous on Σ_K.*

Proof. Let $\theta \in \Sigma_K$ and $(\theta^{(n)})$ a sequence in Σ_K that converges to θ. Then, the probability sequence $(\mu_{\theta^{(n)}})$ converges weakly toward μ_θ. Indeed, for any bounded continuous function φ, one has that $\int \varphi d\mu_{\theta^{(n)}} = \sum_{k=1}^{K} \theta_k^{(n)} \int \varphi d\mu_k$. As $\theta^{(n)} \to_{n \to \infty} \theta$, it follows that

$$\sum_{k=1}^{K} \theta_k^{(n)} \int \varphi d\mu_k \to \sum_{k=1}^{K} \theta_k \int \varphi d\mu_k. = \int \varphi d\mu_\theta.$$

Hence, $(\mu_{\theta^{(n)}})$ weakly converges towards μ_θ. Then, one can also verify that the sequence $(\theta^{(n)})$ is such that for any $x_0 \in \mathcal{X}$, $\int \|x_0 - x\|^2 d\mu_{\theta^{(n)}}$ converges toward $\int \|x_0 - x\|^2 d\mu_\theta$. Therefore, by Corollary 6.11 in Villani [2009], it follows that

$$W_0(\mu_{\theta^{(n)}}, \nu) \to W_0(\mu_\theta, \nu),$$

which shows the continuity of $F : \theta \mapsto W_0(\mu_\theta, \nu)$. □

Since the set Σ_K is compact, the existence of a minimizer of the optimization problem (2.18) follows from Lemma 2.2. We now give sufficient conditions that ensure the strict convexity of the objective function $\theta \mapsto W_0(\mu_\theta, \nu)$.

Lemma 2.3. *Assume that ν is absolutely continuous with respect to the Lebesgue measure. Then, if the model $\{\mu_\theta \mid \theta \in \Sigma_K\}$ is identifiable (in the sense that the mapping $\theta \mapsto \mu_\theta$ is injective), the function $F : \left\{ \begin{array}{ccc} \Sigma_K & \to & \mathbb{R}_+ \\ \theta & \mapsto & W_0(\mu_\theta, \nu) \end{array} \right.$ is strictly convex.*

Proof. Thanks to the assumption that ν is absolutely continuous, Proposition 7.19 in Santambrogio [2015] ensures the strict convexity of the functional $\mu \mapsto W_0(\mu, \nu)$. Let $\theta_0, \theta_1 \in \Sigma_K$ with $\theta_0 \neq \theta_1$ and $t \in (0, 1)$. Then, we have that $F(t\theta_0 + (1-t)\theta_1) = W_0(\mu_{t\theta_0+(1-t)\theta_1}, \nu)$, and $\mu_{t\theta_0+(1-t)\theta_1} = t\mu_{\theta_0} + (1-t)\mu_{\theta_1}$. Since $\theta_0 \neq \theta_1$ and the model $\{\mu_\theta \mid \theta \in \Sigma_K\}$ is supposed to be identifiable, we have that $\mu_{\theta_0} \neq \mu_{\theta_1}$. Therefore, the strict convexity of $\mu \mapsto W_0(\mu, \nu)$ yields

$$W_0(t\mu_{\theta_0} + (1-t)\mu_{\theta_1}, \nu) < tW_0(\mu_{\theta_0}, \nu) + (1-t)W_0(\mu_{\theta_1}, \nu).$$

which proves the strict convexity of $F : \theta \mapsto W_0(\mu_\theta, \nu)$.

□

3 Parametric Wasserstein estimators

In this section, we present the regularized and un-regularized parametric Wasserstein estimators that are considered in this paper, and we compare their convergence rates.

3.1 Definition of the estimators

We aim at estimating θ^* when the distributions μ and ν are only observed through samples. Hence, we assume given the following empirical measures (as defined in Section 1.1)

$$\hat{\mu} = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_k, \text{ where } \hat{\pi}_k = \frac{m_k}{m}, \quad \text{and} \quad \hat{\nu} = \frac{1}{n} \sum_{j=1}^{n} \delta_{Y_j},$$

where each component $\hat{\mu}_k$ corresponds to a known sub-population of cells of size m_k in the source sample $X_1, \ldots, X_m$. Moreover, we recall that $\hat{\mu}_\theta = \sum_{k=1}^{K} \theta_k \hat{\mu}_k$. denotes the empirical version of the re-weighted measure μ_θ.

We can now define the various Wasserstein estimators whose convergence properties are discussed in Section 3.2. Depending on the chosen loss function (either W_λ, S_λ or W_0) and using the empirical measures $\hat{\mu}_\theta$ and $\hat{\nu}$, three estimators of the optimal vector of class proportions can be defined as follows:

$$\hat{\theta}_\lambda \in \widehat{\Theta}_\lambda := \underset{\theta \in \Sigma_K}{\arg\min}\, W_\lambda(\hat{\mu}_\theta, \hat{\nu}), \quad \hat{\theta}^S_\lambda \in \widehat{\Theta}^S_\lambda := \underset{\theta \in \Sigma_K}{\arg\min}\, S_\lambda(\hat{\mu}_\theta, \hat{\nu}), \text{ and } \hat{\theta}_0 \in \widehat{\Theta}_0 := \underset{\theta \in \Sigma_K}{\arg\min}\, W_0(\hat{\mu}_\theta, \hat{\nu}). \tag{3.1}$$

When considering entropy regularized OT, we also propose to study the estimators that are obtained with the Sinkhorn algorithm on the sample distributions after a limited number ℓ of iterations, that are

$$\hat{\theta}^{(\ell)}_\lambda \in \widehat{\Theta}^{(\ell)}_\lambda := \underset{\theta \in \Sigma_K}{\arg\min}\, W^{(\ell)}_\lambda(\hat{\mu}_\theta, \hat{\nu}), \quad \text{and} \quad \hat{\theta}^{S(\ell)}_\lambda \in \widehat{\Theta}^{S(\ell)}_\lambda := \underset{\theta \in \Sigma_K}{\arg\min}\, S^{(\ell)}_\lambda(\hat{\mu}_\theta, \hat{\nu}).$$

In this paper, to assess the performance of a given estimator $\hat{\theta}$ of θ^* based on n samples, we shall consider the following expected excess risk defined as

$$r_n(\mu_{\hat{\theta}}, \nu) = \mathbb{E}\big[W_0(\mu_{\hat{\theta}}, \nu) - W_0(\mu_{\theta^*}, \nu)\big]. \tag{3.2}$$

Remark 3.1. *In our context of parametric Wasserstein estimation, we can interpret the excess risk as the representation error of ν induced by the estimator. Indeed, μ_{θ^*} defined in equation (2.18) is the best representation of ν in the model $\{\mu_\theta \mid \theta \in \Sigma_K\}$ w.r.t. the Wasserstein distance. And, $W_0^{1/2}$ being a distance, under the assumption that the function $\theta \mapsto W_0(\mu_\theta, \nu)$ is bounded on Σ_K, we can write*

$$0 \leq W_0(\mu_{\hat{\theta}}, \nu) - W_0(\mu_{\theta^*}, \nu) \lesssim W_0^{1/2}(\mu_{\hat{\theta}}, \mu_{\theta^*}).$$

This equation shows that the excess risk is closely related to Wasserstein distance between the best representation of ν in the model that is μ_{θ^*} and its estimated version $\mu_{\hat{\theta}}$.

Remark 3.2. *In the case where the function $\theta \mapsto W_0(\mu_\theta, \nu)$ is strongly convex w.r.t. θ^*, that is, when*

$$W_0(\mu_\theta, \nu) - W_0(\mu_{\theta^*}, \nu) \geq \rho \|\theta - \theta^*\|_2^2, \quad \text{for some } \rho > 0 \text{ and all } \theta \in \Sigma_K,$$

then controlling the expected excess risk allows to derive a convergence rate on the expected quadratic risk $\mathbb{E}\left[\|\hat{\theta} - \theta^*\|_2^2\right]$.

Remark 3.3. *We have chosen to derive rates of convergence for the excess risk (3.2) as our statistical analysis allows to treat general classes of parametric Wasserstein estimators that go beyond the setting of class proportions in mixture models considered in this paper. Indeed, our results can be applied to the study of regularized Wasserstein estimators within any parametric family $\mathcal{F} = \{\mu_\theta, \ \theta \in \Theta\}$ of probability measures with compact support in $B(0, R)$ provided that the mapping $\theta \mapsto \mu_\theta$ is continuous. In particular, our approach could be used to extend existing results by G. Biau and M. Sangnier and U. Tanielian [2021] on the statistical analysis of un-regularized Wasserstein Generative Adversarial Networks (WGAN) to the case of entropy regularized WGAN considered by Sanjabi et al. [2018].*

In Section 3.2, we present upper bounds on the above expected risk for the proposed estimators. When the regularization parameter λ is involved, we also propose a decreasing choice $(\lambda_n)_{n \geq 0}$ of its value to ensure that the resulting estimator has an expected excess risk that goes to zero when $n \to +\infty$.

3.2 Convergence rates for the expected excess risk

We now compare the convergence rate of the various estimators introduced in Section 3.1. For every estimator $\hat{\theta}$ considered in this work, to derive rate of convergence toward θ^*, we need the probability distributions to have compact supports.

Assumption A.1. *The supports of μ and ν, respectively denoted by $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of $B(0, R) = \{x \ : \ \|x\| \leq R\}$ for some $R > 0$.*

In the next Section 3.2.1, we introduce two additional assumptions in order to exploit the approximation result established in Chizat et al. [2020] between the Sinkhorn divergence $S_\lambda(\mu, \nu)$ and the Wasserstein distance $W_0(\mu, \nu)$. Indeed, Theorem 1 in Chizat et al. [2020], that we remind in Theorem 3.1 of this paper, allows to bound $|S_\lambda(\mu, \nu) - W_0(\mu, \nu)|$ with a constant that depends on λ, the standard Fisher information $I(\mu)$, $I(\nu)$ of μ, ν, and $I(\mu, \nu)$ that is the Fisher information of the Wasserstein geodesic between μ and ν defined as in Chizat et al. [2020].

3.2.1 Fisher information and approximation error of the Sinkhorn divergence $S_\lambda(\mu, \nu)$

We discuss conditions that ensure a control of the approximation error between the Sinkhorn divergence $S_\lambda(\mu, \nu)$ and $W_0(\mu, \nu)$. These conditions will prove to be useful when considering S_λ as a loss function in Section A.2, for instance in Proposition A.8.

Theorem 3.1. *[Chizat et al., 2020, Theorem 1] Suppose that μ and ν have bounded densities and supports. Then, it holds that*

$$|S_\lambda(\mu, \nu) - W_0(\mu, \nu)| \leq \frac{\lambda^2}{4} \max\{I(\mu, \nu), (I(\mu) + I(\nu))/2\}, \tag{3.3}$$

where $I(\mu)$ refers to the standard Fisher information of μ, and $I(\mu,\nu)$ is the Fisher information of the Wasserstein geodesic between μ and ν as defined in Chizat et al. [2020].

First, we introduce sufficient conditions to ensure that the Fisher information of μ_θ can be upper bounded without dependence on θ.

Assumption A.2. *The probability distributions $\mu_1,\dots,\mu_K$ have finite Fisher information with respective densities $f_1,\dots,f_K$ w.r.t. the Lebesgue measure, and all the components $\mu_1,\dots,\mu_K$ have disjoint supports $\mathcal{X}_1,\dots,\mathcal{X}_K$.*

Proposition 3.1. *Suppose that assumption A.2 holds. Then, one has that*

$$\forall\theta\in\Sigma_K,\ I(\mu_\theta)\leq\max_{k\in\{1,\dots,K\}} I(\mu_k). \tag{3.4}$$

Proof. For simplicity, we first consider the case $d=1$. Let $\theta\in\Sigma_K$. Then, using the assumption that the components μ_k have disjoint supports, we decompose the Fisher information of μ_θ as follows

$$\begin{aligned}I(\mu_\theta)&=\int_{\mathcal{X}}\left(\frac{f_\theta'(x)}{f_\theta(x)}\right)^2 f_\theta(x)dx=\sum_{k=1}^K\theta_k\int_{\mathcal{X}_k}\left(\frac{f_\theta'(x)}{f_\theta(x)}\right)^2 f_k(x)dx\\&=\sum_{k=1}^K\theta_k\int_{\mathcal{X}_k}\left(\frac{\theta_k f_k'(x)}{\theta_k f_k(x)}\right)^2 f_k(x)dx=\sum_{k=1}^K\theta_k I(\mu_k)\leq\max_{k\in\{1,\dots,K\}} I(\mu_k),\end{aligned}$$

which proves Inequality (3.4) for $d=1$. The case $d>1$ can be treated analogously. □

Next, in order to bound the Fisher information of the Wasserstein geodesic between μ_θ and ν with a constant independent of θ, we adapt the assumptions of Proposition 1 from Chizat et al. [2020] to our needs.

Assumption A.3. *The probability distribution ν is absolutely continuous with respect to the Lebesgue measure. Moreover, there exist two constants $m>0$ and $L>0$ such that for all $\theta\in\Sigma_K$ the Brenier potential φ_θ between μ_θ and ν has a L-Lipschitz continuous Hessian satisfying $m\,\mathrm{Id}\leq\nabla^2\varphi_\theta$.*

Proposition 3.2. *Suppose that Assumptions A.2 and A.3 hold. Then, we have the following inequality*

$$\forall\theta\in\Sigma_K,\ I(\mu_\theta,\nu)\leq\frac{2}{m}\left(\max_{k\in\{1,\dots,K\}} I(\mu_k)+\frac{L^2}{3m^2}\right). \tag{3.5}$$

Proof. A straight application of Proposition 1 from Chizat et al. [2020] gives that

$$\forall\theta\in\Sigma_K,\ I(\mu_\theta,\nu)\leq\frac{2}{m}\left(I(\mu_\theta)+\frac{L^2}{3m^2}\right).$$

Then, using $I(\mu_\theta)\leq\max_{k\in\{1,\dots,K\}} I(\mu_k)$ from Proposition 3.1 yields inequality (3.5). □

As a consequence of Theorem 3.1, we have the following result.

Corollary 3.1. *Suppose that Assumptions A.2 and A.3 hold. Then, we have that there exists a constant $M_I>0$ such that*

$$\forall\theta\in\Sigma_K,\ |S_\lambda(\mu_\theta,\nu)-W_0(\mu_\theta,\nu)|\leq M_I\lambda^2, \tag{3.6}$$

where

$$M_I=\frac{1}{4}\max\left\{\frac{2}{m}\left(\max_{k\in\{1,\dots,K\}} I(\mu_k)+\frac{L^2}{3m^2}\right),\frac{\max_{k\in\{1,\dots,K\}} I(\mu_k)+I(\nu)}{2}\right\}. \tag{3.7}$$

Proof. The combination of the upper bounds on $I(\mu_\theta)$ and $I(\mu_\theta,\nu)$ established in Proposition 3.1 and in Proposition 3.2 respectively, as well as the upper bound of Theorem 3.1 yields (3.6). □

3.2.2 Comparison between regularized and un-regularized Wasserstein estimators

We now give the main result of this paper on a comparison of the convergence rates of the three estimators defined by (3.1). For the sake of clarity, we state this result under Assumptions A.1, A.2 and A.3, but it should be noted that the convergence rates for $\hat{\theta}_\lambda$ and $\hat{\theta}_0$ hold under weaker assumptions as detailed later on in Section A. In order to present a concise result, we also assume that the dimension $d > 4$ is even (we write $2\lfloor d/2 \rfloor$ in the general case). We refer the reader to Sections A.1, A.2 and A.3 for precise convergence rates that hold for any $d \geq 1$. The notation $\lesssim$ means inequality up to a multiplicative universal constant.

As classically done in nonparametric statistics, it is natural to decompose the excess risk of an estimator into an estimation error and an approximation error that need to be balanced to derive an optimal choice of the regularization parameter $\lambda \to 0$ as the number of observations tends to infinity. For example, it follows from Lemma A.1 that the excess risk of $\hat{\theta}_\lambda$ is upper bounded as follows

$$W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 2 \underbrace{\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)|}_{\text{Estimation error}} + 2 \underbrace{\sup_{\theta \in \Sigma_K} |W_0(\mu_\theta, \nu) - W_\lambda(\mu_\theta, \nu)|}_{\text{Approximation error}}$$

Then, a first approach to derive an upper bound for the above estimation error is to rely on recent results on the sample complexity of regularized OT from Genevay et al. [2019] and Chizat et al. [2020]. Adapting arguments from these works, we prove that, for a fixed value of λ, the estimation error decays at the parametric rate $n^{-1/2}$ *but with a multiplicative constant that scales at the rate* $\lambda^{-d/2}$. Using this approach, we first arrive at the following result.

Theorem 3.2. *Suppose that Assumptions A.1, A.2 and A.3 hold and that $d > 4$ is even. If for all component μ_k as well as for ν, at least n observations are available, then the two regularized estimators $\hat{\theta}_\lambda$ and $\hat{\theta}^S_\lambda$ defined in (3.1) have the following non-asymptotic rates of convergence:*

(i)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-\frac{1}{d+2}} \log(n), \quad \textit{for} \quad \lambda_n = n^{-\frac{1}{d+2}},$$

(ii)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-\frac{2}{d+4}}, \quad \textit{for} \quad \lambda_n = n^{-\frac{1}{d+4}},$$

To derive these rates of convergence, we exploit the bounds on the expected excess risk of the estimation $\hat{\theta}_\lambda$ and $\hat{\theta}^S_\lambda$ respectively established in Proposition A.6 and Proposition A.8. For these two regularized estimators, we choose λ in order to make a trade-off between the estimation error and the approximation error induced by the parameter λ. To be more specific, we now give a sketch of proof for point (i) and (ii) of Theorem 3.2 by indulging ourselves to exploit the results of Proposition A.6 and Proposition A.8.

Proof. Under the assumptions of Theorem 3.2, using Proposition A.6, for λ small enough, the expected excess risk of $\hat{\theta}_\lambda$ can be bounded as follows

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim \frac{1}{\sqrt{n}\lambda^{d/2}} + \lambda \log(\lambda^{-1}),$$

where the first term corresponds to an estimation error and the second term to an approximation error. To balance those two terms, we are led to choose $\lambda_n = n^{-\frac{1}{d+2}}$. And with this choice of regularization parameter, we recover the rate of convergence announced in point (i) of Theorem 3.2. The line of argumentation is similar for point (ii). With the current assumptions, Proposition A.8 gives

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim \frac{1}{\sqrt{n}\lambda^{d/2}} + \lambda^2.$$

Then, to balance the two terms of the previous equation, we choose $\lambda_n = n^{-\frac{1}{d+4}}$, and one can check that this choice leads to the rate of convergence of point (ii). □

Theorem 3.2 suggests that, for an optimal choice of $\lambda_n \to 0$, the estimators $\hat{\theta}_{\lambda_n}$ and $\hat{\theta}^S_{\lambda_n}$ converge at the sub-optimal rates $\mathcal{O}(n^{-\frac{1}{d+2}} \log(n))$ and $\mathcal{O}(n^{-\frac{2}{d+4}})$ respectively when compared to the convergence rate $\mathcal{O}(n^{-\frac{2}{d}})$ of the un-regularized estimator $\hat{\theta}_0$. Nevertheless, one can derive better convergence rates for the excess risk of $\hat{\theta}_{\lambda_n}$ and $\hat{\theta}^S_{\lambda_n}$ by choosing to upper bound the estimation error by a quantity that decays at the rate $n^{-2/d}$ *but with a multiplicative constant that is independent of* λ. To the best of our knowledge, this result is new, and it follows from the alternative dual formulation introduced in Section 2.2 and Proposition 2.1 as detailed in the proof of Theorem 3.3 below.

Theorem 3.3. *Suppose that Assumptions A.1, A.2 and A.3 hold and that $d > 4$. If for all component μ_k as well as for ν, at least n observations are available, then the following non-asymptotic rates of convergence hold*

(i)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-\frac{2}{d}} \log(n), \quad \textit{for} \quad \lambda_n = n^{-2/d},$$

(ii)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-\frac{2}{d}}, \quad \textit{for} \quad \lambda_n = n^{-\frac{1}{d}},$$

(iii)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_0}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-\frac{2}{d}}.$$

Proof. Let us begin with point (i) of Theorem 3.3. Under our assumptions, using Proposition A.10, and arguing as in the proof of Theorem 3.2, we have that for λ small enough, the expected excess risk of $\hat{\theta}_\lambda$ can be bounded as follows

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-2/d} + \lambda \log(\lambda^{-1}).$$

To balance the above estimation and approximation errors, we are led to choose $\lambda_n = n^{-2/d}$, and we recover the rate of convergence announced in point (i) of Theorem 3.3. Similarly, using Proposition A.12 with our assumptions, we obtain that

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-2/d} + \lambda^2.$$

Therefore, by choosing $\lambda_n = n^{-\frac{1}{d}}$, we obtain the convergence rate claimed in Theorem 3.3 (ii). As point (iii) derives from the un-regularized case, there is no approximation error. Hence, this third point does not require any additional effort and is just the rewriting of Proposition A.10 for $d > 4$ and $\lambda = 0$. □

Hence, Theorem 3.3 shows that, for the choice $\lambda_n = n^{-\frac{1}{d}}$, the estimator $\hat{\theta}^S_{\lambda_n}$ (based on the Sinkhorn divergence) achieves the same convergence rate than the un-regularized estimator $\hat{\theta}_0$. Up to a $\log(n)$ factor, the estimator $\hat{\theta}_{\lambda_n}$ also achieves the same convergence rate than $\hat{\theta}_0$ but at the price of taking a much smaller value of $\lambda_n = n^{-\frac{2}{d}}$, which impacts the convergence of the Sinkhorn algorithm used to compute these regularized estimators as detailed in the following section.

3.2.3 Influence of the number of iterations in the Sinkhorn algorithm

We now discuss the convergence rate of entropy regularized estimators when taking into account the computational complexity of the Sinkhorn algorithm through its number of iterations. To this end, we recall that $W_\lambda^{(\ell)}(\mu,\nu)$ denotes the approximation of the regularized Wasserstein distance that is returned after ℓ iterations of the Sinkhorn algorithm. We also introduce the following approximation of the Sinkhorn divergence

$$S_\lambda^{(\ell)}(\mu,\nu) = W_\lambda^{(\ell)}(\mu,\nu) - \frac{1}{2}(W_\lambda^{(\ell)}(\mu,\mu) + W_\lambda^{(\ell)}(\nu,\nu))$$

after ℓ iterations. As in Section 3.2.2, we first present results derived by bounding the estimator error using the sample complexity of regularized OT that decays, for a fixed value of λ, at the parametric rate $n^{-1/2}$ but with a multiplicative constant that scales at the rate $\lambda^{-d/2}$. For ease of reading, the theorem is given for even d, and we refer to Section A.5 for a more general statement.

Theorem 3.4. *Suppose that Assumptions A.1, A.2 and A.3 hold, and suppose d even. We also suppose that, for all $k \in \{1,\dots,K\}$, m_k samples from μ_k are available, and that n samples are available from ν. Let $\underline{m} = \min(m_1,\dots,m_K)$. Then, for $\hat\theta_\lambda^{(\ell)} = \underset{\theta\in\Sigma_K}{\arg\min}\, W_\lambda^{(\ell)}(\hat\mu_\theta,\hat\nu)$ we have that,*

(i)

$$\mathbb{E}\left[W_0(\mu_{\hat\theta_\lambda^{(\ell)}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim \underbrace{\frac{4M_\lambda}{\sqrt{\underline{m}}} + \frac{4M_\lambda}{\sqrt{n}}}_{\textit{Estimation error}} + \underbrace{8d\lambda\log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right)}_{\textit{Approximation error}} + \underbrace{\frac{32R^4}{\lambda\ell}}_{\textit{Algorithm error}}$$

while for $\hat\theta_\lambda^{S(\ell)} = \underset{\theta\in\Sigma_K}{\arg\min}\, S_\lambda^{(\ell)}(\hat\mu_\theta,\hat\nu)$ we have that,

$$\mathbb{E}\left[W_0(\mu_{\hat\theta_\lambda^{S(\ell)}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim \underbrace{\frac{8M_\lambda}{\sqrt{\underline{m}}} + \frac{8M_\lambda}{\sqrt{n}}}_{\textit{Estimation error}} + \underbrace{8M_I\lambda^2}_{\textit{Approximation error}} + \underbrace{\frac{64R^4}{\lambda\ell}}_{\textit{Algorithm error}},$$

where in both cases, $M_\lambda = M_d \max\left(R^2, \frac{R^{\frac{d}{2}+1}}{\lambda^{\frac{d}{2}}}\right)$, M_d is a constant that depends only on d and M_I is a constant defined in equation (3.7).

(ii) *Making the additional assumptions that, for each μ_k, at least n samples are available, we can propose a choice of λ_n and a number of Sinkhorn iterations ℓ_n for the two estimators $\hat\theta_{\lambda_n}^{(\ell_n)}$ and $\hat\theta_{\lambda_n}^{S(\ell_n)}$ to recover the convergence rates of Theorem 3.2.*

First, for $\hat\theta_{\lambda_n}^{(\ell_n)} = \underset{\theta\in\Sigma_K}{\arg\min}\, W_{\lambda_n}^{(\ell_n)}(\hat\mu_\theta,\hat\nu)$, we have that

$$\mathbb{E}\left[W_0(\mu_{\hat\theta_{\lambda_n}^{(\ell_n)}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim n^{\frac{-1}{d+2}}\log(n) \quad \textit{with} \quad \begin{cases} \lambda_n = n^{\frac{-1}{d+2}}, \\ \ell_n \geq 32R^4 n^{\frac{2}{d+2}}(\log(n))^{-1}. \end{cases}$$

Secondly, for $\hat\theta_{\lambda_n}^{S(\ell_n)} = \underset{\theta\in\Sigma_K}{\arg\min}\, S_{\lambda_n}^{(\ell_n)}(\hat\mu_\theta,\hat\nu)$, we have that

$$\mathbb{E}\left[W_0(\mu_{\hat\theta_{\lambda_n}^{S(\ell_n)}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim n^{\frac{-2}{d+4}}, \quad \textit{with} \quad \begin{cases} \lambda_n = n^{\frac{-1}{d+4}}, \\ \ell_n \geq 64R^4 n^{\frac{3}{d+4}}. \end{cases}$$

These results come from Corollary A.3 and Lemma A.6. Establishing point (i) requires some efforts, and it is thoroughly dealt with in Section A.1 and A.2. Then admitting this first result, we can derive point (ii) quite easily. The idea is to choose λ_n to balance the estimation and approximation error, and then to set the number of Sinkhorn iterations ℓ_n to maintain the algorithm error below this estimation versus approximation trade-off.

As we take into account the algorithmic error induced by Sinkhorn algorithm Theorem 3.4 is a computational version of Theorem 3.2. The analysis of the algorithmic error enables us to propose regularization choice λ, and a limited number of algorithm iterations to achieve the rates of convergence claimed in Theorem 3.2. In the same way, we give our second main result that consists in a computational version of Theorem 3.3. Exploiting results from Theorem 3.3, besides giving a regularization choice $(\lambda_n)_{n\geq 0}$, we also propose a limited number of Sinkhorn iterations to achieve the rates of convergence claimed in Theorem 3.3. Note that the difference with Theorem 3.4 comes from the fact that we have proposed a bound on the estimation error that is independent of λ. Hence, the following result ensures that the implementable estimators $\hat{\theta}_{\lambda_n}^{(\ell_n)}$ and $\hat{\theta}_{\lambda_n}^{S(\ell_n)}$ respectively reach the theoretical rates $n^{-2/d}\log(n)$ and $n^{-2/d}$.

Theorem 3.5. *Suppose that Assumptions A.1, A.2 and A.3 hold and that $d > 4$. If for all component μ_k as well as for ν, at least n observations are available, then the following non-asymptotic rates of convergence hold for the estimators $\hat{\theta}_{\lambda_n}^{(\ell_n)}$ and $\hat{\theta}_{\lambda_n}^{S(\ell_n)}$ that are computed with the Sinkhorn algorithm.*

(i)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_{\lambda_n}^{(\ell_n)}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim n^{-\frac{2}{d}}\log(n), \quad \textit{with} \quad \begin{cases} \lambda_n = n^{-2/d}, \\ \ell_n \geq 32R^4 n^{4/d}. \end{cases}$$

(ii)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_{\lambda_n}^{S(\ell_n)}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim n^{-\frac{2}{d}}, \quad \textit{with} \quad \begin{cases} \lambda_n = n^{-1/d}, \\ \ell_n \geq 64R^4 n^{3/d}. \end{cases}$$

Proof. This is the same proof as for Theorem 3.4, the only difference is to substitute the bound in the estimation error that is of order $\frac{1}{\lambda^{d/2}\sqrt{n}}$ in Theorem 3.4 with a bound of order $n^{-2/d}$. And then to chose the regularization parameter λ and the number of Sinkhorn iterations ℓ to bring the approximation error and the algorithm error bellow the estimation error. □

In Section A.6 we give bounds on the expected excess risks of $\hat{\theta}_{\lambda}^{(\ell)}$ and $\hat{\theta}_{\lambda}^{S(\ell)}$ that hold for any regularization parameter $\lambda > 0$ and any number of Sinkhorn iterations ℓ.

4 Numerical experiments

In this section, using simulated and real data from flow cytometry, we analyze the numerical performances of the various estimators introduced in Section 3. These numerical experiments have been designed to highlight that a relevant choice of the regularization parameter λ may lead to regularized Wasserstein estimators (based either on W_λ or S_λ) with performances similar to those of un-regularized Wasserstein estimators based on the standard OT cost W_0 at a lowest computational cost. For the results reported here, the parameter λ ranges in a finite grid $\Lambda \subset \mathbb{R}_+^*$ from 0.01 to 1.

For a given loss function $\mathcal{L} \in \{W_\lambda, S_\lambda, W_\lambda^{(\ell)}, S_\lambda^{(\ell)}, W_0\}$, we follow the protocol described thereafter. Using either simulated or real data, a single estimator $\hat{\theta}$ of the class proportions in the target dataset is obtained by solving the optimization problem:

$$\hat{\theta} = \underset{\theta\in\Sigma_K}{\arg\min}\, \mathcal{L}(\hat{\mu}_\theta, \hat{\nu}), \tag{4.1}$$

based on the empirical distributions $\hat{\mu}_1, \dots, \hat{\mu}_K$ and $\hat{\nu}$. In order to solve it, we apply a gradient descent algorithm to the function $F : \theta \mapsto \mathcal{L}(\hat{\mu}_\theta, \hat{\nu})$, where the computation of the gradient $\nabla_\theta \mathcal{L}(\hat{\mu}_\theta, \hat{\nu})$ essentially boils down to the resolution of the dual problem stated in Theorem 2.1 between the measures $\hat{\mu}_\theta$ and $\hat{\nu}$. We numerically solve this problem with the Sinkhorn algorithm in the regularized case, and with a linear programming algorithm in the un-regularized case. Repeating this protocol $N = 50$ times, we obtain N independent realizations $\hat{\theta}^{[1]}, \dots, \hat{\theta}^{[N]}$ of a given estimator of the class proportions. Then, we choose to compare the estimators using the expected quadratic risk $\mathbb{E}[\|\hat{\theta} - \theta^*\|^2]$ that is approximated by Monte-Carlo repetitions as classically done in statistical experiments:

$$\mathbb{E}[\|\hat{\theta} - \theta^*\|^2] \approx \frac{1}{N} \sum_{r=1}^{N} \|\hat{\theta}^{[r]} - \theta^*\|^2, \tag{4.2}$$

where θ^* is an optimal vector of class proportions depending on the data that are considered. This protocol is repeated for each value of λ in the grid Λ and each loss function.

Remark 4.1. *In these numerical experiments, we have chosen to focus on the expected error $\mathbb{E}[\|\hat{\theta} - \theta^*\|^2]$ rather than the expected excess risk $r_n(\mu_{\hat{\theta}}, \nu)$ as in flow cytometry the relevant quantity is an accurate estimation of class proportions in the target dataset. Also, notice that the risk $r_n(\mu_{\hat{\theta}}, \nu)$ cannot be computed exactly because it involves the quantity $W_0(\mu_{\hat{\theta}}, \nu)$ for which we have no closed-form formula.*

4.1 Simulated data

We first simulated two Gaussian mixtures of dimension $d = 6$ with the same $K = 5$ components but with different class proportions. Thus, a source data set corresponds to random vectors $X_1, ..., X_n$ sampled with respect to μ and a target data set corresponds to random vectors $Y_1, ..., Y_n$ sampled with respect to the distribution ν, where μ and ν are defined below:

$$\mu = \sum_{k=1}^{5} \pi_k \mathcal{N}(\rho_k, \sigma^2 I_d), \qquad \nu = \sum_{k=1}^{5} \theta_k^* \mathcal{N}(\rho_k, \sigma^2 I_d). \tag{4.3}$$

Because the vector of proportions π and θ^* are not assumed to be equal, we exploit the known classes at the source in order to estimate the class proportions θ^* at the target, based on empirical versions of $\mu_1, \dots, \mu_K$ and ν.

We have same number of samples $m_k = n$ from each source components μ_k than samples from the target distribution ν. This experimentation setting matches the presentation of our theoretical results given in Section 3.2.2. To ease the simulation study, we constrain the number of observations to $m_k = 50$ observations for each class of the source data set. In the target data set, we also constrain the number of observations per class with $n_1 = 20, n_2 = 5, n_3 = 8, n_4 = 7, n_5 = 10$, so $n = 50$ in total. We display in Figure 1 two-dimensional projections of one dataset from the source measure and one dataset from the target measure with their respective clustering. Note that the clustering of the target dataset is then assumed to be unknown.

4.1.1 Unlimited number of Sinkhorn iterations

Through a first series of experiments, we compare the performances of the estimators computed with the losses W_0, W_λ and S_λ. In Figure 2, using a boxplot we display the behavior of the error $\|\hat{\theta}_\lambda - \theta^*\|^2$ for each value of the regularization parameter $\lambda \in \Lambda$. In Figure 3, we also display the estimation of $\mathbb{E}[\|\hat{\theta}_\lambda - \theta\|^2]$ using the Monte-Carlo estimator (4.2). For small values of λ, the regularized losses W_λ and S_λ yield competitive estimators compared to the one obtained with W_0. These numerical findings thus validate our theoretical results from Section 3.2.2.

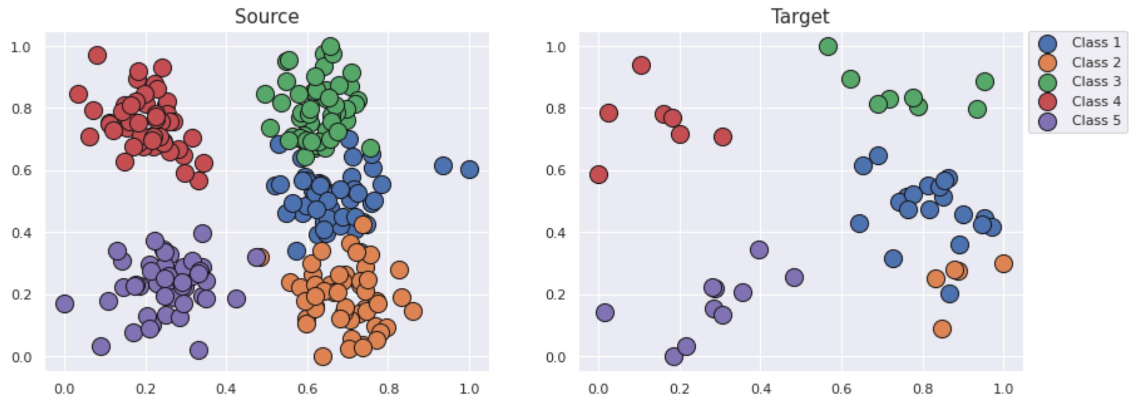

Figure 1: 2D projections of a simulated source data set and a target data set with their clustering.

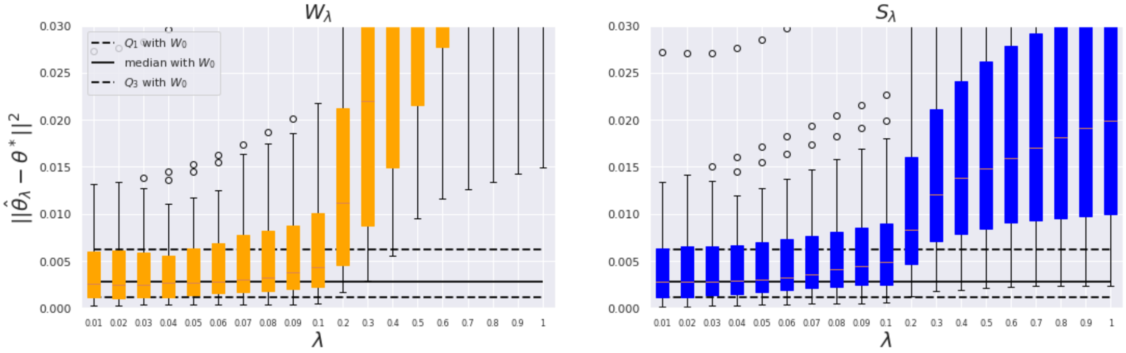

Figure 2: Estimation results on simulated data without limitation on the number of iterations of the Sinkhorn algorithm. We display the error $\|\hat{\theta}_\lambda - \theta^*\|^2$ using either the loss W_λ (left) or S_λ (right). The black line is the median error of the un-regularized estimator $\hat{\theta}_0$ using the standard optimal transport cost W_0, while the dotted lines are the first and third quartiles of the errors of estimation $\|\hat{\theta}_0 - \theta^*\|^2$.

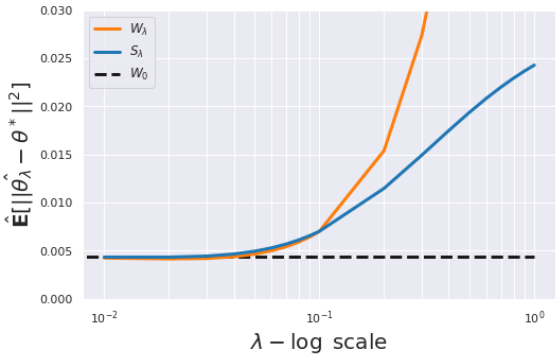

Figure 3: Average error on simulated data of the estimators $\hat{\theta}_\lambda$ (orange) and $\hat{\theta}^S_\lambda$ (blue) as a function of the regularization parameter λ. There is no limitation on the number of iterations, Sinkhorn algorithm runs until convergence is reached. The black dotted line is the average error of the un-regularized estimator $\hat{\theta}_0$.

We also point out that the computational complexity of the Sinkhorn algorithm is highly dependent on the regularization parameter λ as discussed in [Dvurechensky et al., 2018], [Altschuler et al., 2017]. To illustrate this fact, we display in Figure 4 the time (in seconds) required to compute $N = 50$ samples of $\hat{\theta}_\lambda$ depending on the value of λ. As $\nabla_\theta S(\mu_\theta, \nu) = \nabla_\theta W_\lambda(\mu_\theta, \nu) - \frac{1}{2}\nabla_\theta W_\lambda(\mu_\theta, \mu_\theta)$, computing the gradient of $S_\lambda(\mu_\theta, \nu)$, requires to solve the dual problem associated to $W_\lambda(\mu_\theta, \mu_\theta)$ in addition to the dual problem associated to $W_\lambda(\mu_\theta, \nu)$. But as noticed in Feydy et al. [2019] Sinkhorn algorithm converges much faster for the symmetric term $W_\lambda(a, a)$ than in the general case when computing $W_\lambda(a, b)$. We have observed in our experiment that the number of iterations before reaching convergence when computing $W_\lambda(a, a)$ does not seem to be a monotonic function with respect to the regularization parameter λ. This partially accounts for the slightly longer time of computation for $\lambda = 0.02$ in comparison to $\lambda = 0.01$ on the right side of Figure 4, that is when using S_λ as loss function.

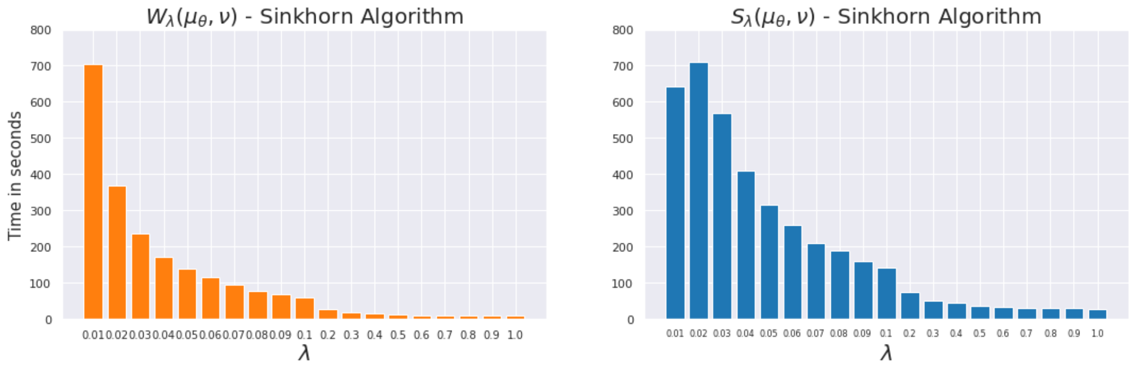

Figure 4: Time required to compute $N = 50$ estimators $\hat{\theta}_\lambda$ (left) and $\hat{\theta}^S_\lambda$ (right) when the number of iterations is unlimited.

4.1.2 Limited number of Sinkhorn iterations

Figure 2 and Figure 4 presents results questionning the trade-off between the computational cost of regularized OT and the quality of statistical estimation. We have repeated the experiments of Section 4.1.1 by now constraining the number iterations of the Sinkhorn algorithm to be equal to $\ell = 5$ for any value λ. In other words, we compute the estimators $\hat{\theta}^{(\ell)}_\lambda$ and $\hat{\theta}^{S(\ell)}_\lambda$ with $\ell = 5$, thus fxing the computational budget. Figure 5 and Figure 6 both present the performances of those estimators: by limiting the number of Sinkhorn iterations, the accuracy of the estimation deteriorates for small values of λ. This degradation comes from $\ell = 5$ being too small a number of iterations for the Sinkhorn algorithm to converge for small values of λ. Yet Figure 5 points to some values of Λ as offering a nice trade-off between the computational cost of small λ and the approximation error induced by larger λ. For such values, the performances of the regularized estimators $\hat{\theta}^{(\ell)}_\lambda$ and $\hat{\theta}^{S(\ell)}_\lambda$ are seen to be comparable to those of the un-regularized estimator $\hat{\theta}_0$.

4.2 Flow cytometry data

We now apply our method of class proportions estimation on flow cytometry data. We demonstrate that the regularization parameter λ has also a significant impact on the estimation of class proportions on real data. As an illustrative example, we apply our technique to flow cytometry data sets from the T-cell panel of the Human Immunology Project Consortium (HIPC) – publicly available on ImmuneSpace [Brusic et al., 2014]. We arbitrarily chose two data sets that comes from cytometry measurements performed in the “Stanford” laboratory center. One data set, that acts as the

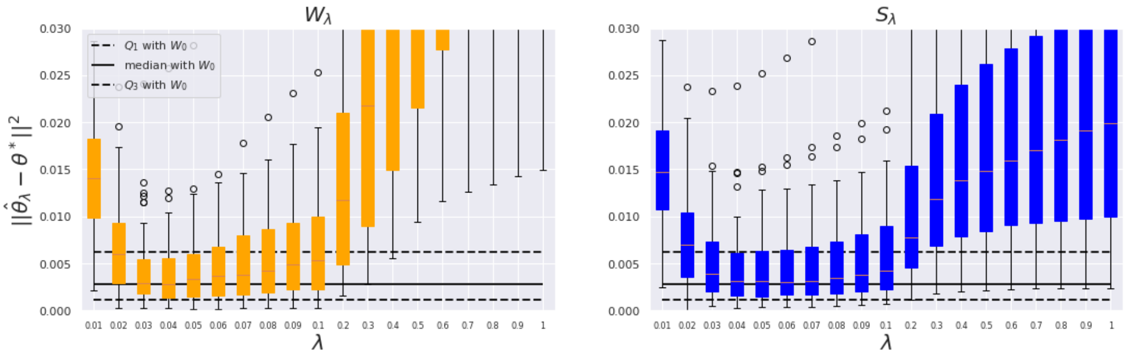

Figure 5: Estimation results on simulated data when the number of iterations of the Sinkhorn algorithm is limited to $\ell = 5$. We display the error $\|\hat{\theta}_\lambda^{(\ell)} - \theta^*\|^2$ using either the loss $W_\lambda^{(\ell)}$ (left) or $S_\lambda^{(\ell)}$ (right). The black line is the median error of the un-regularized estimator $\hat{\theta}_0$ using the standard optimal transport cost W_0, while the dotted lines are the first and third quartiles of $\|\hat{\theta}_0 - \theta^*\|^2$.

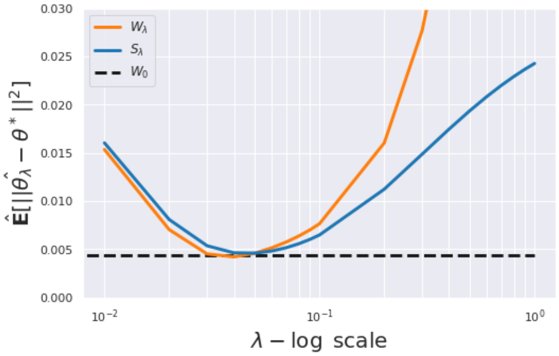

Figure 6: Average error on simulated data of the estimators $\hat{\theta}_\lambda^{(\ell)}$ (orange) and $\hat{\theta}_\lambda^{S(\ell)}$ (blue) as a function of the regularization parameter λ with a limitation on the number of iterations. For all values of λ, Sinkhorn algorithm is limited to $\ell = 5$ iterations. The black dotted line is the average error of the un-regularized estimator $\hat{\theta}_0$.

source measure, is built from observations measured from a biological sample of a certain patient. Another second data set, acting as the target measure, is built from the observations obtained from a biological sample that comes from another patient. After performing cytometry measurements the observations were manually gated into 10 cell populations: CD4 Effector (CD4 E), CD4 Naive (CD4 N), CD4 Central memory (CD4 CM), CD4 Effector memory (CD4 EM), CD4 Activated (CD4 A), CD8 Effector (CD8 E), CD8 Naive (CD8 N), CD8 Central memory (CD8 CM), CD8 Effector memory (CD8 EM) and CD8 Activated (CD8 A). Hence, for these data sets, a manual clustering is at our disposal to evaluate the performances of our method. In this context θ^* is defined as the class proportions defined thanks to the manual gating. For each cell, seven biological markers have been measured, and it thus leads to observations X_i and Y_j that belong to $\mathbb{R}^d$ with $d = 7$. A two-dimensional projection of these datasets is displayed in Figure 7 with the resulting manual clustering.

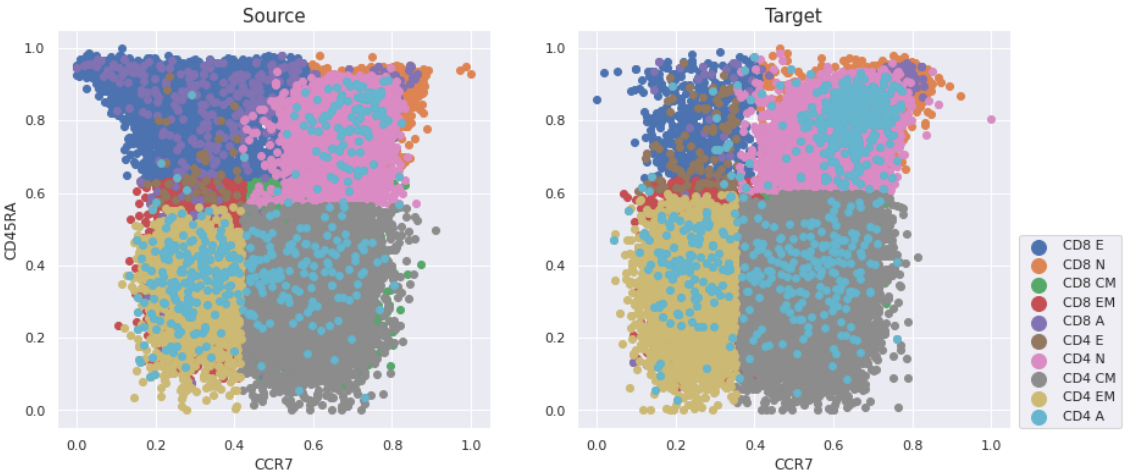

Figure 7: Two-dimensional projection of the flow cytometry datasets used in these numerical experiements with a clustering of the cells into 10 sub-populations.

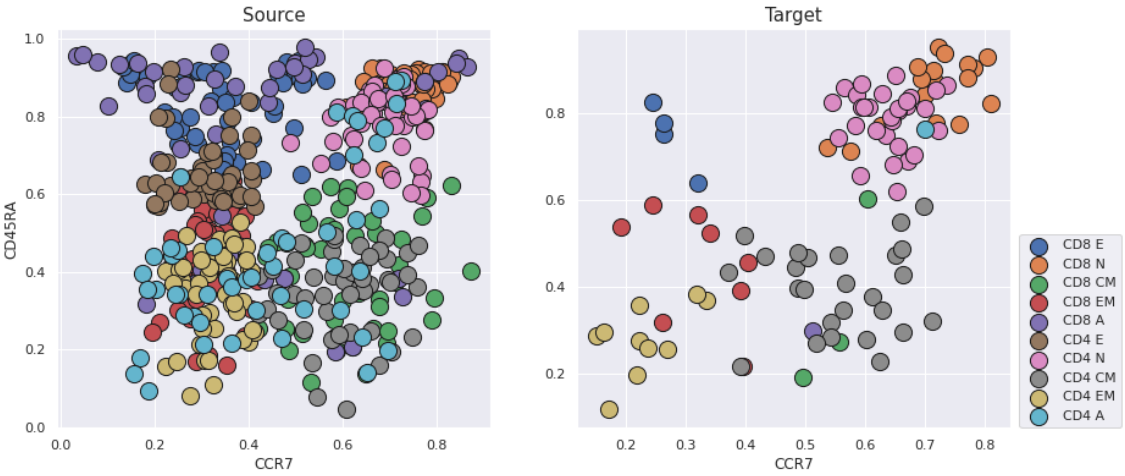

Figure 8: Sub-sample of the source and target flow cytometry data sets. In the source sub-sample, $m_k = m = 50$ elements of each class have been sampled. In the target sub-sample, $n = 100$ observations have been randomly chosen.

4.2.1 Unlimited Sinkhorn iterations

We reproduce the protocol that we have considered in the case of simulated data. To build an empirical distribution of the source distribution when analyzing flow cytometry data, we subsample 50 observations from each class of the source data set in order to construct the empirical measures $\hat{\mu}_1, \dots, \hat{\mu}_K$, and to define $\hat{\mu}_\theta = \sum_{k=1}^K \theta_k \hat{\mu}_k$ for $\theta \in \Sigma_K$. Figure 8 shows two sub-samples from the source and target distributions displayed in Figure 7. We recall again that the clustering of the target dataset is not used in the estimation procedure.

The numerical performances of the estimators computed with the three loss functions W_0, W_λ and S_λ are displayed on Figure 9 and Figure 10. In the context of flow cytometry data, the underlying distributions μ and ν are obviously unknown, and the quantity $\min W_0(\mu_\theta, \nu)$ is thus not accessible. Therefore, we define the optimal vector θ^* of class proportions to be the one in the fully observed (not sub-sampled) target dataset that is displayed in Figure 7. Those results on real data are consistent with the results of simulated data. Indeed, one can observe that for small values of $\lambda \in \Lambda$ the accuracy of the estimation obtained with the loss functions W_λ and S_λ is very similar to the one obtained using W_0.

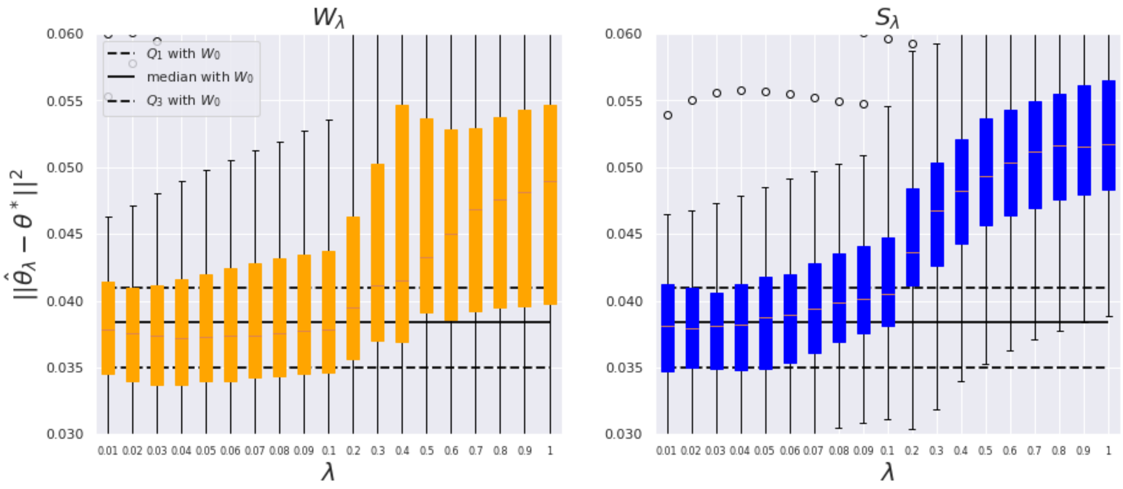

Figure 9: Results on HIPC data without imposing limitations on the number of Sinkhorn iterations. We display the error $\|\hat{\theta}_\lambda - \theta^*\|^2$ using either the loss W_λ (left) or S_λ (right). The black line is the median error of the un-regularized estimator $\hat{\theta}_0$ using the loss W_0, while the dotted lines are the first and third quartiles of $\|\hat{\theta}_0 - \theta^*\|^2$.

4.2.2 Limited Sinkhorn iterations

In order to reduce the computational cost of our estimation method, we limit the number of Sinkhorn iterations to $\ell = 10$. Once again, the results displayed in Figure 11 and Figure 12 show that it is possible to propose a competitive alternative to W_0 at a lower computational cost.

5 Conclusion and discussion

In this work, we have presented a thorough study of Wasserstein estimators based on regularized OT with an emphasis on the influence of the regularization parameter λ. This study was carried out through the example of a mixture model and weights estimation. We derived upper bounds

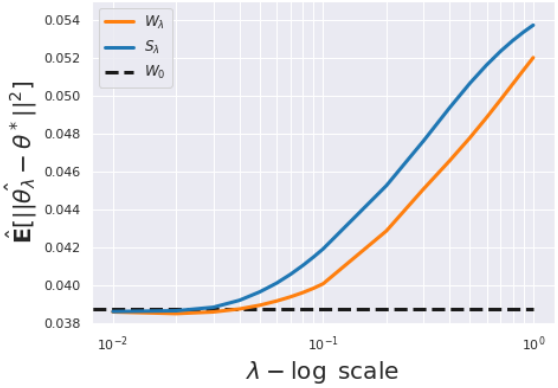

Figure 10: Average error on HIPC data of the estimators $\hat{\theta}_\lambda$ (orange) and $\hat{\theta}_\lambda^S$ (blue) as a function of the regularization parameter λ. There is no limitation on the number of iterations, Sinkhorn algorithm runs until convergence is reached. The black dotted line is the average error of the un-regularized estimator $\hat{\theta}_0$.

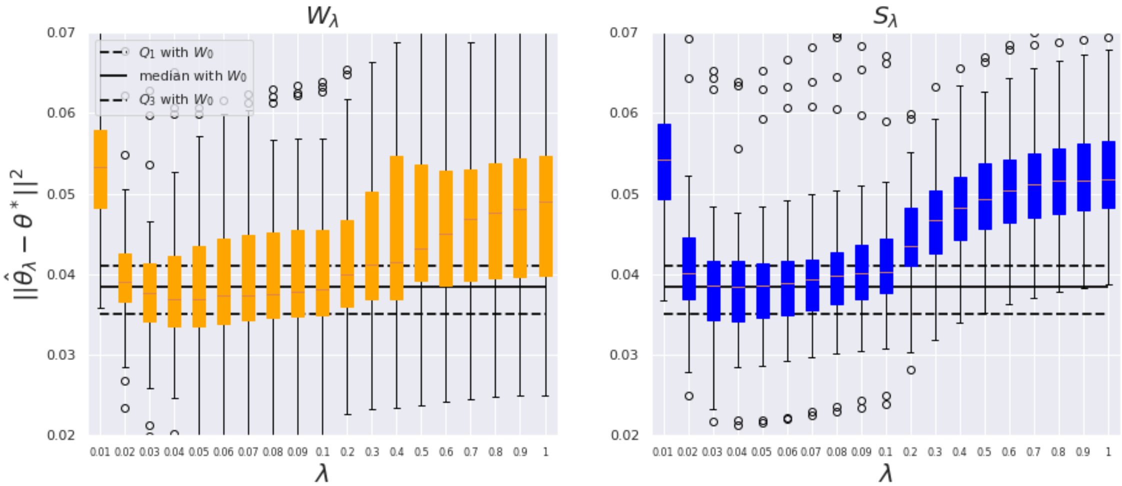

Figure 11: Results on HIPC data when the number of Sinkhorn iterations is limited to $\ell = 10$. We display boxplots of the error $\|\hat{\theta}_\lambda - \theta^*\|^2$ using either the loss W_λ (left) or S_λ (right). The black line is the median error of the un-regularized estimator $\hat{\theta}_0$ using the loss W_0, while the dotted lines are the first and third quartiles of $\|\hat{\theta}_0 - \theta^*\|^2$.

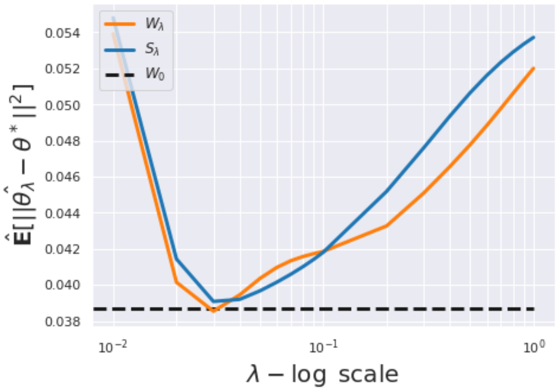

Figure 12: Average error on HIPC data of the estimators $\hat{\theta}_\lambda^{(\ell)}$ (orange) and $\hat{\theta}_\lambda^{S(\ell)}$ (blue) as a function of the regularization parameter λ with a limitation on the number of iterations. For all values of λ, Sinkhorn algorithm is limited to $\ell = 10$ iterations. The black dotted line is the average error of the un-regularized estimator $\hat{\theta}_0$

on the risk of Wasserstein estimators in terms of an estimation error and an approximation error. We assessed the influence of the chosen OT-based loss (among W_λ, S_λ and W_0) on the decay of the estimation and approximation terms. We have also proposed an optimal decay of the regularization parameter $\lambda = \lambda_n$ based on these upper bounds to achieve decreasing rate of $n^{-2/d}$ for the expected excess risk. Secondly, motivated by the sensitive question of the computational cost of regularized OT, we have studied the algorithmic error induced by limiting the number of iterations in the Sinkhorn algorithm. This study resulted in a principled strategy to set the number of Sinkhorn iterations $\ell = \ell_n$ in order to maintain the algorithm error below the statistical error. We have also demonstrated with numerical experiments that an appropriate choice of λ and a limited number of Sinkhorn iterations ℓ allow to equal the performances of the un-regularized estimator at a reduced computational cost.

Based on the results of Manole and Niles-Weed [2021], we believe the rate $n^{-2/d}$ to be near minimax. To the best of our knowledge such a rate of convergence was not established yet for regularized estimators.

We now present a few perspectives for future research. For an estimator $\hat{\theta}_n$ of θ^*, we have derived a control on the excess risk, that is $W_0(\mu_{\hat{\theta}_n}, \nu) - W_0(\mu_{\theta^*}, \nu)$. However, a direct control of the weights estimator, i.e. of the quantity $\|\hat{\theta}_n - \theta^*\|$, would be even more valuable. For instance, a control of $\|\hat{\theta}_n - \theta^*\|$ may allow to develop statistical tests on the estimator $\hat{\theta}_n$. An other possible extension of this work is suggested by our numerical experiments. Figure 2 and Figure 5 indicate that limiting the number of iterations for Sinkhorn algorithm could improve statistical performance. These better results with limited iterations are not accounted by the present work. Hence, further investigation on this observation is an other direction for research.

References

N. Aghaeepour, G. Finak, H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, R.H. Scheuermann, FlowCAP Consortium, Dream Consortium, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228, 2013.

J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

M. Ballu, Q. Berthet, and F. Bach. Stochastic optimization for regularized wasserstein estimators. In *International Conference on Machine Learning*, pages 602–612. PMLR, 2020.

B. Bercu and J. Bigot. Asymptotic distribution and convergence rates of stochastic algorithms for entropic optimal transportation between probability measures. *The Annals of Statistics*, 49(2): 968 – 987, 2021.

E. Bernton, P. E. Jacob, M. Gerber, and C.P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

J. Bigot. Statistical data analysis in the Wasserstein space. *ESAIM: ProcS*, 68:1–19, 2020.

J. Bigot, E. Cazelles, and N. Papadakis. Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration. *Information and Inference: A Journal of the IMA*, 8, 04 2018.

V. Brusic, R. Gottardo, S.H. Kleinstein, M.M. Davis, D.A. Hafler, H. Quill, A.K. Palucka, G.A. Poland, B. Pulendran, E.L. Reinherz, et al. Computational resources for high-dimensional immune analysis from the human immunology project consortium. *Nature biotechnology*, 32(2): 146, 2014.

L. Chizat, P. Roussillon, F. Léger, F.X. Vialard, and G. Peyré. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. In *Proc. NeurIPS'20*, 2020.

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

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn algorithm. In *International conference on machine learning*, pages 1367–1376. PMLR, 2018.

J. Feydy, T. Séjourné, F.X. Vialard, S. Amari, A. Trouve, and G. Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

P. Freulon, J. Bigot, and B. P. Hejblum. Cytopt: Optimal transport with domain adaptation for interpreting flow cytometry data. *arXiv preprint arXiv:2006.09003*, 2020.

G. Biau and M. Sangnier and U. Tanielian. Some Theoretical Insights into Wasserstein GANs. *Journal of Machine Learning Research*, 22(119):1–45, 2021. URL http://jmlr.org/papers/v22/20-553.html.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in neural information processing systems*, pages 3440–3448, 2016.

A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR, 2018.

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

G.J Huizing, G. Peyré, and L. Cantini. Optimal transport improves cell-cell similarity inference in single-cell omics data. *bioRxiv*, 2021.

D. Liu, M. T. Vu, S. Chatterjee, and L. K. Rasmussen. Entropy-regularized optimal transport generative models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3532–3536. IEEE, 2019.

T. Manole and J Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *arXiv preprint arXiv:2106.13181*, 2021.

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

V. M. Panaretos and Y. Zemel. Statistical Aspects of Wasserstein Distances. *Annual Reviews of Statistics and its Applications*, 6:405–431, 2018.

A. Petersen, C. Zhang, and P. Kokoszka. Modeling Probability Density Functions as Data Objects. *Econometrics and Statistics*, 21(C):159–178, 2022.

M. Sanjabi, J. Ba, M. Razaviyayn, and J. D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pages 7091–7101, 2018.

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

A. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics.* New York: Springer, 1996.

R. van Handel. *Probability in High Dimension.* Princeton University, 2016.

C. Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

A Proofs of the main results

In this appendix, we proceed to the proofs of the main results of the paper.

A.1 Proof of Theorem 3.2 - Part (i)

The goal of this section is to derive the rate of convergence of regularized Wasserstein estimators, that is, when considering W_λ as a loss function with $\lambda > 0$. We therefore investigate the behavior of the estimator

$$\hat{\theta}_\lambda \in \widehat{\Theta}_\lambda := \underset{\theta \in \Sigma_K}{\arg\min}\, W_\lambda(\hat{\mu}_\theta, \hat{\nu}), \tag{A.1}$$

where W_λ is defined in equation (2.1). In what follows, we derive an upper bound on the expected excess risk of $\hat{\theta}_\lambda$ for a fixed value of λ. This result finally yields the convergence rate claimed in part (i) of Theorem 3.2 for the choice $\lambda = \lambda_n = n^{-\frac{1}{2\lfloor d/2 \rfloor + 2}}$ as discussed in Section 3.2.2. The results derived in this section hold under Assumption (A.1), i.e., that the distributions μ and ν have compact supports included in $B(0, R)$.

A.1.1 Decomposition of the excess risk

We first detail how the excess risk of $\hat{\theta}_\lambda$ can be upper bounded by the sum of two terms representing a tradeoff between an estimation error and an approximation error. Thanks to [Genevay et al., 2019, Theorem 1] adapted to the squared Euclidean cost $c(x,y) = \|x-y\|^2$ (which is R-Lipschitz on $B(0,R)$ w.r.t. both its variables), we can control the impact of entropic regularization on the approximation of the value of the un-regularized OT cost is as follows.

Proposition A.1. *Assume that $\mathcal{X}, \mathcal{Y}$ are compact subsets of $B(0,R)$. Then, it holds that*

$$0 \le W_\lambda(\mu,\nu) - W_0(\mu,\nu) \le 2d\lambda \log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right), \tag{A.2}$$

and consequently

$$\sup_{\theta\in\Sigma_K} |W_0(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\nu)| \leqslant B(\lambda) \quad \textit{where} \quad B(\lambda) = 2d\lambda\log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right) \tag{A.3}$$

Notice that $B(\lambda)$ goes to zero when $\lambda \to 0$ at the speed

$$B(\lambda) \sim_{\lambda\to 0} 2d\lambda\log\left(1/\lambda\right).$$

Next, a key result is the following decomposition of (an upper bound) of the excess risk into the sum of an estimation error and an approximation error .

Lemma A.1. *For $\lambda \ge 0$, the excess risk of the estimator $\hat{\theta}_\lambda$ defined by (A.1) is bounded as follows:*

$$0 \le W_0(\mu_{\hat{\theta}_\lambda},\nu) - W_0(\mu_{\theta^*},\nu) \le \underbrace{2\sup_{\theta\in\Sigma_K}|W_\lambda(\hat{\mu}_\theta,\hat{\nu}) - W_\lambda(\mu_\theta,\nu)|}_{\textit{Estimation error}} + \underbrace{2B(\lambda)}_{\textit{Approximation error}}, \tag{A.4}$$

with $B(\lambda)$ defined in (A.3) (and by convention $B(0) = 0$).

The proof of Lemma A.1 is deferred to Section B.1 of the Appendix. To control the estimation error of equation (A.4) we write

$$\sup_{\theta\in\Sigma_K}|W_\lambda(\mu_\theta,\nu) - W_\lambda(\hat{\mu}_\theta,\hat{\nu})| \le \sup_{\theta\in\Sigma_K}|W_\lambda(\mu_\theta,\hat{\nu}) - W_\lambda(\mu_\theta,\nu)| + \sup_{\theta\in\Sigma_K}|W_\lambda(\hat{\mu}_\theta,\hat{\nu}) - W_\lambda(\mu_\theta,\hat{\nu})|. \tag{A.5}$$

Hence, controlling the expected excess risk boils down to controlling the (closely related) expected empirical processes

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K}\left|W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})\right|\right], \quad \text{and} \quad \mathbb{E}\left[\sup_{\theta\in\Sigma_K}\left|W_\lambda(\hat{\mu}_\theta,\hat{\nu}) - W_\lambda(\mu_\theta,\hat{\nu})\right|\right]. \tag{A.6}$$

We will begin by controlling the first one, see Proposition A.2 below. The control for the second one is very similar and given in Proposition A.5 whose proof is deferred to Section B.2. Using the dual formulation of the regularized OT, we show that controlling the above expected empirical processes essentially leads to find an upper bound for

$$\mathbb{E}\left[\sup_{\psi\in\mathcal{F}}\left|\int\psi d\nu - \int\psi d\hat{\nu}\right|\right] = \mathbb{E}\left[\sup_{\psi\in\mathcal{F}}\left|\mathbb{E}[\psi(Y)] - \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i)\right|\right] \tag{A.7}$$

for a suitable class $\mathcal{F}$ of smooth functions ψ.

A.1.2 Control of empirical processes for the loss function W_λ

This section aims at controlling the empirical process

$$\sup_{\theta\in\Sigma_K}\left|W_\lambda(\mu_\theta,\nu)-W_\lambda(\mu_\theta,\hat{\nu})\right|. \tag{A.8}$$

The arguments that we use are very much inspired by the works Genevay et al. [2019], Chizat et al. [2020]. For a fixed value of $\lambda>0$, we are going to obtain an upper bound decaying at the rate $\frac{1}{\sqrt{n}}$ where n is the number of samples in the empirical measure $\hat{\nu}$. We will also see how the constants involved in this upper bound depend on the regularizing parameter λ with a power that depends on the dimension d of the data. For a bounded subset $\mathcal{Z}$ of $\mathbb{R}^d$, we shall denote by $\mathscr{C}^{\mathscr{K}}(\mathcal{Z})$ the set of $\mathscr{C}^{\mathscr{K}}$ functions on $\mathcal{Z}$ equipped with the norm $\|f\|_{\mathscr{K}}=\max_{|\kappa|\leqslant\mathscr{K}}\|\partial^\kappa f\|_\infty$ and

$$\mathscr{C}^{\mathscr{K}}_M(\mathcal{Z})=\{f\in\mathscr{C}^{\mathscr{K}}(\mathcal{Z})\mid\|f\|_{\mathscr{K}}\leqslant M\}, \tag{A.9}$$

where the notation $|\kappa|\leqslant\mathscr{K}$ denotes any multi-index κ of differentiation of length $|\kappa|$ at most $\mathscr{K}$.

Proposition A.2. *Suppose that Assumption A.1 holds true. Then, we can bound the expectation of the empirical process (A.8) as follows*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K}\left|W_\lambda(\mu_\theta,\nu)-W_\lambda(\mu_\theta,\hat{\nu})\right|\right]\lesssim\frac{M_\lambda}{\sqrt{n}}, \tag{A.10}$$

with

$$M_\lambda=M_d\max\left(R^2,\frac{R^{\lfloor d/2\rfloor+1}}{\lambda^{\lfloor d/2\rfloor}}\right), \tag{A.11}$$

and M_d is a constant that depends only on d.

This proof is built upon the following lemmas. A first step is to study the regularity of the optimal potentials of the dual formulation (2.2). To this end, we adapt the analysis of Genevay et al. [2019] to the setting where the measure μ belongs to the parametric model $\{\mu_\theta\mid\theta\in\Sigma_K\}$. This result is adapted from Genevay et al. [2019] and relies on the fact that an optimal potential can be chosen as the c-transform of $\varphi\in L^\infty(\mathcal{X})$. The choice of the squared Euclidean cost allows for a very clear description of the regularity of an optimal potential. We recall below the definition of the c-transform of $\varphi\in L^\infty(\mathcal{X})$ that we denote by $\varphi^{c,\lambda}_\mu$ as $\lambda>0$. With the aim of manipulating functions defined on a convex and bounded subset of $\mathbb{R}^d$, the c-transforms are defined on $B(0,R)$. Hence, even if integrated only against μ or ν that have support $\mathcal{X}$ and $\mathcal{Y}$ respectively, the c-transform are defined on $B(0,R)$. The expression of $\varphi^{c,\lambda}_\mu\in L^\infty(B(0,R))$ is given by

$$\forall y\in B(0,R),\ \varphi^{c,\lambda}_\mu(y)=-\lambda\log\int e^{-\frac{\|x-y\|^2-\varphi(x)}{\lambda}}d\mu(x). \tag{A.12}$$

Lemma A.2. *Suppose that Assumption A.1 holds true. Then, for all $\theta\in\Sigma_K$, there exists a couple of dual potentials (φ,ψ) with respect to $W_\lambda(\mu_\theta,\nu)$, that satisfies $\psi(0)=0$, and $\psi=\varphi^{c,\lambda}_{\mu_\theta}$. Moreover, ψ belongs to $\mathscr{C}^\infty(B(0,R))$, and for each $\mathscr{K}>0$, there exists a constant $M_{\mathscr{K}}>0$ that depends only on $\mathscr{K}$ such that*

$$\|\psi\|_{\mathscr{K}}\leqslant M_{\mathscr{K}}\max\left(R^2,\frac{R^{\mathscr{K}}}{\lambda^{\mathscr{K}-1}}\right). \tag{A.13}$$

The sup norm is taken over $B(0,R)$.

This lemma will be proved in Section E. Two observations on this Lemma A.2 will reveal useful in the sequel.

Remark A.1. *We stress that $\psi=\varphi^{c,\lambda}_{\mu_\theta}$ is a regularized c-transform with respect to μ_θ (with $\theta\in\Sigma_K$), and that the constant $M_{\mathscr{K}}>0$ that appears in Lemma A.2 does not depend on θ.*

Remark A.2. *If we denote by φ the dual potential of Lemma A.2 such $\psi = \varphi_{\mu_\theta}^{c,\lambda}$ where ψ meets the requirements of the Lemma A.2, this variable φ can be chosen as an optimal potential for the semi-dual formulation (2.5) of the regularized optimal transport problem.*

In the sequel of this section, we make a repeating use of the constant of Lemma A.2. Thus, we introduce the notation $M_{\lambda,\mathscr{K}}$ to refer to the upper bound of equation (A.13) which is defined as

$$M_{\lambda,\mathscr{K}} := M_{\mathscr{K}} \max\left(R^2, \frac{R^{\mathscr{K}}}{\lambda^{\mathscr{K}-1}}\right), \tag{A.14}$$

and

$$M_{\mathscr{K}} > 0 \text{ is a constant that depends only on } \mathscr{K}. \tag{A.15}$$

The next proposition links the estimation error $\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu,\hat{\nu})|$ to the regularity of the c-transforms established in Lemma A.2. This result is established using the semi-dual formulation of $W_\lambda(\mu,\nu)$ that one can find in equation (2.5) in the introduction of the present paper.

Proposition A.3. *Suppose that Assumption A.1 holds. Then, for $\mathscr{K} > 0$, we have the following upper bound*

$$\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| \leqslant \sup_{\psi\in\mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))} \int \psi d(\nu - \hat{\nu}), \tag{A.16}$$

with $M_{\lambda,\mathscr{K}} > 0$ a constant defined in equation (A.14).

Proof. Let $\theta \in \Sigma_K$, $\mathscr{K} > 0$ and introduce φ, ψ (resp. $\hat{\varphi}, \hat{\psi}$) two optimal potentials for the dual formulation of $W_\lambda(\mu_\theta,\nu)$ (resp. $W_\lambda(\mu_\theta,\hat{\nu})$) chosen as in Lemma A.2. In particular $\psi, \hat{\psi} \in \mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))$, and the potentials φ and $\hat{\varphi}$ are respectively optimal potentials for the semi-dual formulation of $W_\lambda(\mu_\theta,\nu)$ and $W_\lambda(\mu_\theta,\hat{\nu})$ as precised in Remark A.2. We can thus write

$$\begin{aligned} W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu}) &= \int \varphi d\mu_\theta + \int \psi d\nu - \int \hat{\varphi} d\mu_\theta - \int \hat{\psi} d\hat{\nu} \\ &= \int \psi d\nu - \int \psi d\hat{\nu} \\ &+ \underbrace{\left(\int \varphi d\mu_\theta + \int \psi d\hat{\nu} - \int \hat{\varphi} d\mu_\theta - \int \hat{\psi} d\hat{\nu}\right)}_{\leqslant 0}. \end{aligned}$$

By optimality of $\hat{\varphi}$ for the semi dual formulation of $W_\lambda(\mu_\theta,\hat{\nu})$, the last term in the above parenthesis is non-positive. Using a symmetric optimality argument for $W_\lambda(\mu_\theta,\nu)$ and its optimal potential φ for the semi-dual formulation, we get

$$\int \hat{\psi} d(\nu - \hat{\nu}) \leqslant W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu}) \leqslant \int \psi d(\nu - \hat{\nu}). \tag{A.17}$$

As ψ and $\hat{\psi}$ belong to $\mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))$, we can write

$$|W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| \leqslant \sup_{\psi\in\mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))} \left|\int \psi d(\nu - \hat{\nu})\right| = \sup_{\psi\in\mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))} \int \psi d(\nu - \hat{\nu}). \tag{A.18}$$

The set $\mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))$ being independent of θ, we get

$$\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| \leqslant \sup_{\psi\in\mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))} \int \psi d(\nu - \hat{\nu}). \tag{A.19}$$

□

Therefore, the search for a control over $\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta, \nu) - W_\lambda(\mu, \hat{\nu})|$ leads us to the study of the following empirical process

$$\sup_{\psi\in\mathscr{C}^{\mathscr{K}}_{M_\lambda,\mathscr{K}}(B(0,R))} Y_\psi \quad \text{with} \quad Y_\psi = \int \psi d(\hat{\nu} - \nu) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i) - \int \psi d\nu. \tag{A.20}$$

In order to bound the empirical process (A.20), we will need several ingredients. First, we show that this empirical process has a sub-Gaussian behavior (see the definition in van Handel [2016]).

Lemma A.3. *Under Assumption A.1, if $Y = \{Y_1, ..., Y_n\}$ where $Y_1, ..., Y_n$ are independent random samples from ν, the empirical process $(Y_\psi)_{\psi\in\mathscr{C}^{\mathscr{K}}_{M_\lambda,\mathscr{K}}(B(0,R))}$ defined in equation (A.20) has zero mean and is subgaussian w.r.t. $2n^{-\frac{1}{2}}\|\cdot\|_\infty$. In other terms,*

$$\forall\varphi,\psi \in \mathscr{C}^{\mathscr{K}}_{M_\lambda,\mathscr{K}}(B(0,R)), \quad \forall s\in\mathbb{R}, \quad \mathbb{E}[e^{s(Y_\varphi - Y_\psi)}] \leqslant e^{\frac{2s^2}{n}\|\varphi-\psi\|^2_\infty} = e^{\frac{s^2}{2}(2n^{-\frac{1}{2}}\|\varphi-\psi\|_\infty)^2}. \tag{A.21}$$

Proof. Let us set $\varphi,\psi \in \mathscr{C}^{\mathscr{K}}_{M_\lambda,\mathscr{K}}(B(0,R))$ and consider the increment $Y_\varphi - Y_\psi$ defined by

$$Y_\varphi - Y_\psi = \frac{1}{n}\sum_{i=1}^{n} \varphi(Y_i) - \psi(Y_i) - \int_{\mathcal{Y}} \varphi(y) - \psi(y) d\nu(y). \tag{A.22}$$

We are going to show that $Y_\varphi - Y_\psi$ subgaussian. Denote by Δ_i the variable defined by

$$\Delta_i = \frac{1}{n}\left(\varphi(Y_i) - \psi(Y_i) - \int_{\mathcal{Y}} \varphi(y) - \psi(y) d\nu(y)\right) \tag{A.23}$$

Denote by $(\mathscr{F}_i)_{1\leq i\leq n}$ the filtration defined by $\mathscr{F}_i = \sigma(Y_1, \ldots, Y_i)$. The two following facts hold true. First, $\mathbb{E}\left[\Delta_i|\mathscr{F}_{i-1}\right] = 0$. Second,

$$-2\frac{\|\varphi-\psi\|_\infty}{n} \leq \Delta_i \leq 2\frac{\|\varphi-\psi\|_\infty}{n}. \tag{A.24}$$

We can thus apply Azuma-Hoeffding inequality van Handel [2016][Corollary 3.9] to derive that $Y_\varphi - Y_\psi$ is subgaussian with variance proxy

$$\frac{1}{4}\sum_{i=1}^{n}\left(\frac{4\|\varphi-\psi\|_\infty}{n}\right)^2 = \frac{4\|\varphi-\psi\|^2_\infty}{n}. \tag{A.25}$$

We thus have

$$\forall s\in\mathbb{R}, \quad \mathbb{E}[e^{s(Y_\varphi - Y_\psi)}] \leqslant e^{\frac{s^2 4}{2n}\|\varphi-\psi\|^2_\infty} = e^{\frac{s^2}{2}(2n^{-\frac{1}{2}}\|\varphi-\psi\|_\infty)^2},$$

as claimed in Lemma A.3. □

We can now use Dudley's entropy integral inequality, which we recall now.

Theorem A.1. *(Dudley's entropy integral inequality) Let $(Y_\varphi)_{\varphi\in\Phi}$ be a zero mean stochastic process which is sub-Gaussian with respect to the distance induced by a norm $\|\cdot\|$ on the indexing set Φ. Then*

$$\mathbb{E}\left[\sup_{\varphi\in\Phi} Y_\varphi\right] \leq 12\int_0^\infty \sqrt{\log(N(\varepsilon,\Phi,\|\cdot\|))}d\varepsilon,$$

where $N(\varepsilon,\Phi,\|\cdot\|)$ is the covering number of Φ by balls of radius ε with respect to the norm $\|\cdot\|$.

A classical bound on the covering number for smooth functions (see e.g. [van der Vaart and Wellner, 1996, Theorem 2.7.1]) will prove highly valuable.

Theorem A.2. *If $\mathcal{Z}$ is a bounded convex subset of $\mathbb{R}^d$ with nonempty interior, then there exists a constant $L(\mathscr{K}, d)$ such that*

$$\forall \varepsilon > 0, \quad \log N(\varepsilon, \mathscr{C}_1^{\mathscr{K}}(\mathcal{Z}), \|\cdot\|_\infty) \leqslant L(\mathscr{K}, d)|\mathcal{Z} + B(0,1)|\frac{1}{\varepsilon^{d/\mathscr{K}}}. \tag{A.26}$$

where $N(\varepsilon, \mathscr{C}_1^{\mathscr{K}}(\mathcal{Z}), \|\cdot\|_\infty)$ denotes the covering number of $\mathscr{C}_1^{\mathscr{K}}(\mathcal{Z})$ (by balls of radius ε) with respect to the ℓ_∞ norm, and where $|\mathcal{Z} + B(0,1)|$ is the Lebesgue measure of $\mathcal{Z} + B(0,1)$

We now have all the ingredients to bound the expectation of the empirical process (A.20).

Proposition A.4. *Suppose that Assumption A.1 holds. Then, we have the following upper bound*

$$\mathbb{E}\left[\sup_{\psi \in \mathscr{C}^{d'}_{M_\lambda}(B(0,R))} \int \psi d(\hat{\nu} - \nu)\right] \lesssim \frac{M_\lambda}{\sqrt{n}} \quad \textit{with} \quad M_\lambda = M_d \max\left(R^2, \frac{R^{\lfloor d/2 \rfloor + 1}}{\lambda^{\lfloor d/2 \rfloor}}\right), \tag{A.27}$$

and $M_d > 0$ is a constant that depends only on d.

Proof. Set $\mathscr{K} > 0$. We denote the empirical process under study by

$$(Y_\psi)_{\psi \in \mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))} \quad \text{with} \quad Y_\psi = \int \psi d(\hat{\nu} - \nu) = \frac{1}{n}\sum_{i=1}^n \psi(Y_i) - \int \psi d\nu,$$

and $M_{\lambda,\mathscr{K}} > 0$ the constant defined in equation (A.14). Dudley's inequality with the entropy integral (see e.g. van Handel [2016]) gives

$$\mathbb{E}\left[\sup_{\psi \in \mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R))} Y_\psi\right] \leqslant 12 \int_0^\infty \sqrt{\log N(\varepsilon, \mathscr{C}^{\mathscr{K}}_{M_{\lambda,\mathscr{K}}}(B(0,R)), 2n^{-\frac{1}{2}}\|\cdot\|_\infty)} d\varepsilon \tag{A.28}$$

$$\leqslant 12 \int_0^\infty \sqrt{\log N\left(\frac{1}{2}\sqrt{n} M^{-1}_{\lambda,\mathscr{K}}\varepsilon, \mathscr{C}_1^{\mathscr{K}}(B(0,R)), \|\cdot\|_\infty\right)} d\varepsilon \tag{A.29}$$

$$\leqslant \frac{24 M_{\lambda,\mathscr{K}}}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\varepsilon, \mathscr{C}_1^{\mathscr{K}}(B(0,R)), \|\cdot\|_\infty)} d\varepsilon \tag{A.30}$$

$$\lesssim \frac{M_{\lambda,\mathscr{K}}}{\sqrt{n}} \int_0^{1/2} \varepsilon^{-\frac{d}{2\mathscr{K}}} d\varepsilon. \tag{A.31}$$

This integral is finite as soon as $\mathscr{K} > d/2$. As $M_{\lambda,\mathscr{K}} = M_{\mathscr{K}} \max\left(R^2, \frac{R^{\mathscr{K}}}{\lambda^{\mathscr{K}-1}}\right)$ and λ will be chosen little in the sequel, we set $\mathscr{K} = \lfloor d/2 \rfloor + 1$ in order to have the quantity $M_{\lambda,\mathscr{K}}$ as small as possible. From now on, we denote by $d' := \lfloor d/2 \rfloor + 1$, and with this choice of $\mathscr{K} = d'$, we can substitute the constant $M_{\lambda,\mathscr{K}}$ of Lemma A.2 with a new constant $M_\lambda := M_{\lambda,d'}$ that reads

$$M_\lambda = M_d \max\left(R^2, \frac{R^{\lfloor d/2 \rfloor + 1}}{\lambda^{\lfloor d/2 \rfloor}}\right),$$

where M_d is a constant that depends only on d. Finally, we have the following bound for the empirical process under study

$$\mathbb{E}\left[\sup_{\psi \in \mathscr{C}^{d'}_{M_\lambda}(B(0,R))} Y_\psi\right] \lesssim \frac{M_\lambda}{\sqrt{n}}. \tag{A.32}$$

□

Proof of Proposition A.2 Gathering the results established since Lemma A.2, we are in a favorable position to prove Proposition A.2.

Proof. Indeed, by combining inequality (A.27) from Proposition A.4 with upper bound (A.16), we derive

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K}\left|W_\lambda(\mu_\theta,\nu)-W_\lambda(\mu_\theta,\hat{\nu})\right|\right]\lesssim\frac{M_\lambda}{\sqrt{n}}. \tag{A.33}$$

And the above inequality is the result claimed in Proposition A.2. □

The second empirical process in (A.6) can be upper bounded in a similar manner, as shown in the next proposition.

Proposition A.5. *Suppose that Assumption A.1 holds. Make the additional assumption that for all $k\in\{1,...,K\}$, m_k samples from μ_k are available, and denote by $\underline{m}=\min(m_1,\dots,m_K)$. Then, the following inequality holds*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K}|W_\lambda(\hat{\mu}_\theta,\hat{\nu})-W_\lambda(\mu_\theta,\hat{\nu})|\right]\lesssim\frac{M_\lambda}{\sqrt{\underline{m}}}, \tag{A.34}$$

where M_λ is defined in equation (A.11).

For the proof of Proposition A.5 we refer to Section B.2 of the Appendix.

A.1.3 Expected excess risk of regularized Wasserstein estimators

We now gather the results from the previous section to obtain an upper bound on the expected excess risk of our regularized Wasserstein estimators.

Proposition A.6. *Suppose that Assumption A.1 holds true. If, for all $k\in\{1,...,K\}$, m_k samples from μ_k are available and n samples from ν are available, denoting $\underline{m}=\min(m_1,\dots,m_K)$, we have*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda},\nu)-W_0(\mu_{\theta^*},\nu)\right]\lesssim\underbrace{\frac{2M_\lambda}{\sqrt{\underline{m}}}+\frac{2M_\lambda}{\sqrt{n}}}_{\textit{Estimation error}}\underbrace{+4d\lambda\log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right)}_{\textit{Approximation error}}, \tag{A.35}$$

where $M_\lambda=M_d\max\left(R^2,\frac{R^{\lfloor d/2\rfloor+1}}{\lambda^{\lfloor d/2\rfloor}}\right)$, and M_d is a constant that depends only on d.

Proof. Gathering the results on the approximation error for the regularized OT cost in Proposition A.1 and the upper bounds from Proposition A.2 and Proposition A.5 on the empirical processes defined in (A.6), we obtain the convergence rate claimed in (A.35). □

From the upper bound on the expected excess risk of $\hat{\theta}_\lambda$ established in the previous Proposition A.6, we can propose a regularization policy in order to balance the estimation error and the approximation error. This regularization policy and the corresponding rate of convergence are given in the next Corollary.

Corollary A.1. *Suppose that Assumption A.1 holds true. Make the additional assumption that for all the distributions μ_k and for ν, at least n samples are available. Then, choosing $\lambda_n=n^{\frac{-1}{2\lfloor d/2\rfloor+2}}$ we get*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda},\nu)-W_0(\mu_{\theta^*},\nu)\right]\lesssim n^{-\frac{1}{2\lfloor d/2\rfloor+2}}\log(n), \tag{A.36}$$

where $\lesssim$ hides a constant that depends on R and d.

Proof. In order to drive the approximation term towards 0, the regularization parameter will converge towards 0. And in this case, $\lambda \log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right) \sim_{\lambda\to 0} \lambda \log(\lambda^{-1})$. Next as we have assumed that all the distributions have n samples, the estimation term equals $\frac{M_\lambda}{\sqrt{n}} = \frac{M_d R^{\lfloor d/2 \rfloor+1}}{\lambda^{\lfloor d/2 \rfloor}\sqrt{n}}$. To balance these two terms we set $\lambda_n = n^{\frac{-1}{2\lfloor d/2 \rfloor+2}}$, and with this choice of regularization parameter, there exists a constant $M_{R,d} > 0$ such that for n sufficiently large,

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \leq M_{R,d} n^{\frac{-1}{2\lfloor d/2 \rfloor+2}} \log(n). \tag{A.37}$$

□

A.2 Proof of Theorem 3.2 - Part (ii)

In this section, for a fixed regularization parameter $\lambda > 0$, we substitute the regularized OT cost $W_\lambda(\mu,\nu)$ by its de-biased counterpart, the Sinkhorn divergence $S_\lambda(\mu,\nu)$ as defined by (2.8). In other terms, we now investigate the behavior of the following estimator

$$\hat{\theta}_\lambda^S = \underset{\theta\in\Sigma_K}{\arg\min}\, S_\lambda(\hat{\mu}_\theta, \hat{\nu}). \tag{A.38}$$

We will derive an upper bound on the expected excess risk of the resulting estimator $\hat{\theta}_\lambda^S$. Then, this upper bound allows to derive the convergence rate claimed in part (ii) of Theorem 3.2 using a tradeoff between approximation and estimation errors as discussed in Section 3.2.2. The results of this section rely in part on the approximation error of the Sinkhorn divergence S_λ with respect to W_0 established in Section 3.2.1 which is an adaptation of Theorem 1 from Chizat et al. [2020]. This adaptation is reminded in the next corollary.

Corollary A.2. *Suppose that assumptions A.2 and A.3 hold. Then, there exists a constant $M_I > 0$ such that*

$$\forall \theta \in \Sigma_K,\ \ |S_\lambda(\mu_\theta, \nu) - W_0(\mu_\theta, \nu)| \leq M_I \lambda^2. \tag{A.39}$$

The explicit value of M_I is given in equation (3.7) and the motives for Assumptions A.2 and A.3 can be found in Section 3.2.1.

A.2.1 Decomposition of the excess risk

To analyze the excess risk of the above estimator, a first step is to use an upper bound that can be interpreted as a decomposition of the excess risk between an estimation error and an approximation error.

Lemma A.4. *Suppose that Assumptions A.2 and A.3 hold true. Then, the following inequality holds*

$$0 \leq W_0(\mu_{\hat{\theta}_\lambda^S}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq \underbrace{2 \sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\hat{\mu}_\theta, \hat{\nu})|}_{\textit{Estimation error}} + \underbrace{2M_I\lambda^2,}_{\textit{Approximation error}} \tag{A.40}$$

where M_I is a positive constant defined in equation (3.7).

The proof of Lemma A.4 can be found in Section C.1 of the Appendix. To control the estimation error, we decompose this error into two terms

$$\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})| \leq \sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})| + \sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\hat{\nu}) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})|. \tag{A.41}$$

A.2.2 Control of the empirical processes for the loss function S_λ

From Lemma A.4 and equation (A.41), we are led to obtain upper bounds on the expectation of the following empirical processes

$$\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})|, \quad \text{and} \quad \sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\hat{\nu}) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})|,$$

that are given in the proposition below.

Proposition A.7. *Suppose that Assumption A.1 holds true.*

(i) *If n samples from ν are available, we have*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} \left|S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})\right|\right] \lesssim \frac{2M_\lambda}{\sqrt{n}}, \tag{A.42}$$

where M_λ is defined in equation (A.11).

(ii) *If for all $k \in \{1,\ldots,K\}$, m_k samples from μ_k are available, we have*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} \left|S_\lambda(\mu_\theta,\hat{\nu}) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})\right|\right] \lesssim \frac{2M_\lambda}{\sqrt{\underline{m}}}, \tag{A.43}$$

where $\underline{m} = \min(m_1,\ldots,m_K)$, and M_λ is defined in equation (A.11).

The proof of Proposition A.7 can be found in Section C.2.

A.2.3 Expected excess risk of the Sinkhorn divergence S_λ

Thanks to Proposition A.7 we can deduce the following upper bound on the convergence rate for the expected excess risk of the estimator $\hat{\theta}^S_\lambda$.

Proposition A.8. *Suppose that Assumptions A.1, A.2, and A.3 hold. If for all $k \in \{1,\ldots,K\}$, m_k samples from μ_k are available, and n samples are available from ν,*

(i) *we can propose the following bound for the expected excess risk of $\hat{\theta}^S_\lambda$.*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_\lambda},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim \underbrace{\frac{4M_\lambda}{\sqrt{\underline{m}}} + \frac{4M_\lambda}{\sqrt{n}}}_{\textit{Estimation error}} + \underbrace{2M_I\lambda^2}_{\textit{Approximation error}}, \tag{A.44}$$

where $\underline{m} = \min(m_1,\ldots,m_K)$, M_I is a positive constant defined in equation (3.7), and M_λ is defined in equation (A.11).

(ii) *Make the additional assumption that for each μ_k, at least n samples are available, then setting $\lambda_n = n^{\frac{-1}{2\lfloor d/2\rfloor+4}}$ we get the following rate of convergence for the expected excess risk of $\hat{\theta}^S_{\lambda_n}$.*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_{\lambda_n}},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim n^{\frac{-2}{2\lfloor d/2\rfloor+4}}, \tag{A.45}$$

we $\lesssim$ hides a constant that depends on R and d.

Proof. Let us begin with point (i) of Proposition A.8. Combining inequalities (A.40) and (A.41) with the upper bounds (A.42) and (A.43), we reach the upper bound for the expected excess risk that is proposed in equation (A.44).

For point (ii) we follow the same reasoning as in Corollary A.1. That is, we set λ in order to balance the estimation and the approximation error. And solving

$$\frac{1}{\lambda^{\lfloor d/2\rfloor}\sqrt{n}} = \lambda^2 \tag{A.46}$$

yields $\lambda_n = n^{\frac{-1}{2\lfloor d/2\rfloor+4}}$. Injecting this value in equation (A.44) we recover the rate of convergence of equation (A.45). □

A.3 Proof of Theorem 3.3 (i) and (iii)

To reach the rate of convergence claimed in Theorem 3.2 we relied on the following bound

$$\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| \leqslant \sup_{\psi\in\mathscr{C}^{d'}_{M_\lambda}(\mathcal{Y})} \int \psi d(\nu-\hat{\nu}), \tag{A.47}$$

where M_λ and d' are both defined in equation (A.16). We now exploit an other bound on the estimation error that is independent of the regularization parameter λ. The proofs presented on this section all relied on variations of the following inequality

$$\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| \leq \sup_{\varphi\in\mathcal{F}_R} \left|\int_{\mathcal{Y}} \varphi d(\hat{\nu}-\nu)\right| + \int_{\mathcal{Y}} \|y\|^2 d(\hat{\nu}-\nu), \tag{A.48}$$

where $\mathcal{F}_R$ is the set of concave and R-Lipschitz functions on $B(0,R)$. From this last bound (A.48) we derive the following upper bound for the estimation error.

Proposition A.9. *Let $\lambda \geq 0$. Suppose that Assumption A.1 holds.*

(i) *If n samples from ν are available, then it holds that*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})|\right] \lesssim \begin{cases} R^2 n^{-1/2} & \textit{if} \quad d<4, \\ R^2 n^{-1/2}\log(n) & \textit{if} \quad d=4, \\ R^2 n^{-2/d} & \textit{if} \quad d>4. \end{cases} \tag{A.49}$$

(ii) *If for each distribution μ_k, m_k samples are available, then*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\hat{\nu}) - W_\lambda(\hat{\mu}_\theta,\hat{\nu})|\right] \lesssim \begin{cases} R^2 \underline{m}^{-1/2} & \textit{if} \quad d<4, \\ R^2 \underline{m}^{-1/2}\log(\underline{m}) & \textit{if} \quad d=4, \\ R^2 \underline{m}^{-2/d} & \textit{if} \quad d>4, \end{cases} \tag{A.50}$$

where $\underline{m} = \min(m_1,\ldots,m_K)$.

The proof of Proposition A.9 can be found in Section D. We have the elements to get the estimation error under control. Next, we aim for a bound on the empirical excess risk of $\hat{\theta}_\lambda$ for $\lambda \geq 0$.

Proposition A.10. *Set $\lambda \geq 0$ and suppose that Assumption A.1 holds true. Also assume that n samples are drawn from ν and denote by $\underline{m} = \min(m_1,\ldots,m_K)$ where m_k is the number of samples from μ_k. Then, the expected excess risk of the estimator $\hat{\theta}_\lambda$ defined in equation (3.1) can be upper bounded by*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim \left\{\begin{array}{lll} R^2\min(\underline{m},n)^{-1/2} & \textit{if} & d<4, \\ R^2\min(\underline{m},n)^{-1/2}\log(\min(\underline{m},n)) & \textit{if} & d=4, \\ R^2\min(\underline{m},n)^{-2/d} & \textit{if} & d>4. \end{array}\right\} + B(\lambda) \tag{A.51}$$

where $B(\lambda)$ is the bias induced by the regularization term and whose expression can be found in equation (A.3).

Proof. Set $\lambda \geq 0$. Using Lemma A.1 we have

$$W_0(\mu_{\hat{\theta}_\lambda},\nu) - W_0(\mu_{\theta^*},\nu) \leq 2\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| + 2B(\lambda). \tag{A.52}$$

And

$$\sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\hat{\mu}_\theta,\hat{\nu})| \leq \sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| + \sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\hat{\nu}) - W_\lambda(\hat{\mu}_\theta,\hat{\nu})|.$$

After taking the expectation of this last inequality, a straight application of Proposition A.9 allows us to control the estimation term of decomposition (A.52). We conclude by adding the bias term $B(\lambda)$. □

Remark A.3. *Note that this last result holds for $\lambda \geq 0$. We can thus use it to prove point (i) and point (iii) of Theorem 3.3.*

A.4 Proof of Theorem 3.3 (ii)

The proof of point (ii) of Theorem 3.3 is very similar to the proof of point (i) and (iii), it relies on the next Proposition.

Proposition A.11. *Set $\lambda > 0$. Suppose that Assumption A.1 holds.*

(i) *If n samples from ν are available, then it holds that*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})|\right] \lesssim \begin{cases} R^2 n^{-1/2} & \text{if} \quad d < 4, \\ R^2 n^{-1/2}\log(n) & \text{if} \quad d = 4, \\ R^2 n^{-2/d} & \text{if} \quad d > 4. \end{cases} \tag{A.53}$$

(ii) *If for each distribution μ_k, m_k samples are available, then*

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\hat{\nu}) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})|\right] \lesssim \begin{cases} R^2 \underline{m}^{-1/2} & \text{if} \quad d < 4, \\ R^2 \underline{m}^{-1/2}\log(\underline{m}) & \text{if} \quad d = 4, \\ R^2 \underline{m}^{-2/d} & \text{if} \quad d > 4, \end{cases} \tag{A.54}$$

where $\underline{m} = \min(m_1,\ldots,m_K)$.

Proof. For concision, we prove only point (i) and assume $d > 4$. For $\lambda > 0$ set, we have

$$\begin{aligned}\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})| \le{}& \sup_{\theta\in\Sigma_K} |W_\lambda(\mu_\theta,\nu) - W_\lambda(\mu_\theta,\hat{\nu})| \\ &+ \frac{1}{2}\left(|W_\lambda(\nu,\nu) - W_\lambda(\nu,\hat{\nu})| + |W_\lambda(\nu,\hat{\nu}) - W_\lambda(\hat{\nu},\hat{\nu})|\right).\end{aligned} \tag{A.55}$$

Then, using the results established for W_λ in Proposition A.9, we get

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})|\right] \lesssim R^2 n^{-2/d}. \tag{A.56}$$

This is the result claimed for point (i) of Proposition A.11 in the case $d > 4$. Point (ii) can be deduced with the same reasoning. □

Proposition A.12. *Set $\lambda > 0$ and suppose that Assumptions A.1, A.2 and A.3 hold true. Also assume that n samples are drawn from ν and denote by $\underline{m} = \min(m_1,\ldots,m_K)$ where m_k is the number of samples from μ_k. Then, the expected excess risk of the estimator $\hat{\theta}^S_\lambda$ defined in equation (3.1) can be upper bounded by*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^S_\lambda},\nu) - W_0(\mu_{\theta^*},\nu)\right] \lesssim \left\{\begin{array}{ll} R^2 \min(\underline{m},n)^{-1/2} & \text{if} \quad d < 4, \\ R^2 \min(\underline{m},n)^{-1/2}\log(\min(\underline{m},n)) & \text{if} \quad d = 4, \\ R^2 \min(\underline{m},n)^{-2/d} & \text{if} \quad d > 4. \end{array}\right\} + M_I\lambda^2. \tag{A.57}$$

where M_I is a constant defined in equation (3.7).

Proof. For $\lambda > 0$ set, Lemma A.4 gives

$$W_0(\mu_{\hat{\theta}^S_\lambda},\nu) - W_0(\mu_{\theta^*},\nu) \le 2\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})| + 2M_I\lambda^2. \tag{A.58}$$

Next, we break down the estimation error as follows

$$\sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})| \le \sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\nu) - S_\lambda(\mu_\theta,\hat{\nu})| + \sup_{\theta\in\Sigma_K} |S_\lambda(\mu_\theta,\hat{\nu}) - S_\lambda(\hat{\mu}_\theta,\hat{\nu})|.$$

Finally, Proposition A.11 allows to bound in expectation the last equation. Hence, we derive the rate of convergence announced. □

A.5 Proof of Theorem 3.4

In this section, we take into account the Sinkhorn algorithm error depending on the number of iterations, and how it impacts the rate of convergence of regularized estimators when computed with this specific algorithm.

A.5.1 Proof of Theorem 3.4 - Part (i)

For $a = \sum_{i=1}^{I} a_i \delta_{x_i}$, and $b = \sum_{j=1}^{J} b_j \delta_{y_j}$ two discrete distributions, we denote by

$$W_\lambda^{(\ell)}(a,b) = \sum_{i=1}^{I} a_i \varphi_i^{(\ell)} + \sum_{j=1}^{J} b_j \psi_j^{(\ell)}, \tag{A.59}$$

the approximation of the regularized OT cost $W_\lambda(a,b)$ that is returned by the Sinkhorn approximation after ℓ iterations (see Definition 2.3). The variables $\varphi^{(\ell)}$ and $\psi^{(\ell)}$ denote the dual variables after ℓ iterations of the Sinkhorn algorithm. We thus consider the estimator used in our numerical experiments that is defined as

$$\hat{\theta}_\lambda^{(\ell)} = \arg\min_{\theta \in \Sigma_K} W_\lambda^{(\ell)}(\hat{\mu}_\theta, \hat{\nu}). \tag{A.60}$$

The computational complexity of Sinkhorn algorithm has been studied in Chizat et al. [2020] and we remind the error after ℓ iterations of the Sinkhorn algorithm with respect to the regularized OT cost.

Proposition A.13. *[Chizat et al., 2020, Proposition 2]. Assume that $\lambda > 0$. For $a = \sum_{i=1}^{I} a_i \delta_{x_i}$ and $b = \sum_{j=1}^{J} b_j \delta_{y_j}$ two discrete distributions and a ground cost set to $c(x,y) = \|x-y\|^2$ on $\mathbb{R}^d$. The approximation of the regularized OT cost after ℓ iterations of the Sinkhorn algorithm satisfies:*

$$|W_\lambda^{(\ell)}(a,b) - W_\lambda(a,b)| \leq \frac{\|c\|_\infty^2}{\lambda \ell} \tag{A.61}$$

where $\|c\|_\infty = \max_{(i,j)} \|x_i - y_j\|^2$.

Remark A.4. *If the discrete distributions a and b have both their supports subsets of $B(0,R)$, it implies that $\max_{(i,j)} \|x_i - y_j\|^2 \leq 4R^2$. An then, the quantity $\|c\|_\infty^2$ can be upper bounded by $\|c\|_\infty^2 \leq 16R^4$.*

The next Lemma yields an upper bound for the excess risk of the estimator $\hat{\theta}_\lambda^{(\ell)}$ after ℓ iterations. This upper bound can be decomposed into an estimation error, an approximation and an algorithmic error.

Lemma A.5. *Suppose that Assumption A.1 holds true. For $\hat{\theta}_\lambda^{(\ell)}$ the Sinkhorn estimator defined in (A.60), its excess risk can be upper bounded in the following manner:*

$$W_0(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq \underbrace{4 \sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)|}_{\textit{Estimation error}} + \underbrace{4B(\lambda)}_{\textit{Approximation error}} + \underbrace{\frac{2\|c\|_\infty^2}{\lambda \ell}}_{\textit{Algorithm error}}. \tag{A.62}$$

where $B(\lambda)$ is defined in (A.3).

The proof of Lemma A.5 is deferred to Section B.3 in the Appendix.

Corollary A.3. *Suppose that Assumption A.1 holds true, that for every μ_k we have m_k samples available, and that for ν we have n samples. For $\hat{\theta}_\lambda^{(\ell)}$ the regularized estimator defined in A.60, the expected excess risk can be upper bounded in the following manner*

(i)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^{(\ell)}_{\lambda}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim \underbrace{\frac{4M_\lambda}{\sqrt{\underline{m}}} + \frac{4M_\lambda}{\sqrt{n}}}_{\textit{Estimation error}} \underbrace{+ 8d\lambda \log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right)}_{\textit{Approximation error}} + \underbrace{\frac{32R^4}{\lambda\ell}}_{\textit{Algorithm error}}, \tag{A.63}$$

where $\underline{m} = \min(m_1, \ldots, m_K)$, and M_λ is defined in equation (A.11).

(ii) *Make the additional assumption that for each component μ_k at least n samples are available. Then, we propose a parameter λ_n and a number of Sinkhorn iterations ℓ_n that allow to recover the rate of convergence of Corollary A.1 for $\hat{\theta}^{(\ell_n)}_{\lambda_n}$ while taking into account the algorithm error.*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^{(\ell_n)}_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}} \log(n) \quad \textit{with} \quad \begin{cases} \lambda_n = n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}}, \\ \ell_n \geq 32R^4 n^{\frac{2}{2\lfloor d/2 \rfloor + 2}} (\log(n))^{-1}. \end{cases} \tag{A.64}$$

Proof. Considering point (i), the two bounds on

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} |W_\lambda(\mu_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)|\right], \quad \text{and} \quad \mathbb{E}\left[\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu})|\right],$$

established in Proposition A.2 and in Proposition A.5 respectively, as well as the inequality established in Lemma A.5 yield the upper bound (A.63). Let us now prove point (ii).

We begin as in Corollary A.1. First, we set $\lambda_n = n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}}$ to balance the estimation error and the approximation error. With this choice of regularization parameter the expected excess risk reads

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^{(\ell)}_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}} \log(n) + \frac{32R^4}{n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}} \ell}.$$

The issue is now to bring the algorithm error $\frac{32R^4}{n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}} \ell}$ below the desired rate of $n^{\frac{-1}{2\lfloor d/2 \rfloor + 2}} \log(n)$. To do so, we set $\ell_n \geq 32R^4 n^{\frac{2}{2\lfloor d/2 \rfloor + 2}} (\log(n))^{-1}$. And then, we recover the rate of convergence claimed in equation (A.64) for $\hat{\theta}^{(\ell_n)}_{\lambda_n}$. □

A.5.2 Proof of Theorem 3.4 - Part (ii)

We now adapt the proof of the previous paragraph to the case of the Sinkhorn divergence obtained after ℓ iterations of the Sinkhorn algorithm. For two discrete distributions $a = \sum_{i=1}^I a_i \delta_{x_i}$, and $b = \sum_{j=1}^J b_j \delta_{y_j}$, we denote by

$$S^{(\ell)}_\lambda(a, b) = W^{(\ell)}_\lambda(a, b) - \frac{1}{2}\left(W^{(\ell)}_\lambda(a, a) + W^{(\ell)}_\lambda(b, b)\right). \tag{A.65}$$

where $W^{(\ell)}_\lambda(a, b)$ is the approximation of the regularized OT cost returned by the Sinkhorn algorithm after ℓ iterations defined in equation (A.59). Using Proposition A.13, we get

$$|S^{(\ell)}_\lambda(a, b) - S_\lambda(a, b)| \leq \frac{2\|c\|^2_\infty}{\lambda\ell}. \tag{A.66}$$

We will then analyze the estimator

$$\hat{\theta}^{S(\ell)}_\lambda = \arg\min_{\theta \in \Sigma_K} S^{(\ell)}_\lambda(\hat{\mu}_\theta, \hat{\nu}), \tag{A.67}$$

Lemma A.6. *Suppose that Assumptions A.2 and A.3 hold, and denote by* $\|c\|_\infty = \max_{i,j} \|X_i - Y_j\|^2$*, where* (X_i) *and* (Y_j) *refer to the samples draw from the distributions* μ, ν *respectively.*

(i) *In this case, the excess risk of* $\hat{\theta}_\lambda^{S(\ell)}$ *is bounded by*

$$0 \leq W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq \underbrace{4 \sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)|}_{\text{Estimation error}} + \underbrace{8M_I\lambda^2}_{\text{Approximation error}} + \underbrace{\frac{4\|c\|_\infty^2}{\lambda\ell}}_{\text{Algorithm error}} \tag{A.68}$$

(ii) *Make the additional Assumption A.1, and suppose that for all* $k \in \{1, \ldots, K\}$*,* m_k *samples from* μ_k *are available, and* n *samples are available from* ν*. Then,*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim \underbrace{\frac{8M_\lambda}{\sqrt{\underline{m}}} + \frac{8M_\lambda}{\sqrt{n}}}_{\text{Estimation error}} + \underbrace{8M_I\lambda^2}_{\text{Approximation error}} + \underbrace{\frac{64R^4}{\lambda\ell}}_{\text{Algorithm error}} \tag{A.69}$$

where $\underline{m} = \min(m_1, \ldots, m_K)$*,* M_I *is a positive constant defined in equation (3.7), and* M_λ *is defined in equation (A.11).*

(iii) *Finally assume that for each component* μ_k*, at least* n *samples are available. Then, the estimator* $\hat{\theta}_{\lambda_n}^{S(\ell_n)}$ *admits the following rate of convergence*

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_{\lambda_n}^{S(\ell_n)}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{\frac{-2}{2\lfloor d/2 \rfloor + 4}}, \quad \text{with} \quad \begin{cases} \lambda_n = n^{\frac{-1}{2\lfloor d/2 \rfloor + 4}}, \\ \ell_n \geq 64R^4 n^{\frac{3}{2\lfloor d/2 \rfloor + 4}}. \end{cases} \tag{A.70}$$

Where $\lesssim$ *hides a constant that depends on* R *and* d

The proof of Lemma A.6 can be found in Section C.3 of the Appendix.

A.6 Proof of Theorem 3.5

This short section gives the elements to understand Theorem 3.5. The next property is a more general statement of Theorem 3.5 when the regularization parameter λ and the number of Sinkhorn iterations ℓ are not chosen to decrease at the same pace as the estimation error.

Proposition A.14. *Suppose that Assumptions A.1, A.2, and A.3 hold true and that the dimension is such that* $d > 4$*. If* n *samples are available from* ν *and for every* μ_k*,* m_k *samples are available, then, the expected excess risk the estimator* $\hat{\theta}_\lambda^{(l)}$ *can be upper bounded as follows.*

(i)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim \underbrace{\min(\underline{m}, n)^{-2/d}}_{\text{Estimation error}} + \underbrace{8d\lambda \log\left(\frac{8\exp(2)R^2}{\sqrt{d}\lambda}\right)}_{\text{Approximation error}} + \underbrace{\frac{32R^4}{\lambda\ell}}_{\text{Algorithm error}}, \tag{A.71}$$

where $\underline{m} = \min(m_1, \ldots, m_K)$.

And for $\hat{\theta}_\lambda^{S(l)}$*, we have*

(ii)

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim \underbrace{\min(\underline{m}, n)^{-2/d}}_{\text{Estimation error}} + \underbrace{8M_I\lambda^2}_{\text{Approximation error}} + \underbrace{\frac{64R^4}{\lambda\ell}}_{\text{Algorithm error}}, \tag{A.72}$$

where $\underline{m} = \min(m_1, \ldots, m_K)$.

At this point of the article, proving Proposition A.14 more or less consists in gathering already established results.

Proof. For point (i), Lemma A.5 gives

$$W_0(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 4 \sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)| + 4B(\lambda) + \frac{2\|c\|_\infty^2}{\lambda \ell}.$$

Proposition A.9, gives $\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)| \leq \min(\underline{m}, n)^{-2/d}$. Then, we substitute $B(\lambda)$ by its expression and write $\|c\|_\infty^2 \leq 16R^4$ thanks to Assumption A.1.

For point (ii), we exploit Lemma A.6 to write

$$W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 4 \sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)| + 8M_I \lambda^2 + \frac{4\|c\|_\infty^2}{\lambda \ell}.$$

Finally, Proposition A.11 allows us to control the expectation of $\sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)|$, which concludes the proof. □

B Proofs of Section A.1

This appendix contains auxiliary results related to the case where W_λ is the loss function.

B.1 Proof of Lemma A.1

Proof. We begin with the decomposition

$$\begin{aligned} W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu) = {} & W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu) + W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu) - W_\lambda(\mu_{\theta^*}, \nu) \\ & + W_\lambda(\mu_{\theta^*}, \nu) - W_0(\mu_{\theta^*}, \nu). \end{aligned} \tag{B.1}$$

The first and the last differences of equation (B.1) are controlled by the approximation error $B(\lambda)$ of the regularized Wasserstein distance W_λ with respect to W_0. We thus have

$$W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu) - W_\lambda(\mu_{\theta^*}, \nu) + 2B(\lambda). \tag{B.2}$$

And we can rewrite

$$\begin{aligned} W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu) - W_\lambda(\mu_{\theta^*}, \nu) = {} & W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu) - W_\lambda(\hat{\mu}_{\hat{\theta}_\lambda}, \hat{\nu}) + W_\lambda(\hat{\mu}_{\hat{\theta}_\lambda}, \hat{\nu}) - W_\lambda(\hat{\mu}_{\theta^*}, \hat{\nu}) \\ & + W_\lambda(\hat{\mu}_{\theta^*}, \hat{\nu}) - W_\lambda(\mu_{\theta^*}, \nu). \end{aligned} \tag{B.3}$$

For the second difference of equation (B.3), we have $W_\lambda(\hat{\mu}_{\hat{\theta}_\lambda}, \hat{\nu}) - W_\lambda(\hat{\mu}_{\theta^*}, \hat{\nu}) \leq 0$ as $\hat{\theta}_\lambda \in \arg\min_{\theta \in \Sigma_K} W_\lambda(\hat{\mu}_\theta, \hat{\nu})$. Next, we bound the first and last differences of the same equation (B.3) by $\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)|$. Hence, we get

$$W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu) - W_\lambda(\mu_{\theta^*}, \nu) \leq 2 \sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)|.$$

Injecting this last inequality in equation (B.2) we finally derive

$$W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 2 \sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \nu)| + 2B(\lambda). \tag{B.4}$$

which is the result claimed in Lemma A.1.

□

B.2 Proof Proposition A.5

This proof is based on a very similar reasoning to the one of proof of Proposition A.3 as well as the reuse of results from Section A.1.2.

Proof. We begin by setting $\theta \in \Sigma_K$. Let us denote by φ (resp.$\hat{\varphi}$) an optimal potential chosen as in Lemma A.2 when considering the semi dual formulation of the regularized optimal transport problem between μ_θ and $\hat{\nu}$ (resp. $\hat{\mu}_\theta$ and $\hat{\nu}$). Thus,

$$\begin{aligned}W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu}) &= \int \hat{\varphi}_{\hat{\nu}}^{c,\lambda} d\hat{\mu}_\theta + \int \hat{\varphi} d\hat{\nu} - \left(\int \varphi_{\hat{\nu}}^{c,\lambda} d\mu_\theta + \int \varphi d\hat{\nu} \right) \\ &= \int \hat{\varphi}_{\hat{\nu}}^{c,\lambda} d\hat{\mu}_\theta - \int \hat{\varphi}_{\hat{\nu}}^{c,\lambda} d\mu_\theta \\ &+ \underbrace{\left(\int \hat{\varphi}_{\hat{\nu}}^{c,\lambda} d\mu_\theta + \int \hat{\varphi} d\hat{\nu} - \left(\int \varphi_{\hat{\nu}}^{c,\lambda} d\mu_\theta + \int \varphi d\hat{\nu} \right) \right)}_{\leq 0}\end{aligned}$$

The optimality of the variable φ with respect to the measures μ_θ and $\hat{\nu}$ ensures the last term of the previous equation to be non positive. Hence,

$$W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu}) \leq \int \hat{\varphi}_{\hat{\nu}}^{c,\lambda} d(\hat{\mu}_\theta - \mu_\theta).$$

With a slight modification of the last argument we get

$$W_\lambda(\mu_\theta, \hat{\nu}) - W_\lambda(\hat{\mu}_\theta, \hat{\nu}) \leq \int \varphi_{\hat{\nu}}^{c,\lambda} d(\mu_\theta - \hat{\mu}_\theta).$$

Combining these last inequalities, we have

$$|W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu})| \leq \sup_{\psi \in L^\infty(\mathcal{Y})} \left| \sum_{k=1}^K \theta_k \int \psi_{\hat{\nu}}^{c,\lambda} d(\mu_k - \hat{\mu}_k) \right| \leq \sum_{k=1}^K \theta_k \sup_{\psi \in L^\infty(\mathcal{Y})} \left| \int \psi_{\hat{\nu}}^{c,\lambda} d(\mu_k - \hat{\mu}_k) \right|.$$

Next, the application of Lemma A.2 when computing a c-transform w.r.t. $\hat{\nu}$ gives that $\psi_{\hat{\nu}}^{c,\lambda} \in \mathscr{C}_{M_\lambda}^{d'}(B(0,R))$ with d' and M_λ both defined in equation (A.11). Hence, for $k \in \{1, ..., K\}$, using the same ingredients as in Proposition A.3 , we reach the study of an empirical process indexed by the class of functions $\mathscr{C}_{M_\lambda}^{d'}(B(0,R))$. Finally, the straight application of Proposition A.4 yields

$$\mathbb{E}\left[\sup_{f \in \mathscr{C}_{M_\lambda}^{d'}((B(0,R))} \left| \int f d(\mu_k - \hat{\mu}_k) \right| \right] \lesssim \frac{M_\lambda}{\sqrt{m_k}},$$

where m_k is the number of observations sampled from distribution μ_k. As $\sum_{k=1}^K \theta_k = 1$, it follows that

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu})| \right] \lesssim \frac{M_\lambda}{\sqrt{\underline{\mathrm{m}}}}, \tag{B.5}$$

where $\underline{\mathrm{m}} = \min\{m_k : 1 \leq k \leq K\}$. □

B.3 Proof of Lemma A.5

Proof. We first write $W_0(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu) = W_0(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_0(\mu_{\hat{\theta}_\lambda}, \nu) + W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)$.

The second term $W_0(\mu_{\hat{\theta}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu)$ being controlled with lemma A.1, we can focus on

$$W_0(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_0(\mu_{\hat{\theta}_\lambda}, \nu) \leq 2B(\lambda) + W_\lambda(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - W_\lambda(\mu_{\hat{\theta}_\lambda}, \nu).$$

Then,

$$\begin{aligned}W_\lambda(\mu_{\hat\theta_\lambda^{(\ell)}},\nu)-W_\lambda(\mu_{\hat\theta_\lambda},\nu)&\leq W_\lambda(\mu_{\hat\theta_\lambda^{(\ell)}},\nu)-W_\lambda(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)+W_\lambda(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)-W_\lambda(\hat\mu_{\hat\theta_\lambda},\hat\nu)\\&\quad+W_\lambda(\hat\mu_{\hat\theta_\lambda},\hat\nu)-W_\lambda(\mu_{\hat\theta_\lambda},\nu)\\&\leq W_\lambda(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)-W_\lambda(\hat\mu_{\hat\theta_\lambda},\hat\nu)+2\sup_{\theta\in\Sigma_K}|W_\lambda(\hat\mu_\theta,\hat\nu)-W_\lambda(\mu_\theta,\nu)|.\end{aligned}$$

We finally introduce the approximation of the regularized Wasserstein distance that is the Sinkhorn output after ℓ iterations.

$$\begin{aligned}W_\lambda(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)-W_\lambda(\hat\mu_{\hat\theta_\lambda},\hat\nu)&=W_\lambda(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)-W_\lambda^{(\ell)}(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)+W_\lambda^{(\ell)}(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)-W_\lambda^{(\ell)}(\hat\mu_{\hat\theta_\lambda},\hat\nu)\\&\quad+W_\lambda^{(\ell)}(\hat\mu_{\hat\theta_\lambda},\hat\nu)-W_\lambda(\hat\mu_{\hat\theta_\lambda},\hat\nu)\\&\leq\frac{2\|c\|_\infty^2}{\lambda\ell},\end{aligned}$$

as $W_\lambda^{(\ell)}(\hat\mu_{\hat\theta_\lambda^{(\ell)}},\hat\nu)-W_\lambda^{(\ell)}(\hat\mu_{\hat\theta_\lambda},\hat\nu)\leq 0$ due to the definition of $\hat\theta_\lambda^{(\ell)}=\arg\min_{\theta\in\Sigma_K}W_\lambda^{(\ell)}(\hat\mu_\theta,\hat\nu)$. We also made use of Proposition A.13, to bound the first and last difference. Gathering the previous inequalities, we get

$$W_0(\mu_{\hat\theta_\lambda^{(\ell)}},\nu)-W_0(\mu_{\hat\theta_\lambda},\nu)\leq 2B(\lambda)+2\sup_{\theta\in\Sigma_K}|W_\lambda(\hat\mu_\theta,\hat\nu)-W_\lambda(\mu_\theta,\nu)|+\frac{2\|c\|_\infty^2}{\lambda\ell}.$$

And Lemma A.1 gives

$$W_0(\mu_{\hat\theta_\lambda},\nu)-W_0(\mu_{\theta^*},\nu)\leq 2\sup_{\theta\in\Sigma_K}|W_\lambda(\hat\mu_\theta,\hat\nu)-W_\lambda(\mu_\theta,\nu)|+2B(\lambda).$$

Putting all the pieces together, we finally get

$$W_0(\mu_{\hat\theta_\lambda^{(\ell)}},\nu)-W_0(\mu_{\theta^*},\nu)\leq 4B(\lambda)+\frac{2\|c\|_\infty^2}{\lambda\ell}+4\sup_{\theta\in\Sigma_K}|W_\lambda(\hat\mu_\theta,\hat\nu)-W_\lambda(\mu_\theta,\nu)|,$$

as announced in Lemma A.5. □

C Proofs of Section A.2

This appendix contains auxiliary results related to the case where the loss function is S_λ.

C.1 Proof of Lemma A.4

Proof. We start with

$$\begin{aligned}0\leq W_0(\mu_{\hat\theta_\lambda^S},\nu)-W_0(\mu_{\theta^*},\nu)&=W_0(\mu_{\hat\theta_\lambda^S},\nu)-S_\lambda(\mu_{\hat\theta_\lambda^S},\nu)+S_\lambda(\mu_{\hat\theta_\lambda^S},\nu)-S_\lambda(\mu_{\theta^*},\nu)\\&\quad+S_\lambda(\mu_{\theta^*},\nu)-W_0(\mu_{\theta^*},\nu)\\&\leq 2M_I\lambda^2+S_\lambda(\mu_{\hat\theta_\lambda^S},\nu)-S_\lambda(\mu_{\theta^*},\nu).\end{aligned}\tag{C.1}$$

We now study the remaining term $S_\lambda(\mu_{\hat\theta_\lambda^S},\nu)-S_\lambda(\mu_{\theta^*},\nu)$ thanks to the decomposition

$$\begin{aligned}S_\lambda(\mu_{\hat\theta_\lambda^S},\nu)-S_\lambda(\mu_{\theta^*},\nu)&=S_\lambda(\mu_{\hat\theta_\lambda^S},\nu)-S_\lambda(\hat\mu_{\hat\theta_\lambda^S},\hat\nu)+S_\lambda(\hat\mu_{\hat\theta_\lambda^S},\hat\nu)-S_\lambda(\hat\mu_{\theta^*},\hat\nu)\\&\quad+S_\lambda(\hat\mu_{\theta^*},\hat\nu)-S_\lambda(\mu_{\theta^*},\nu)\\&\leq 2\sup_{\theta\in\Sigma_K}|S_\lambda(\mu_\theta,\nu)-S_\lambda(\hat\mu_\theta,\hat\nu)|.\end{aligned}\tag{C.2}$$

As $\hat{\theta}_\lambda^S \in \underset{\theta \in \Sigma_K}{\arg\min}\, S_\lambda(\hat{\mu}_\theta, \hat{\nu})$, we have upper bounded $S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^S}, \hat{\nu}) - S_\lambda(\hat{\mu}_{\theta^*}, \hat{\nu}) \leq 0$ to derive inequality (C.2). Gathering inequality (C.1) and inequality (C.2) we get

$$0 \leq W_0(\mu_{\hat{\theta}_\lambda^S}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 2M_I \lambda^2 + 2 \sup_{\theta \in \Sigma_K} |S_\lambda(\mu_\theta, \nu) - S_\lambda(\hat{\mu}_\theta, \hat{\nu})|, \tag{C.3}$$

as claimed in Lemma A.4.

□

C.2 Proof of Proposition A.7

We begin with the first point of Proposition A.7.

Proof. First, let us set $\theta \in \Sigma_K$ and write

$$\begin{aligned} S_\lambda(\mu_\theta, \nu) - S_\lambda(\mu_\theta, \hat{\nu}) &= W_\lambda(\mu_\theta, \nu) - \frac{1}{2}\left(W_\lambda(\mu_\theta, \mu_\theta) + W_\lambda(\nu, \nu)\right) - \left(W_\lambda(\mu_\theta, \hat{\nu}) - \frac{1}{2}\left(W_\lambda(\mu_\theta, \mu_\theta) + W_\lambda(\hat{\nu}, \hat{\nu})\right)\right) \\ &= W_\lambda(\mu_\theta, \nu) - W_\lambda(\mu_\theta, \hat{\nu}) + \frac{1}{2}\left(W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\nu, \nu)\right). \end{aligned}$$

Therefore,

$$|S_\lambda(\mu_\theta, \nu) - S_\lambda(\mu_\theta, \hat{\nu})| \leq |W_\lambda(\mu_\theta, \nu) - W_\lambda(\mu_\theta, \hat{\nu})| + \frac{1}{2}|\left(W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\nu, \nu)\right)|. \tag{C.4}$$

We have already derived an upper bound for $\mathbb{E}\left[\sup_{\theta \in \Sigma_K} \left|W_\lambda(\mu_\theta, \nu) - W_\lambda(\mu_\theta, \hat{\nu})\right|\right]$ in Section A.1.2. Indeed, we established that

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} \left|W_\lambda(\mu_\theta, \nu) - W_\lambda(\mu_\theta, \hat{\nu})\right|\right] \lesssim \frac{M_\lambda}{\sqrt{n}}$$

where the value of M_λ can be found in equation (A.11). Then, to bound $\mathbb{E}\left[|W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\nu, \nu)|\right]$, we use the triangle inequality

$$|W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\nu, \nu)| \leq |W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\hat{\nu}, \nu)| + |W_\lambda(\hat{\nu}, \nu) - W_\lambda(\nu, \nu)|.$$

To bound $|W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\hat{\nu}, \nu)|$, we follow the same line of argumentation as in the proof of Proposition A.3. First, using the regularity of the c-transform yields

$$|W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\hat{\nu}, \nu)| \leq \sup_{\varphi \in \mathscr{C}^{d'}_{M_\lambda}(\mathcal{Y})} \left|\int \varphi d(\nu - \hat{\nu})\right|,$$

where $d' = \lfloor d/2 \rfloor + 1$. Then, Proposition A.4 gives

$$\mathbb{E}\left[|W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\hat{\nu}, \nu)|\right] \lesssim \frac{M_\lambda}{\sqrt{n}}. \tag{C.5}$$

We control $|W_\lambda(\hat{\nu}, \nu) - W_\lambda(\nu, \nu)|$ exactly as $|W_\lambda(\hat{\nu}, \hat{\nu}) - W_\lambda(\hat{\nu}, \nu)|$. Hence,

$$\mathbb{E}\left[|W_\lambda(\hat{\nu}, \nu) - W_\lambda(\nu, \nu)|\right] \lesssim \frac{M_\lambda}{\sqrt{n}}. \tag{C.6}$$

The two terms of equation (C.4) are now under control in expectation. Thus, we reach the first point of Proposition A.7, which is

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} \left|S_\lambda(\mu_\theta, \nu) - S_\lambda(\mu_\theta, \hat{\nu})\right|\right] \lesssim \frac{2M_\lambda}{\sqrt{n}}. \tag{C.7}$$

□

The proof of the second point of Proposition A.7 is very similar to the previous proof.

Proof of the second point of Proposition A.7

Proof. For $\theta \in \Sigma_K$ we write

$$\begin{aligned} S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \hat{\nu}) &= W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - \frac{1}{2}\left(W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) + W_\lambda(\hat{\nu}, \hat{\nu})\right) - \left(W_\lambda(\mu_\theta, \hat{\nu}) - \frac{1}{2}\left(W_\lambda(\mu_\theta, \mu_\theta) + W_\lambda(\hat{\nu}, \hat{\nu})\right)\right) \\ &= W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu}) + \frac{1}{2}\left(W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\mu_\theta, \mu_\theta)\right). \end{aligned}$$

Therefore,

$$|S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \hat{\nu})| \leq |W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\mu_\theta, \hat{\nu})| + \frac{1}{2}|\left(W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\mu_\theta, \mu_\theta)\right)|. \tag{C.8}$$

We already derived an upper bound for $\mathbb{E}\left[\sup_{\theta\in\Sigma_K}\left|W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\hat{\mu}_\theta, \nu)\right|\right]$ in Proposition A.5, that is:

$$\mathbb{E}\left[\sup_{\theta\in\Sigma_K}\left|W_\lambda(\hat{\mu}_\theta, \hat{\nu}) - W_\lambda(\hat{\mu}_\theta, \nu)\right|\right] \lesssim \frac{M_\lambda}{\sqrt{\underline{\mathrm{m}}}}.$$

where $\underline{\mathrm{m}} = \min\{m_k : 1 \leq k \leq K\}$ and the expression M_λ can be found in (A.11).

To handle the difference between $W_\lambda(\mu_\theta, \mu_\theta)$ and its empirical counterpart $W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta)$, we write:

$$|W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\mu_\theta, \mu_\theta)| \leq |W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\hat{\mu}_\theta, \mu_\theta)| + |W_\lambda(\hat{\mu}_\theta, \mu_\theta) - W_\lambda(\mu_\theta, \mu_\theta)|$$

Let us denote by φ a optimal potential of the semi-dual problem associated to $W_\lambda(\hat{\mu}_\theta, \mu_\theta)$ and $\hat{\varphi}$ associated to $W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta)$ and write

$$\begin{aligned} W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\hat{\mu}_\theta, \mu_\theta) &= \int \hat{\varphi} d\hat{\mu}_\theta + \int \hat{\varphi}^{c,\lambda}_{\hat{\mu}_\theta} d\hat{\mu}_\theta - \left(\int \varphi d\hat{\mu}_\theta + \int \varphi^{c,\lambda}_{\hat{\mu}_\theta} d\mu_\theta\right) \\ &= \int \hat{\varphi}^{c,\lambda}_{\hat{\mu}_\theta} d\hat{\mu}_\theta - \int \hat{\varphi}^{c,\lambda}_{\hat{\mu}_\theta} d\mu_\theta \\ &\quad + \underbrace{\int \hat{\varphi} d\hat{\mu}_\theta + \int \hat{\varphi}^{c,\lambda}_{\hat{\mu}_\theta} d\mu_\theta - \left(\int \varphi d\hat{\mu}_\theta + \int \varphi^{c,\lambda}_{\hat{\mu}_\theta} d\mu_\theta\right)}_{\leq 0}. \end{aligned}$$

As φ is optimal with respect to $W_\lambda(\hat{\mu}_\theta, \mu_\theta)$ the last term is non positive. Consequently,

$$W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\hat{\mu}_\theta, \mu_\theta) \leq \int \hat{\varphi}^{c,\lambda}_{\hat{\mu}_\theta} d(\hat{\mu}_\theta - \mu_\theta).$$

With a similar argument, we have

$$W_\lambda(\hat{\mu}_\theta, \mu_\theta) - W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) \leq \int \varphi^{c,\lambda}_{\hat{\mu}_\theta} d(\mu_\theta - \hat{\mu}_\theta).$$

Thus

$$|W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\hat{\mu}_\theta, \mu_\theta)| \leq \sup_{\psi \in L^\infty(\mathcal{X})}\left|\int \psi^{c,\lambda}_{\hat{\mu}_\theta} d(\hat{\mu}_\theta - \mu_\theta)\right|.$$

And using Lemma A.2, we have that $\psi^{c,\lambda}_{\hat{\mu}_\theta}$ belongs to $\mathscr{C}^{d'}_{M_\lambda}(B(0,R))$, where $d' = \lfloor d/2 \rfloor + 1$ and the value of M_λ can be found in equation (A.11). We remind that the constant M_λ has been chosen to be independent of θ.

Therefore, we are led to bound $\sup_{\varphi \in \mathscr{C}^{d'}_{M_\lambda}(B(0,R))} \int \varphi d(\hat{\mu}_\theta - \mu_\theta)$. This empirical process has been studied in the proof of Proposition A.5, that one can find in Section B.2 of the Appendix. Therefore, we can right away propose the following upper bound

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \hat{\mu}_\theta) - W_\lambda(\hat{\mu}_\theta, \mu_\theta)|\right] \lesssim \frac{M_\lambda}{\sqrt{\underline{\mathrm{m}}}}, \tag{C.9}$$

where $\underline{\mathrm{m}} = \min\{m_k : 1 \leq k \leq K\}$. The same reasoning shows that

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} |W_\lambda(\hat{\mu}_\theta, \mu_\theta) - W_\lambda(\mu_\theta, \mu_\theta)|\right] \lesssim \frac{M_\lambda}{\sqrt{\underline{\mathrm{m}}}}. \tag{C.10}$$

Remembering inequality (C.8), we finally get

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} \Big| S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\hat{\mu}_\theta, \nu) \Big|\right] \lesssim \frac{2M_\lambda}{\sqrt{\underline{\mathrm{m}}}}, \tag{C.11}$$

which is the second point of Proposition A.7. □

C.3 Proof of Lemma A.6

Proof. We first write

$$W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\theta^*}, \nu) = W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\hat{\theta}_\lambda^{S}}, \nu) + W_0(\mu_{\hat{\theta}_\lambda^{S}}, \nu) - W_0(\mu_{\theta^*}, \nu).$$

The second term being controlled with lemma A.4, we can focus on the first term $W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\hat{\theta}_\lambda^{S}}, \nu)$ that we upper bound as follows

$$W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\hat{\theta}_\lambda^{S}}, \nu) \leq 2M_I\lambda^2 + S_\lambda(\mu_{\hat{\theta}_\lambda^{(\ell)}}, \nu) - S_\lambda(\mu_{\hat{\theta}_\lambda^{S}}, \nu),$$

thanks to the approximation error of the Sinkhorn divergence established in Corollary A.2. Then,

$$\begin{aligned}
S_\lambda(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - S_\lambda(\mu_{\hat{\theta}_\lambda^{S}}, \nu) &\leq S_\lambda(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) + S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) - S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) \\
&\quad + S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) - S_\lambda(\mu_{\hat{\theta}_\lambda^{S}}, \nu) \\
&\leq S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) - S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) + 2\sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)|.
\end{aligned}$$

Finally, we introduce the Sinkhorn algorithm output after ℓ iterations that approximate the Sinkhorn divergence by writing

$$\begin{aligned}
S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) - S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) &= S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) - S_\lambda^{(\ell)}(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) + \underbrace{S_\lambda^{(\ell)}(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) - S_\lambda^{(\ell)}(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu})}_{\leq 0} \\
&\quad + S_\lambda^{(\ell)}(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) - S_\lambda(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) \\
&\leq \frac{4\|c\|_\infty^2}{\lambda\ell}.
\end{aligned}$$

Indeed, $S_\lambda^{(\ell)}(\hat{\mu}_{\hat{\theta}_\lambda^{S(\ell)}}, \hat{\nu}) - S_\lambda^{(\ell)}(\hat{\mu}_{\hat{\theta}_\lambda^{S}}, \hat{\nu}) \leq 0$ due to the definition of $\hat{\theta}_\lambda^{S(\ell)} \in \arg\min_{\theta \in \Sigma_K} S_\lambda^{(\ell)}(\hat{\mu}_\theta, \hat{\nu})$. We also made use of Proposition A.13, to bound the first and last difference. Thus gathering the previous inequalities, we get

$$W_0(\mu_{\hat{\theta}_\lambda^{S(\ell)}}, \nu) - W_0(\mu_{\hat{\theta}_\lambda^{S}}, \nu) \leq 2M_I\lambda^2 + 2\sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)| + \frac{4\|c\|_\infty^2}{\lambda\ell}.$$

And Lemma A.4 gives that:

$$W_0(\mu_{\hat{\theta}^S_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 2 \sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)| + 2M_I \lambda^2.$$

Putting all the pieces together, we finally get

$$W_0(\mu_{\hat{\theta}^{S(\ell)}_\lambda}, \nu) - W_0(\mu_{\theta^*}, \nu) \leq 4M_I\lambda^2 + \frac{4\|c\|_\infty^2}{\lambda \ell} + 4 \sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)|, \tag{C.12}$$

which is the inequality claimed at equation (A.68). Using this previous inequality (C.12), with the bounds on the estimation error of the Sinkhorn divergence established in Proposition A.7, we bound $\mathbb{E}\left[\sup_{\theta \in \Sigma_K} |S_\lambda(\hat{\mu}_\theta, \hat{\nu}) - S_\lambda(\mu_\theta, \nu)|\right]$. Next, we exploit assumption A.1 to get $\|c\|_\infty^2 \leq 16R^4$ and the upper bound claimed in equation (A.69) for the expected excess risk follows, which proves point (ii) of Lemma A.6.

For the last point of Lemma A.6 we begin as in Proposition A.8, that is, we set $\lambda_n = n^{-1/(2\lfloor d/2 \rfloor + 4)}$ to balance the estimation error and the approximation error. With this choice of regularization parameter the expected excess risk reads

$$\mathbb{E}\left[W_0(\mu_{\hat{\theta}^{(\ell)}_{\lambda_n}}, \nu) - W_0(\mu_{\theta^*}, \nu)\right] \lesssim n^{-2/(2\lfloor d/2 \rfloor + 4)} + \frac{64R^4}{n^{-1/(2\lfloor d/2 \rfloor + 4)}\ell}.$$

The issue is now to bring the algorithm error $\frac{64R^4}{n^{-1/(2\lfloor d/2 \rfloor + 4)}\ell}$ below the rate of $n^{-2/(2\lfloor d/2 \rfloor + 4)}$. To do so, we set $\ell_n \geq 64R^4 n^{3/(2\lfloor d/2 \rfloor + 4)}$. And then, we recover the rate of convergence claimed in equation (A.70) for $\hat{\theta}^{(\ell_n)}_{\lambda_n}$.

□

D Proof of Proposition A.9

This section is devoted to the control of empirical processes that appear in Proposition A.9. We exploit two bounds established in Chizat et al. [2020], which are recalled in the next lemma (by symmetry, we can write it for concave functions instead of convex functions).

Lemma D.1. *[Chizat et al., 2020, Lemma 4 and proof of Theorem 2] Assume that Assumption A.1 holds and that n samples for ν are available. Then it holds that*

$$\mathbb{E}\left[\sup_{\varphi \in \mathcal{F}_R} \left| \int \varphi d(\nu - \hat{\nu}) \right| \right] \lesssim \begin{cases} R^2 n^{-1/2} & \text{if} \quad d < 4, \\ R^2 n^{-1/2}\log(n) & \text{if} \quad d = 4, \\ R^2 n^{-2/d} & \text{if} \quad d > 4. \end{cases} \tag{D.1}$$

where $\lesssim$ hides a constant that depends only on d and $\mathcal{F}_R$ denotes the class of concave and R-Lipschitz functions on $B(0, R)$. In the same paper, the authors established that

$$\mathbb{E}\left[\left| \int_{\mathcal{Y}} \|y\|^2 d(\nu - \hat{\nu})(y) \right|\right] \leq 4R^2 n^{-1/2}. \tag{D.2}$$

We now prove Proposition A.9.

Proof. The key point is to exploit the alternative dual formulation of regularized OT that has been introduced in Section 2.2. Using relation (2.12), we remark that for any $\theta \in \Sigma_K$,

$$\begin{aligned} W_\lambda(\mu_\theta, \nu) - W_\lambda(\mu_\theta, \hat{\nu}) &= \int_{\mathcal{Y}} \|y\|^2 d\nu(y) - \int_{\mathcal{Y}} \|y\|^2 d\hat{\nu}(y) + W_\lambda^s(\mu_\theta, \nu) - W_\lambda^s(\mu_\theta, \hat{\nu}) \\ &= \int_{\mathcal{Y}} \|y\|^2 d(\nu - \hat{\nu})(y) + W_\lambda^s(\mu_\theta, \nu) - W_\lambda^s(\mu_\theta, \hat{\nu}). \end{aligned} \tag{D.3}$$

Now, let us denote by φ and $\hat{\varphi}$ two optimal dual potentials respectively associated to $W_\lambda^s(\mu_\theta,\nu)$ and $W_\lambda^s(\mu_\theta,\hat{\nu})$ when exploiting the semi-dual formulation (2.15). We can thus write

$$\begin{aligned}
W_\lambda^s(\mu_\theta,\nu) - W_\lambda^s(\mu_\theta,\hat{\nu}) &= \int \varphi(x) d\mu_\theta(x) + \int \varphi^s(y) d\nu(y) - \left(\int \hat{\varphi}(x) d\mu_\theta(x) + \int \hat{\varphi}^s(y) d\hat{\nu}(y)\right) \\
&= \int \varphi^s(y) d\nu(y) - \int \varphi^s(y) d\hat{\nu}(y) \\
&+ \underbrace{\int \varphi(x) d\mu_\theta(x) + \int \varphi^s(y) d\hat{\nu}(y) - \left(\int \hat{\varphi}(x) d\mu_\theta(x) + \int \hat{\varphi}^s(y) d\hat{\nu}(y)\right)}_{\leq 0} \\
&\leq \int \varphi^s(y) d(\nu - \hat{\nu})(y),
\end{aligned}$$

where the last inequality derives from the optimality of $\hat{\varphi}$ for the semi-dual formulation of $W_\lambda^s(\mu_\theta,\hat{\nu})$. A similar reasoning yields

$$W_\lambda^s(\mu_\theta,\hat{\nu}) - W_\lambda^s(\mu_\theta,\nu) \leq \int \hat{\varphi}^s(y) d(\hat{\nu} - \nu)(y)$$

As φ^s and $\hat{\varphi}^s$ are both s-transform, Proposition 2.2 ensures that both φ^s and $\hat{\varphi}$ gives the upper bound

$$|W_\lambda^s(\mu_\theta,\nu) - W_\lambda^s(\mu_\theta,\hat{\nu})| \leq \sup_{\varphi \in \mathcal{F}_R} \left| \int \varphi d(\nu - \hat{\nu}) \right|, \tag{D.4}$$

where $\mathcal{F}_R$ denotes the class of concave and R-Lipschitz functions on $B(0,R)$. The part (i) of Proposition A.9 then follows from Lemma D.1. The part (ii) of Proposition A.9 can be obtained with a similar reasoning, by decomposing the mixture as in the beginning of Section B.2, one can see that

$$\begin{aligned}
|W_\lambda(\mu_\theta,\hat{\nu}) - W_\lambda(\hat{\mu}_\theta,\hat{\nu})| &\leq \sup_{\varphi \in \mathcal{F}_R} \left| \int \varphi d(\mu_\theta - \hat{\mu}_\theta) \right| + \left| \int_{\mathcal{X}} \|x\|^2 d(\hat{\mu}_\theta - \mu_\theta)(y) \right| \\
&\leq \sum_{k=1}^K \theta_k \left(\sup_{\varphi \in \mathcal{F}_R} \left| \int \varphi d(\mu_k - \hat{\mu}_k) \right| + \left| \int_{\mathcal{X}} \|x\|^2 d(\hat{\mu}_k - \mu_k)(y) \right| \right).
\end{aligned}$$

Therefore, applying Lemma D.1 to the probability distribution μ_k, we obtain

$$\mathbb{E}\left[\sup_{\varphi_k \in \mathcal{F}_R} \left| \int \varphi_k d(\mu_k - \hat{\mu}_k) \right| \right] + \mathbb{E}\left[\left| \int_{\mathcal{Y}} \|y\|^2 d(\mu_k - \hat{\nu}_k)(y) \right| \right] \lesssim \begin{cases} R^2 m_k^{-1/2} & \text{if} \quad d < 4, \\ R^2 m_k^{-1/2} \log(m_k) & \text{if} \quad d = 4, \\ R^2 m_k^{-2/d} & \text{if} \quad d > 4. \end{cases}$$

It follows that for every $\theta \in \Sigma_K$, we have

$$\mathbb{E}\left[\sum_{k=1}^K \theta_k \sup_{\varphi_k \in \mathcal{F}_R} \left| \int \varphi_k d(\mu_k - \hat{\mu}_k) \right| + \left| \int_{\mathcal{Y}} \|y\|^2 d(\mu_k - \hat{\nu}_k)(y) \right| \right] \lesssim \begin{cases} R^2 \underline{\mathrm{m}}^{-1/2} & \text{if} \quad d < 4, \\ R^2 \underline{\mathrm{m}}^{-1/2} \log(\underline{\mathrm{m}}) & \text{if} \quad d = 4, \\ R^2 \underline{\mathrm{m}}^{-2/d} & \text{if} \quad d > 4, \end{cases}$$

where $\underline{\mathrm{m}} = \min(m_1, \ldots, m_K)$. We used the fact that $\sum_{k=1}^K \theta_k = 1$ to get the last inequality. We can now write

$$\mathbb{E}\left[\sup_{\theta \in \Sigma_K} |W_\lambda(\mu_\theta,\hat{\nu}) - W_\lambda(\hat{\mu}_\theta,\hat{\nu})| \right] \lesssim \begin{cases} R^2 \underline{\mathrm{m}}^{-1/2} & \text{if} \quad d < 4, \\ R^2 \underline{\mathrm{m}}^{-1/2} \log(\underline{\mathrm{m}}) & \text{if} \quad d = 4, \\ R^2 \underline{\mathrm{m}}^{-2/d} & \text{if} \quad d > 4, \end{cases}$$

which gives the last inequality of Proposition A.9.

□

E Proof of Lemma A.2

In this section we give a precise bound on the derivatives of a c-transform. Some results of the same flavor had already been established, for instance in Genevay et al. [2019, Lemma 1, Lemma 2]. The specificity of our result is to exploit the particular cost function $c(x,y) = \|x-y\|^2$ to give a precise description of the bound and to ensure that it is independent of the parameter θ. To prove Lemma A.2, we first give a rescaling argument to reduce our study to the case $\lambda = 1$. Then, in Proposition E.1 we give a clear description of the derivatives of a c-transform when $\lambda = 1$. Next, using this last proposition we give a uniform bound on the derivatives of the c-transform. And in section E.3 we prove Lemma A.2.

E.1 Rescaling argument to link the case $\lambda = 1$ to any positive value

We use a rescaling argument given in Mena and Niles-Weed [2019] to link the case $\lambda = 1$ to any value of the regularization parameter λ.

Lemma E.1. *Let us set $\lambda > 0$, and denote by $T_{\lambda\#}\mu$ and $T_{\lambda\#}\nu$, the push-forward measures of μ and ν by the map $T_\lambda : x \mapsto \lambda^{-1/2}x$. We have the following relation*

$$W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu) = \frac{1}{\lambda} W_\lambda(\mu,\nu). \tag{E.1}$$

And, denoting $\eta_\lambda, \rho_\lambda$ optimal potentials with respect to $W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu)$ the functions $\varphi_\lambda, \psi_\lambda$ defined as

$$\varphi_\lambda(x) = \lambda\eta_\lambda(\lambda^{-1/2}x) \quad \text{and} \quad \psi_\lambda(x) = \lambda\rho_\lambda(\lambda^{-1/2}x), \tag{E.2}$$

are two optimal potentials with respect to $W_\lambda(\mu,\nu)$,

In our context, $W_1(\mu,\nu)$ is not the 1-Wasserstein distance but the regularized Wasserstein distance as defined in equation (2.1) with regularization parameter $\lambda = 1$.

Proof. Denote by T_λ^{-1} the function $T_\lambda^{-1} : x \mapsto \lambda^{1/2}x$, and $\mathcal{J}_\lambda^{\mu,\nu} : L^\infty(\mathcal{X}) \times L^\infty(\mathcal{Y}) \to \mathbb{R}$ the objective function of the dual regularized problem (2.2). For $(f,g) \in L^\infty(\mathcal{X}) \times L^\infty(\mathcal{Y})$, we write,

$$\begin{aligned}
\mathcal{J}_\lambda^{\mu,\nu}(f,g) &= \int f d\mu + \int g d\nu - \lambda \int \exp\left(\frac{f(x)+g(y)-\|x-y\|^2}{\lambda}\right) d\mu(x)d\nu(y) + \lambda \\
&= \int f \circ T_\lambda^{-1} \circ T_\lambda d\mu + \int g \circ T_\lambda^{-1} \circ T_\lambda d\nu \\
&- \lambda \int \exp\left(\frac{f \circ T_\lambda^{-1} \circ T_\lambda(x) + g \circ T_\lambda^{-1} \circ T_\lambda(y) - \|T_\lambda^{-1} \circ T_\lambda(x) - T_\lambda^{-1} \circ T_\lambda(y)\|^2}{\lambda}\right) d\mu(x)d\nu(y) + \lambda \\
&= \int f \circ T_\lambda^{-1} dT_{\lambda\#}\mu + \int g \circ T_\lambda^{-1} dT_{\lambda\#}\nu \\
&- \lambda \int \exp\left(\frac{f \circ T_\lambda^{-1}(x) + g \circ T_\lambda^{-1}(y) - \|T_\lambda^{-1}(x) - T_\lambda^{-1}(y)\|^2}{\lambda}\right) d(T_{\lambda\#}\mu \otimes T_{\lambda\#}\nu)(x,y) + \lambda \\
&= \lambda \Bigg(\int \frac{1}{\lambda} f \circ T_\lambda^{-1} dT_{\lambda\#}\mu + \int \frac{1}{\lambda} g \circ T_\lambda^{-1} dT_{\lambda\#}\nu \Bigg) \\
&- \lambda \exp\left(\frac{1}{\lambda} f \circ T_\lambda^{-1}(x) + \frac{1}{\lambda} g \circ T_\lambda^{-1}(y) - \|x-y\|^2\right) d(T_{\lambda\#}\mu \otimes T_{\lambda\#}\nu)(x,y) + \lambda \\
&\leq \lambda \left(\int \eta_\lambda dT_{\lambda\#}\mu + \int \rho_\lambda dT_{\lambda\#}\nu - \int \exp\left(\eta_\lambda(x) + \rho_\lambda(y) - \|x-y\|^2\right) d(T_{\lambda\#}\mu \otimes T_{\lambda\#}\nu)(x,y) + 1 \right) \\
&= \lambda W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu).
\end{aligned}$$

The last inequality comes from the optimality of η_λ and ρ_λ with respect to $W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu)$. So far, we have derived

$$\sup_{f,g} \int_{\mathcal{X}} f d\mu + \int_{\mathcal{Y}} g d\nu - \lambda \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{f(x)+g(y)-\|x-y\|^2}{\lambda}\right) d\mu(x)d\nu(y) + \lambda \leq \lambda W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu).$$

Hence, $W_\lambda(\mu,\nu) \leq \lambda W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu)$. On the other hand, the optimality of $\eta_\lambda, \rho_\lambda$ with respect $W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu)$ allows to write

$$\begin{aligned}
\lambda W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu) &= \lambda \int \eta_\lambda(x) d(T_{\lambda\#}\mu) + \int \rho_\lambda(y) d(T_{\lambda\#}\nu) \\
&\quad - \lambda \int \exp\left(\eta_\lambda(x) + \rho_\lambda(y) - \|x-y\|^2\right) dT_{\lambda\#}\mu \otimes T_{\lambda\#}\nu + \lambda \\
&= \int \lambda\eta_\lambda(\lambda^{-1/2}x) d\mu + \int \lambda\rho_\lambda(\lambda^{-1/2}y) d\nu \\
&\quad - \lambda \int \exp\left(\eta_\lambda(\lambda^{-1/2}x) + \rho_\lambda(\lambda^{-1/2}y) - \|\lambda^{-1/2}x - \lambda^{-1/2}y\|^2\right) d\mu(x)d\nu(y) + \lambda \\
&= \int \lambda\eta_\lambda(\lambda^{-1/2}x) d\mu + \int \lambda\rho_\lambda(\lambda^{-1/2}y) d\nu \\
&\quad - \lambda \int \exp\left(\frac{\lambda\eta_\lambda(\lambda^{-1/2}x) + \lambda\rho_\lambda(\lambda^{-1/2}y) - \|x-y\|^2}{\lambda}\right) d\mu(x)d\nu(y) + \lambda \\
&= \int \varphi_\lambda(x) d\mu + \int \psi_\lambda(y) d\nu - \lambda \int \exp\left(\frac{\varphi_\lambda(x) + \psi_\lambda(y) - \|x-y\|^2}{\lambda}\right) d\mu(x)d\nu(y) + \lambda,
\end{aligned}$$

with $\varphi_\lambda(x) = \lambda\eta_\lambda(\lambda^{-1/2}x)$ and $\psi_\lambda(x) = \lambda\rho_\lambda(\lambda^{-1/2}x)$. This last computation shows that

$$\lambda W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu) \leq \sup_{f,g} \int_{\mathcal{X}} f d\mu + \int_{\mathcal{Y}} g d\nu - \lambda \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{f(x)+g(y)-\|x-y\|^2}{\lambda}\right) d\mu(x)d\nu(y) + \lambda.$$

Hence $\lambda W_1(T_{\lambda\#}\mu, T_{\lambda\#}\nu) = W_\lambda(\mu,\nu)$ and it shows that φ_λ and ψ_λ are optimal potentials with respect to $W_\lambda(\mu,\nu)$.

□

E.2 Differentiation of a c-transform $\psi = \varphi_\mu^{c,\lambda}$ in the case $\lambda = 1$

In this section, we set $\theta \in \Sigma_K$ and address the specific case $\lambda = 1$. We denote by $\mu = \mu_\theta$ the probability distribution that belongs to our parametric model that is for the moment set.

We precise the notations previously introduced in equation (A.9). For a multi-index $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{N}^d$, we denote by $|\kappa| = \sum_{i=1}^d \kappa_i$ and D^κ the differential operator defined as follows

$$D^\kappa = \frac{\partial^{|\kappa|}}{\partial x_1^{\kappa_1} \ldots \partial x_d^{\kappa_d}}. \tag{E.3}$$

We also denote by $e_i \in \mathbb{N}^d$ the multi index to refer to the partial derivative with respect to x_i. In other terms

$$(e_i)_j = \begin{cases} 1 & \text{if} \qquad j = i \\ 0 & \text{otherwise.} \end{cases} \tag{E.4}$$

For $\mathscr{K} \in \mathbb{N}$, we say that a function $f : \mathcal{Y} \to \mathbb{R}$ belongs to $\mathscr{C}^{\mathscr{K}}(\mathcal{Y})$ if for all $\kappa \in \mathbb{N}^d$ such that $|\kappa| \leq \mathscr{K}$, the function $D^\kappa f : \mathcal{Y} \to \mathbb{R}$ is well defined and continuous on $\mathcal{Y}$. Then, for a bounded function f, we denote by $\|f\|_\infty = \sup_{y\in\mathcal{Y}} |f(y)|$. And for $f \in \mathscr{C}^{\mathscr{K}}(\mathcal{Y})$, we denote by $\|f\|_{\mathscr{K}} = \max_{\kappa \leq \mathscr{K}} \|D^\kappa f\|_\infty$.

Proposition E.1. *Set $\lambda = 1$ and denote by $\psi = \varphi_\mu^{c,\lambda}$ the c-transform of φ. Then, ψ belongs to $\mathscr{C}^\infty$, and for $\mathscr{K} \geq 1$, there exist two finite sequences $(a_l^{\mathscr{K}})_{1 \leq l \leq N_{\mathscr{K}}} \subset \mathbb{N}$ and $(\alpha_{l,n}^{\mathscr{K}})_{1 \leq l \leq N_{\mathscr{K}},\ 1 \leq n \leq \mathscr{K}} \subset \mathbb{N}$ such that for every multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$, there exists a sequence of multi-index $(\sigma_{l,n}^{\kappa})_{1 \leq l \leq N_{\mathscr{K}},\ 1 \leq n \leq \mathscr{K}} \subset \mathbb{N}^d$ such that when we denote by*

$$P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^{\kappa}}(\psi)(y) = \sum_{l=1}^{N_{\mathscr{K}}} a_l^{\mathscr{K}} \left(\prod_{n=1}^{\mathscr{K}} (D^{\sigma_{l,n}^{\kappa}} \psi(y))^{\alpha_{l,n}^{\mathscr{K}}} \right),$$

$$L_\kappa(y) = \left(D^\kappa \psi(y) + P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^{\kappa}}(\psi)(y) \right) e^{-\psi(y)}, \tag{E.5}$$

and

$$R_\kappa(y) = \int_{\mathcal{X}} \left(D^\kappa c(x,y) + P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^{\kappa}}(c)(x,y) \right) e^{\varphi(x) - c(x,y)} d\mu(x), \tag{E.6}$$

the following equality holds

$$\forall y \in B(0,R),\ L_\kappa(y) = R_\kappa(y). \tag{E.7}$$

Additionally, the sequences $(\alpha_{l,n}^{\mathscr{K}})_{1 \leq l \leq N_{\mathscr{K}},\ 1 \leq n \leq \mathscr{K}}$ and $(\sigma_{l,n}^{\kappa})_{1 \leq l \leq N_{\mathscr{K}},\ 1 \leq n \leq \mathscr{K}}$ follow the three conditions,

$$\sum_{n=1}^{\mathscr{K}} |\sigma_{l,n}^{\kappa}| \alpha_{l,n}^{\mathscr{K}} \leq |\kappa| = \mathscr{K},\ \sum_{n=1}^{\mathscr{K}} \alpha_{l,n}^{\mathscr{K}} \leq \mathscr{K},\ \textit{and } |\sigma_{l,n}^{\kappa}| \leq \mathscr{K} - 1. \tag{E.8}$$

Proof. We proceed by induction with the following induction hypothesis: For $\mathscr{K} \geq 1$, there exist two finite sequences $(a_l^{\mathscr{K}})_{1 \leq l \leq N_{\mathscr{K}}} \subset \mathbb{N}$ and $(\alpha_{l,n}^{\mathscr{K}})_{1 \leq l \leq N_{\mathscr{K}},\ 1 \leq n \leq \mathscr{K}} \subset \mathbb{N}$ such that for every multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$ there exists a sequence $(\sigma_{l,n}^{\kappa})_{1 \leq l \leq N_{\mathscr{K}},\ 1 \leq n \leq \mathscr{K}} \subset \mathbb{N}^d$. Such that equations (E.7) and (E.8) hold. We refer to this induction assumption as $(H_{\mathscr{K}})$

Base case: Let us show (H_1). As $\psi = \varphi_\mu^{c,\lambda}$ is the c-transform of φ we have

$$\forall y \in B(0,R), \quad \exp(-\psi(y)) = \int \exp\left(\varphi(x) - c(x,y)\right) d\mu(x). \tag{E.9}$$

A consequence of this relation (see e.g. Feydy et al. [2019]), is that such a potential ψ inherits the regularity properties of the cost function c. The cost function being $c(x,y) = \|x-y\|^2$ and $\mathcal{X}$ being compact, the potential ψ is $\mathscr{C}^\infty(B(0,R))$. Next, setting $\kappa \in \mathbb{N}^d$ such that $|\kappa| = 1$ and applying the operator of differentiation D^κ on both sides of equation (E.9) we get

$$\forall y \in B(0,R), \quad D^\kappa \psi(y) \exp(-\psi(y)) = \int D^\kappa c(x,y) \exp\left(\varphi(x) - c(x,y)\right) d\mu(x). \tag{E.10}$$

In other words,

$$\forall y \in B(0,R), \quad \frac{\partial \psi}{\partial y_i}(y) \exp(-\psi(y)) = \int \frac{\partial c}{\partial y_i}(x,y) \exp\left(\varphi(x) - c(x,y)\right) d\mu(x). \tag{E.11}$$

Where i is the only index such that $\kappa_i > 0$. Equation (E.10) shows that (H_1) is true. In order to ease the understanding of Proposition E.1, we also give the explicit computations in the case $\mathscr{K} = 2$. For $y \in B(0,R)$, differentiating once again on both sides of equation (E.11) with respect to y_j we reach:

$$\left(\frac{\partial^2 \psi}{\partial y_j \partial y_i}(y) - \frac{\partial \psi}{\partial y_i}(y) \frac{\partial \psi}{\partial y_j}(y) \right) \exp(-\psi(y)) = \int \left(\frac{\partial^2 c}{\partial y_j \partial y_i}(x,y) - \frac{\partial c}{\partial y_i}(x,y) \frac{\partial c}{\partial y_j}(x,y) \right) \exp\left(\varphi(x) - c(x,y)\right) d\mu(x). \tag{E.12}$$

With the operator notation E.3, we can rewrite the previous equation as follows. For $y \in B(0,R)$,

$$\left(D^{\sigma_1} \psi(y) - D^{\sigma_2} \psi(y) D^{\sigma_3} \psi(y)\right) \exp(-\psi(y)) = \int \left(D^{\sigma_1} c(x,y) - D^{\sigma_2} c(x,y) D^{\sigma_3} c(x,y)\right) \exp\left(\varphi(x) - c(x,y)\right) d\mu(x). \tag{E.13}$$

Induction step: Set $\mathscr{K} \geq 1$ and assume that $(H_{\mathscr{K}})$ holds true. First, we point out that the symmetry between the left side and the right side of equation (E.7), that one can observe at equation (E.10) and at equation (E.12) allows us to write the computation only for the left side G_κ. We mention that the theorem of differentiation under the integral enables us to derive the right side of equation (E.7) as the subset of integration is compact.

We then set a multi-index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$.
Due to $(H_{\mathscr{K}})$ equation (E.7) holds with $L_\kappa(y) = \left(D^\kappa\psi(y) + P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^\kappa}(\psi)(y)\right)e^{-\psi(y)}$. Hence, differentiating the left side of equation (E.7) with respect to y_i yields

$$\begin{aligned}\frac{\partial L_\kappa}{\partial y_i}(y) = &\left(\frac{\partial D^\kappa\psi}{\partial y_i}(y) + \frac{\partial P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^\kappa}(\psi)}{\partial y_i}(y)\right)e^{-\psi(y)} \\ &- \left(D^\kappa\psi(y)\frac{\partial\psi}{\partial y_i}(y) + P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^\kappa}(\psi)(y)\frac{\partial\psi}{\partial y_i}(y)\right)e^{-\psi(y)}.\end{aligned} \tag{E.14}$$

During this induction step, for a multi index $\sigma \in \mathbb{N}^d$, we denote by $\tilde{\sigma}$ the multi index defined by

$$\tilde{\sigma} = \sigma + e_i,$$

where e_i is defined in equation (E.4). Notice that $\tilde{\sigma} = |\sigma| + 1$. With these notations, we can rewrite equation (E.14) as

$$\begin{aligned}\frac{\partial L_\kappa}{\partial y_i}(y) = &\left(D^{\tilde{\kappa}}\psi(y) + \frac{\partial P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^\kappa}(\psi)}{\partial y_i}(y)\right)e^{-\psi(y)} \\ &- \left(D^\kappa\psi(y)D^{e_i}\psi(y) + P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^\kappa}(\psi)(y)D^{e_i}\psi(y)\right)e^{-\psi(y)}.\end{aligned}$$

Using the formula to derive a product of multiple factors, we get that the second term of equation (E.14) equals

$$\begin{aligned}\frac{\partial P_{a^{\mathscr{K}},\alpha^{\mathscr{K}},\sigma^\kappa}(\psi)}{\partial y_i}(y) &= \sum_{l=1}^{N_{\mathscr{K}}} a_l^{\mathscr{K}} \sum_{n=1}^{\mathscr{K}} \alpha_{l,n}^{\mathscr{K}} \frac{\partial D^{\sigma_{l,n}^\kappa}\psi}{\partial y_i}(y)(D^{\sigma_{l,n}^\kappa}\psi(y)(y))^{\alpha_{l,n}^{\mathscr{K}}-1}\left(\prod_{\substack{j=1,\\ j\neq n}}^{\mathscr{K}} (D^{\sigma_{l,j}^\kappa}\psi(y))^{\alpha_{l,j}^{\mathscr{K}}}\right) \\ &= \sum_{l=1}^{N_{\mathscr{K}}} a_l^{\mathscr{K}} \sum_{n=1}^{\mathscr{K}} \alpha_{l,n}^{\mathscr{K}} D^{\tilde{\sigma}_{l,n}^\kappa}\psi(y)(D^{\sigma_{l,n}^\kappa}\psi(y)(y))^{\alpha_{l,n}^{\mathscr{K}}-1}\left(\prod_{\substack{j=1,\\ j\neq n}}^{\mathscr{K}} (D^{\sigma_{l,j}^\kappa}\psi(y))^{\alpha_{l,j}^{\mathscr{K}}}\right) \\ &= \sum_{\substack{1\leq l\leq N_{\mathscr{K}},\\ 1\leq n\leq \mathscr{K}}} a_l^{\mathscr{K}}\,\alpha_{l,n}^{\mathscr{K}} \prod_{j=1}^{\mathscr{K}+1} (D^{\tau_{l,n,j}}\psi(y))^{\beta_{l,n,j}}.\end{aligned}$$

Let us set $(l,n) \in \{1,\ldots,N_\kappa\}\times\{1,\ldots,\mathscr{K}\}$, and study the sequences $(\tau_{l,n,j})_{1\leq j\leq \mathscr{K}+1}$ and $(\beta_{l,n,j})_{1\leq j\leq \mathscr{K}+1}$ that we have introduced in the last equality of the previous calculation.
We have:

$$\tau_{l,n,j} = \begin{cases} \tilde{\sigma}_{l,n}^\kappa = \sigma_{l,n}^\kappa + e_i & \text{if} \qquad j = \mathscr{K}+1, \\ \sigma_{l,j}^\kappa & \text{otherwise.} \end{cases}$$

and

$$\beta_{l,n,j} = \begin{cases} 1 & \text{if} \qquad j = \mathscr{K}+1, \\ \alpha_{l,n}^{\mathscr{K}} - 1 & \text{if} \qquad j = n, \\ \alpha_{l,j}^{\mathscr{K}} & \text{otherwise.} \end{cases}$$

After pointing out that for all (l, n) the sequence $(\beta_{l,n,j})_{1\leq j\leq \mathscr{K}+1}$ is independent of the multi index κ and y_i, we check that the sequences $(\tau_{l,n})$ and $(\beta_{l,n})$ satisfy conditions (E.8).

$$\begin{aligned}
\sum_{j=1}^{\mathscr{K}+1} |\tau_{l,n,j}|\beta_{l,n,j} &= |\tilde{\sigma}^{\kappa}_{l,n}| \times 1 + \sum_{\substack{j=1,\\ j\neq n}}^{\mathscr{K}} |\sigma^{\kappa}_{l,j}|\alpha^{\mathscr{K}}_{l,j} + |\sigma^{\kappa}_{l,n}|(\alpha^{\mathscr{K}}_{l,n} - 1) \\
&= |\sigma^{\kappa}_{l,n}| + 1 + \sum_{\substack{j=1,\\ j\neq n}}^{\mathscr{K}} |\sigma^{\kappa}_{l,j}|\alpha^{\mathscr{K}}_{l,j} + |\sigma^{\kappa}_{l,n}|\alpha^{\mathscr{K}}_{l,n} - |\sigma^{\kappa}_{l,n}| \\
&= \sum_{j=1}^{\mathscr{K}} |\sigma^{\kappa}_{l,j}|\alpha^{\mathscr{K}}_{l,j} + 1 \\
&\leq \mathscr{K} + 1 \quad \text{thanks to } H_{\mathscr{K}}.
\end{aligned}$$

A similar computation shows that $\sum_{j=1}^{\mathscr{K}+1} |\beta_{l,n,j}| \leq \mathscr{K}$. Hence, the sequences $(\tau_{l,n})$ and $(\beta_{l,n})$ satisfy conditions (E.8). Then, the third term $D^{\kappa}(\psi)(y)D^{e_i}(\psi)(y)$ of equation (E.14) clearly reads

$$\prod_{j=1}^{\mathscr{K}+1} (D^{\eta_j}\psi(y))^{\gamma_j} \quad \text{with} \quad \sum_{j=1}^{\mathscr{K}+1} |\eta_j|\gamma_j = 1 \times 1 + \mathscr{K} \times 1 \leq \mathscr{K} + 1.$$

And the sequences (η) and (γ) also satisfy the other conditions of equation (E.8). The last term of equation (E.14) can be written as

$$\sum_{l=1}^{N_{\mathscr{K}}} a^{\mathscr{K}}_l \left(\prod_{n=1}^{\mathscr{K}+1} (D^{\xi_{l,n}}(\psi)(y))^{\delta_{l,n}} \right),$$

with

$$\xi_{l,n} = \begin{cases} e_i & \text{if} \qquad n = \mathscr{K} + 1, \\ \sigma^{\kappa}_{l,n} & \text{otherwise.} \end{cases}$$

And

$$\delta_{l,n} = \begin{cases} 1 & \text{if} \qquad n = \mathscr{K} + 1, \\ \alpha^{\mathscr{K}}_{l,n} & \text{otherwise.} \end{cases}$$

Hence, for $l \in \{1, \ldots, N_{\mathscr{K}}\}$,

$$\sum_{n=1}^{\mathscr{K}+1} |\xi_{l,n}|\delta_{l,n} = \sum_{n=1}^{\mathscr{K}} |\sigma^{\kappa}_{l,n}|\alpha^{\mathscr{K}}_{l,n} + |e_i| \times 1 = \sum_{n=1}^{\mathscr{K}} |\sigma^{\kappa}_{l,n}|\alpha^{\mathscr{K}}_{l,n} + 1 \leq \mathscr{K} + 1 \quad \text{thanks to } (H_{\mathscr{K}}).$$

Similar computations shows that the others conditions (E.8) are also true. As the choice of the multi-index κ such that $|\kappa| = \mathscr{K}$, and the choice to differentiate with respect to y_i are arbitrary. Hence, the computation of the terms of equation (E.14) shows that there exist two finite sequences $(a^{\mathscr{K}+1}_l)_{1\leq l\leq N_{\mathscr{K}+1}} \subset \mathbb{N}$ and $(\alpha^{\mathscr{K}+1}_{l,n})_{1\leq l\leq N_{\mathscr{K}+1}, 1\leq n\leq \mathscr{K}+1} \subset \mathbb{N}$ such that for every multi-index $\tilde{\kappa} \in \mathbb{N}^d$ such that $|\tilde{\kappa}| = \mathscr{K} + 1$ there exists a sequence of multi-index $(\sigma^{\tilde{\kappa}}_{l,n})_{1\leq l\leq N_{\mathscr{K}+1}, 1\leq n\leq \mathscr{K}+1} \subset \mathbb{N}^d$ such that

$$L_{\tilde{\kappa}}(y) = (D^{\tilde{\kappa}}(\psi)(y) + P_{a^{\mathscr{K}+1}, \alpha^{\mathscr{K}+1}, \sigma^{\tilde{\kappa}}}(\psi)(y))e^{\psi(y)},$$

with the sequences $(\alpha^{\mathscr{K}+1}_{l,n})_{1\leq l\leq N_{\mathscr{K}+1}, 1\leq n\leq \mathscr{K}+1} \subset \mathbb{N}$ and $(\tilde{\sigma}^{\kappa}_{l,n})_{1\leq l\leq N_{\mathscr{K}+1}, 1\leq n\leq \mathscr{K}+1} \subset \mathbb{N}^d$ that satisfy the conditions (E.8).

With the exact same computations with $c(x,.)$ instead of ψ, and differentiating with respect to y_i under the integral (that is possible as $\mathcal{X}$ is compact), one can show that the derivative of the right hand side of equation (E.7) reads

$$R_{\tilde{\kappa}}(y) = \int_{\mathcal{X}} \left(D^{\tilde{\kappa}}(c)(x,y) + P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^{\tilde{\kappa}}}(c)(x,y)\right) e^{\varphi(x)-c(x,y)} d\mu(x),$$

which proves $(H_{\mathscr{K}+1})$.

□

In what follows, we denote by

$$\gamma(x,y) = \exp(\varphi(x) + \psi(y) - c(x,y)), \tag{E.15}$$

where c is still the squared euclidean cost, and φ, ψ are defined as in Proposition E.1. In particular ψ is the c-transform of φ. In the next proposition, we make sure that we can uniformly bound ψ and its derivatives with a constant that **does not** depend of the parameter θ.

Proposition E.2. *Assume that Assumption A.1 holds true. Denote by $\psi = \varphi_\mu^{c,\lambda}$ the c-transform of a certain function $\varphi \in L^\infty(\mathcal{X})$ with regularization parameter $\lambda = 1$. Then, for $\mathscr{K} \geq 1$, there exists a constant $M_{\mathscr{K}}$ that depends only on $\mathscr{K}$, such that for every multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$ we have that*

$$\|D^\kappa \psi\|_\infty \leq M_{\mathscr{K}} R^{\mathscr{K}}.$$

Proof. $\psi = \varphi_\mu^{c,\lambda}$ the c-transform of a certain $\varphi \in L^\infty(\mathcal{X})$. We proceed by induction to show that for $\mathscr{K} \geq 1$ there exists $M_{\mathscr{K}}$ that depends only on $\mathscr{K}$ such that for every multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$, we have $\|D^\kappa\psi\|_\infty \leq M_{\mathscr{K}} R^{\mathscr{K}}$ $(H_{\mathscr{K}})$.

Base case: Set $\mathscr{K} = 1$. Using equation (E.10), for $i \in \{1, \dots, d\}$ we can write that for $y \in \mathcal{Y}$,

$$D^{e_i}(\psi)(y) = \int D^{e_i}(c)(x,y)\gamma(x,y)d\mu(x),$$

where e_i is defined in equation (E.4). As c is the squared Euclidean cost, we have that $\|D^{e_i}(c)\|_\infty \leq 4R$. Thus

$$|D^{e_i}(\psi)(y)| \leq 4R \int \gamma(x,y)d\mu(x).$$

And using equation (E.9) we have that $\int \gamma(x,y)d\mu(x) = 1$. Thus $\|D^{e_i}(\psi)\|_\infty \leq 4R$, which proves that (H_1) is true.

Induction step: set $\mathscr{K} \geq 1$ and assume $(H_1), \dots, (H_{\mathscr{K}})$. Set a multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K} + 1$. For $y \in \mathcal{Y}$, Proposition E.1 allows us to write

$$D^\kappa\psi(y) = P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^\kappa}(\psi)(y) + \int D^\kappa c(x,y) + P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^\kappa}(c)(x,y)\gamma(x,y)d\mu(x).$$

Thus,

$$\begin{aligned} |D^\kappa(\psi)(y)| \leq & |P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^\kappa}(\psi)(y)| \\ & + \int |D^\kappa c(x,y) + P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^\kappa}(c)(x,y)\gamma(x,y)|d\mu(x). \end{aligned} \tag{E.16}$$

For the first term of equation (E.16), using Proposition E.1 we can write

$$\begin{aligned}|P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^\kappa}(\psi)(y)| &\leq \sum_{l=1}^{N_{\mathscr{K}+1}} a_l^{\mathscr{K}+1}\left(\prod_{n=1}^{\mathscr{K}+1}|D^{\sigma^\kappa_{l,n}}(\psi)(y)|^{\alpha^{\mathscr{K}+1}_{l,n}}\right)\\ &\leq \sum_{l=1}^{N_{\mathscr{K}+1}} a_l^{\mathscr{K}+1}\left(\prod_{n=1}^{\mathscr{K}+1}\left(M_{|\sigma^\kappa_{l,n}|}R^{|\sigma^\kappa_{l,n}|}\right)^{\alpha^{\mathscr{K}+1}_{l,n}}\right)\\ &\leq \sum_{l=1}^{N_{\mathscr{K}+1}} a_l^{\mathscr{K}+1}\left(\prod_{n=1}^{\mathscr{K}+1}(M_{\mathscr{K}})^{\alpha^{\mathscr{K}+1}_{l,n}}\right)R^{\sum_{n=1}^{\mathscr{K}+1}|\sigma^\kappa_{l,n}|\alpha^{\mathscr{K}+1}_{l,n}}.\end{aligned}$$

We exploited $(H_1),\dots,(H_{\mathscr{K}})$ to derive second inequality. Then, we exploited the conditions on the $|\sigma^\kappa_{l,n}|$'s and the $\alpha^\kappa_{l,n}$'s from Proposition E.1 that are $|\sigma^\kappa_{l,n}|\leq\mathscr{K}$, and $\sum_{n=1}^{\mathscr{K}+1}|\sigma^\kappa_{l,n}|\alpha^\kappa_{l,n}\leq\mathscr{K}+1$. Hence, there exists a constant $M^{(1)}_{\mathscr{K}+1}$ that depends only on $\mathscr{K}+1$ such that

$$|P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1}}(\psi^{(1)}(y),\dots,\psi^{(\mathscr{K}))}(y))|\leq M^{(1)}_{\mathscr{K}+1}R^{\mathscr{K}+1}. \tag{E.17}$$

Regarding the second term of equation (E.16), remind that for $c(x,y)=\|x-y\|^2$, we have that $\forall\kappa\in\mathbb{N}^d$ such that $|\kappa|\geq 1$, $\|D^\kappa(c)\|_\infty\leq 4R$. Thus, for every $x\in\mathcal{X}$, using Proposition E.1, we have

$$\begin{aligned}|D^\kappa(c)(x,y)+P_{a^{\mathscr{K}+1},\alpha^{\mathscr{K}+1},\sigma^\kappa}(c)(x,y)| &\leq 4R+\sum_{l=1}^{N_{\mathscr{K}+1}}a_l^{\mathscr{K}+1}\left(\prod_{n=1}^{\mathscr{K}+1}(4R)^{\alpha^{\mathscr{K}+1}_{l,n}}\right) &\text{(E.18)}\\ &\leq M^{(2)}_{\mathscr{K}+1}R^{\sum_{n=1}^{\mathscr{K}+1}\alpha^{\mathscr{K}+1}_{l,n}} &\text{(E.19)}\\ &\leq M^{(2)}_{\mathscr{K}+1}R^{\mathscr{K}+1}, &\text{(E.20)}\end{aligned}$$

using condition (E.8) from proposition E.1 to get the last inequality. We can thus upper bound $|D^\kappa(\psi)(y)|$ as follows.

$$\begin{aligned}|D^\kappa(\psi)(y)| &\leq M^{(1)}_{\mathscr{K}+1}R^{\mathscr{K}+1}+M^{(2)}_{\mathscr{K}+1}R^{\mathscr{K}+1}\int_{\mathcal{X}}\gamma(x,y)d\mu(x)\\ &\leq M^{(1)}_{\mathscr{K}+1}R^{\mathscr{K}+1}+M^{(2)}_{\mathscr{K}+1}R^{\mathscr{K}+1}\underbrace{\int_{\mathcal{X}}\gamma(x,y)d\mu(x)}_{=1}\\ &\leq M_{\mathscr{K}+1}R^{\mathscr{K}+1}.\end{aligned}$$

Hence $\|D^\kappa(\psi)\|_\infty\leq M_{\mathscr{K}+1}R^{\mathscr{K}+1}$, which proves $(H_{\mathscr{K}+1})$.

□

E.3 Conclusion of the proof of Lemma A.2

Set $\lambda>0$. Exploiting Proposition 2 in Feydy et al. [2019], we have that there exists two optimal potentials (η,ρ) with respect to $W_1(T_{\lambda\#}\mu,T_{\lambda\#}\nu)$ such that $\rho=\eta^{c,\lambda}_{T_{\lambda\#}\mu}$. Moreover, we can choose $\rho(0)=0$ as the potentials are defined up to an additive constant. Next, using Lemma E.1, we have that $\psi(y) = \lambda\rho(\lambda^{-1/2}y)$ and $\varphi(x) = \lambda\eta(\lambda^{-1/2}x)$. We can thus write,

$$\begin{aligned}
\psi(y) &= \lambda\rho(\lambda^{-1/2}y) \\
&= -\lambda \log\left(\int \exp\left(\eta(x) - \|x - \lambda^{-1/2}y\|^2\right) dT_{\lambda\#}\mu(x)\right) \\
&= -\lambda \log\left(\int \exp\left(\eta(\lambda^{-1/2}x) - \|\lambda^{-1/2}x - \lambda^{-1/2}y\|^2\right) d\mu(x)\right) \\
&= -\lambda \log\left(\int \exp\left(\eta(\lambda^{-1/2}x) - \|\lambda^{-1/2}x - \lambda^{-1/2}y\|^2\right) d\mu(x)\right) \\
&= -\lambda \log\left(\int \exp\left(\frac{\varphi(x) - \|x - y\|^2}{\lambda}\right) d\mu(x)\right).
\end{aligned}$$

Hence, $\psi = \varphi_\mu^{c,\lambda}$. The application of Proposition 12 from Feydy et al. [2019] ensures that a c-transform inherits the Lipschitz constant of the cost function. The cost function being $c(x, y) = ||x - y||^2$ with $x, y \in B(0, R)$, we have that ψ is $4R$-Lipschitz. And as $\psi(0) = \rho(0) = 0$, we can write,

$$\forall y \in B(0, R), \ \|\psi(y)\| \leq 4R||y||.$$

Hence $\|\psi\|_\infty \leq 4R^2$. We now handle the derivatives of ψ. To do so, we set an arbitrary $\mathscr{K} \geq 1$ and an arbitrary multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$, using relation $\psi(y) = \lambda\rho(\lambda^{-1/2}y)$ we have

$$\forall y \in \mathcal{Y}, \ |D^\kappa(\psi_\lambda)(y)| = \lambda^{1-\frac{\mathscr{K}}{2}}|D^\kappa(\rho_\lambda)(\lambda^{-1/2}y)|. \tag{E.21}$$

Next, using the fact that ρ is a c-transform with regularization parameter $\lambda = 1$ and distribution $T_{\lambda\#_\mu}\mu$ whose support is subset of $B(0, \frac{R}{\lambda})$, we are under the assumptions of Proposition E.2. We can thus write

$$\|D^\kappa(\psi_\lambda)(y)\| \leq \lambda^{1-\frac{\mathscr{K}}{2}} M_{\mathscr{K}} \left(\frac{R}{\sqrt{\lambda}}\right)^{\mathscr{K}}.$$

Finally, we get that for every multi index $\kappa \in \mathbb{N}^d$ with $|\kappa| = \mathscr{K}$ there exists a constant $M_{\mathscr{K}}$ that depends only on $\mathscr{K}$ such that:

$$\|D^\kappa(\psi_\lambda)\|_\infty \leq M_{\mathscr{K}} \lambda^{1-\mathscr{K}} R^{\mathscr{K}}. \tag{E.22}$$

Hence, we can finally write that

$$\|\psi_\lambda\|_{\mathscr{K}} \leq M_{\mathscr{K}} \max\left(R^2, \lambda^{1-\mathscr{K}} R^{\mathscr{K}}\right). \tag{E.23}$$

As $\psi(0) = 0$ and (φ, ψ) are optimal potentials with respect to $W_\lambda(\mu, \nu)$, it concludes the proof of Lemma A.2.