Sample Complexity for Regularized Optimal Transport

Valeria Ambrosio, Bjarne Bergh, Tobias Freidling and Lauritz Streck
Supervised by Marcello Carioni and Carola-Bibiane Schönlieb

DPMMS and DAMPT, University of Cambridge

1 Introduction

The problem of transporting people, masses, resources etc. in an optimal way is an intuitive mathematical question and naturally arises in many applications. Consequently, it is not surprising that its first description dates back as far as Monge (1781). Still mainly concerned with practical issues, researchers in the 1930s and 1940s, foremost Kantorovich (1942), lay the foundations for the theory of optimal transport. Since mathematicians discovered the deep connections to partial differential equations, statistics and optimization, e.g. Brenier (1991), optimal transport became an active field of research and flourished especially due to its application in machine learning and data science. In particular, the Wasserstein metric $W(\alpha, \beta)$ is a popular tool to describe the distance between two distributions α and β as it only requires a minimum of assumptions, for example the support of α and β can differ.

The computation of the Wasserstein metric is involved as the underlying optimization problem is not convex and the convergence speed suffers from the curse of dimensionality. Many applications, however, showed that adding an entropic regularization term is a successful strategy to circumvent these problems. Genevay, Chizat, et al. (2019) provided a rigorous treatment of this phenomenon and established a bound on the sample complexity that decays as $\mathcal{O}(n^{-1/2})$. Meanwhile, Marino and Gerolin (2020) generalized the existing theory around the entropy-regularized optimal transport problem to a class of convex regularization functions. Yet, results on the sample complexity for this more general problem are still a blind spot. Our work shows that it is indeed possible to achieve sample complexities of order $\mathcal{O}(n^{-1/2})$, given that suitable regularity conditions are fulfilled.

Section 2 introduces the optimal transport problem and its regularized version and Sections 4 and 5 generalize Genevay, Chizat, et al.'s sample complexity result to a more general class of functions. In Section 6 we summarize algorithms to compute Sinkhorn divergences. Finally, we conduct experiments that confirm our theoretical findings in Section 7.

2 Optimal Transport

We base our description of optimal transport on Peyré and Cuturi (2019). In its original formulation, Monge considered the optimal transport problem as determining a permutation that assigns the source points $x_1, \ldots, x_m$ to target points $y_1, \ldots, y_n$ such that a cost function c is minimized. This so-called Monge problem can be extended to the case of transporting an arbitrary mass to another mass, represented by two probability measures, by expressing it in terms of finding a cost-minimal map. However, both the former and the latter problem are hard to solve as they are combinatorial and non-convex, respectively.

For this reason, Kantorovich (1942) proposed the following relaxed version.

Definition 2.1 (Discrete Optimal Transport Problem)**.** Let a and b be two weight vectors. The set of coupling matrices is given by

$$U(a,b) = \{\mathrm{P} \in \mathbb{R}_+^{n\times m} \colon \mathrm{P}\mathbb{1}_m = a, \mathrm{P}^T\mathbb{1}_n = b\}.$$

Let $\mathrm{C} \in \mathbb{R}^{n\times m}$ be a cost matrix. Then the *discrete optimal transport problem* is

$$\mathrm{OT}_{\mathrm{disc}}(a,b) = \min_{\mathrm{P}\in U(a,b)} \sum_{i,j} \mathrm{P}_{ij}\mathrm{C}_{ij}.$$

This formulation allows, unlike Monge's version, to split up a source mass a_i and transport it to different target masses. Figure 1 illustrates one example of discrete optimal transport. The generalization from discrete mass to continuous probability measures is straightforward, see also Figure 2.

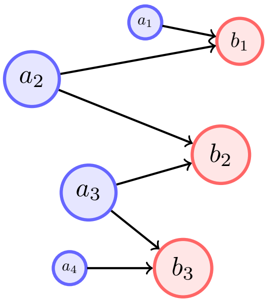


Figure 1: Discrete Optimal Transport

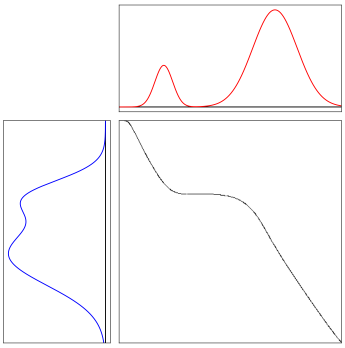


Figure 2: Continuous Optimal Transport

Definition 2.2 (Optimal Transport Problem)**.** Denote the space of positive Radon probability measures on a domain $\mathcal{D}$ by $\mathcal{M}_+^1(\mathcal{D})$. Let α and β be probability measures on the metric spaces X and Y. The space of couplings is given by

$$\Pi(\alpha,\beta) := \{\pi \in \mathcal{M}_+^1(X\times Y) \colon \pi(A\times Y) = \alpha(A), \pi(X\times B) = \beta(B) \ \forall A,B \subset X,Y\}.$$

Let $c\colon X\times Y \to \mathbb{R}$ be a cost function. Then the *optimal transport problem* is

$$\mathrm{OT}(\alpha,\beta) = \min_{\pi\in\Pi(\alpha,\beta)} \int_{X\times Y} c(x,y)\, d\pi(x,y).$$

Definition 2.3 (Wasserstein distance)**.** In the setting of Definition 2.2, assume that $X = Y$ is endowed with a distance d. If $c(x,y) = d(x,y)^p, p \geq 1$, the *$p$-Wasserstein distance* is defined as

$$W_p(\alpha,\beta) = \big(\mathrm{OT}(\alpha,\beta)\big)^{1/p}.$$

The Wasserstein distance becomes increasingly popular in applications, such as machine learning, to compare distributions. In particular, it does not require that the two measures have common support and also allows to compute the distance between a discrete and continuous distribution. However, its computation is expensive as it suffers from the curse of dimensionality.

To avoid this bottleneck, many authors add a regularization term, a technique that was first used by Wilson (1969). This renders the problem strictly convex and decreases computation time considerably, cf. Cuturi (2013). Algorithms like Sinkhorn and variations of it are easy to implement and yield good convergence behaviour which contribute to the increasing use of Wasserstein distances. They are presented in more detail in Section 6.

We define the general regularized optimal transport problem according to Marino and Gerolin (2020).

Definition 2.4 (Regularized Optimal Transport)**.** Let X and Y be two bounded subsets of $\mathbb{R}^d$ and let α and β be probability measures on X and Y, respectively. Let $c\colon X \times Y \to \mathbb{R}$ be a bounded cost function and let $\varepsilon > 0$. Assume that $\Phi\colon [0,\infty] \to [0,\infty]$ is convex, lower semi-continuous and superlinear at infinity, $\Phi \in C^1((0,\infty))$ and $\Phi(1) = \Phi'(1) = 0$. The *optimal transport problem with convex regularization* is given by

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \inf_{\pi\in\Pi(\alpha,\beta)} \int_{X\times Y} c(x,y)\, d\pi(x,y) + \varepsilon\, G(\pi|\alpha\otimes\beta), \tag{1}$$

where G is defined as

$$G(\pi|\alpha\otimes\beta) = \begin{cases} \int_{X\times Y} \Phi\left(\frac{d\pi}{d(\alpha\otimes\beta)}\right) d(\alpha\otimes\beta), & \text{if } \pi \ll \alpha\otimes\beta, \\ +\infty, & \text{otherwise.} \end{cases} \tag{2}$$

Example 2.1 (Entropic regularization)*.* The most common choice for the regularization term is $\Phi(z) = z(\log(z) - 1) + 1$. This yields $G(\pi|\alpha\otimes\beta) = \mathrm{KL}(\pi \,\|\, \alpha\otimes\beta)$ and is consequently called entropic regularization.

Example 2.2 (Tsallis entropy)*.* Let $q > 1$. The Tsallis entropy, cf. Muzellec et al. (2017), is given by

$$\Phi(z) = \frac{1}{q(q-1)}\big(z^q - q(z-1)\big). \tag{3}$$

Example 2.3 (Quadratic regularization)*.*

$$\Phi(z) = \frac{1}{2}|z|^2 \tag{4}$$

Similar to the unregularized case, we can define a distance based on the solution of the regularized optimal transport problem.

Definition 2.5. In the setting of Definition 2.4, we define the Φ-divergence between α and β as

$$\overline{\mathrm{OT}}_\varepsilon = \mathrm{OT}_\varepsilon\big(\alpha,\beta\big) - \frac{1}{2}\big(\mathrm{OT}_\varepsilon(\alpha,\alpha) + \mathrm{OT}_\varepsilon(\beta,\beta)\big).$$

For entropic regularization, we refer to $\overline{\mathrm{OT}}_\varepsilon$ as Sinkhorn-divergence and can prove non-negativity.

The optimization problem (1) has a dual formulation that is a consequence of Fenchel-Rockafellar's Theorem. Genevay, Cuturi, et al. (2016) proved this result for entropic regularization and Marino and Gerolin (2020) contributed the result for general convex regularization functions.

Theorem 2.6 (Dual formulation)**.** *In the setting of Definition 2.4, the Legendre conjugate Ψ of Φ is given by*

$$\Psi(t) = \sup_{s\in[0,\infty]} (st - \Phi(s)), \quad t \in [0,\infty].$$

The dual formulation of the regularized optimal transport problem is

$$\mathrm{D}_\varepsilon(\alpha,\beta) = \sup_{\substack{u\in C(X),\\ v\in C(Y)}} \int_X u d\alpha + \int_Y v d\beta - \varepsilon \int_{X\times Y} \Phi\left(\frac{u(x) + v(y) - c(x,y)}{\varepsilon}\right) d(\alpha\otimes\beta)$$

and $\mathrm{OT}_\varepsilon(\alpha,\beta) = \mathrm{D}_\varepsilon(\alpha,\beta)$.

3 Sample Complexity for the Classical Entropy

When the Sinkhorn algorithm is applied to continuous measures in practice, these measures usually are known only empirically. That is, if we want to solve the regularized optimal transport problem $\mathrm{OT}_\varepsilon(\alpha, \beta)$, the input is only the finite dataset $X_1, \dots, X_n,\ Y_1, \dots, Y_n$ with $\mathrm{law}(X_i) = \alpha$, $\mathrm{law}(Y_i) = \beta$, all independent. The only thing we can solve is then the problem $\mathrm{OT}_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)$, where $\hat{\alpha}_n = \frac{1}{n}\sum_{i=n}^n \delta_{X_i}$, $\hat{\beta}_n = \frac{1}{n}\sum_{i=n}^n \delta_{Y_i}$ are the discrete sample measures. A crucial question is thus what kind of convergence we can be expected on average in the number of samples. That is, one wants to derive quantitative estimates on

$$\mathbb{E}_{X_1,\dots,Y_n}\left|\mathrm{OT}_\varepsilon(\alpha,\beta) - \mathrm{OT}_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)\right|$$

independently of the initial distributions α, β.

In the case of the classical entropy with Legendre conjugate $\Psi(t) = \exp(t)$, Genevay, Chizat, et al. (2019) derive the following result:

Theorem 3.1. *(Genevay, Chizat, et al. (2019) , Theorem 3) Let α, β be probability measures supported in bounded subsets in $X, Y \subset \mathbb{R}^d$. Let $c\colon X \times Y \to \mathbb{R}$ be a smooth, L-Lipschitz cost function. Then*

$$\mathbb{E}_{X_1,\dots,Y_n}\left|\mathrm{OT}_\varepsilon(\alpha,\beta) - \mathrm{OT}_\varepsilon(\hat{\alpha}_n, \hat{\beta}_n)\right| = \mathcal{O}\left(\frac{\Psi\left(\kappa\varepsilon^{-1}\right)}{\sqrt{n}}\left(1 + \varepsilon^{-\lceil\frac{d}{2}\rceil}\right)\right)$$

where $\kappa := \mathrm{diam}(X)L + \|c\|_\infty$ and the implicit constant depends only on $\mathrm{diam}(X), \mathrm{diam}(Y), d$ and $E := \sup_{s=0,\dots,\lceil\frac{d}{2}\rceil}\|c^{(s)}\|_\infty$.

They proved the theorem by showing that the optimizers u, v of $W_\varepsilon(\alpha, \beta)$ have bounded s-Sobolev norm for any s, independent of α and β. This allowed them to apply "standard" PAC-learning results in reproducing kernel Hilbert spaces (RKHS) to derive the bound above. The main step towards the proof of Theorem 3.1 was thus to establish the following proposition.

Proposition 3.2. *Let α, β be probability measures supported in bounded subsets in $X, Y \subset \mathbb{R}^d$. Let $c\colon X \times Y \to \mathbb{R}$ be a smooth, L-Lipschitz cost function. Let u, v be optimizers of $\mathrm{OT}_\varepsilon(\alpha, \beta)$. Then u is in $\mathcal{C}^\infty$ and L-Lipschitz, $u(x) \in \mathrm{ran}[c(x,\cdot) - v(\cdot)]$ and*

$$\|u^{(s)}\|_\infty = \mathcal{O}\left(1 + \varepsilon^{-(s-1)}\right)$$

for all s, with the constant depending only on the first s derivatives of c and on X.

The main ingredient in the proof of this proposition is the equation

$$\exp(-u(x)) = \int \exp\left(\frac{v(y) - c(x,y)}{\varepsilon}\right)\, d\beta \tag{5}$$

derived from the optimality condition.

In the next part we will derive the analogue of Equation (5) for general Ψ and prove a version of Proposition 3.2 under certain conditions on Ψ. We then copy the reasoning of Genevay et al. to derive our main result, Theorem 5.4, which is an analogue of Theorem 3.1 for more general regularization functions.

4 Regularity of the Maximizers

Recall that we aim to study the maximizers of the dual problem

$$\mathrm{D}_\varepsilon(u,v)=\int u\,d\alpha+\int v\,d\beta-\varepsilon\iint\Psi\big((u(x)+v(y)-c(x,y))/\varepsilon\big)\,d\alpha(x)d\beta(y),$$

where Ψ is the Legendre transform of Φ. Existence of the maximizers was shown in Marino and Gerolin (2020).

We will start by finding the correct analogue of Equation (5). For this, we study a pair of maximizers (u,v) through the maximizing condition. For any $\varphi\in C(X)$, the function $g:t\mapsto \mathrm{D}_\varepsilon(u+t\varphi,v)$ has a local maximum at 0. Thus

$$0=\frac{dg}{dt}(0)=\int\varphi d\alpha-\int\varphi(x)\int\Psi'\big((u(x)+v(y)-c(x,y))/\varepsilon\big)\,d\alpha(x)d\beta(y)$$

and as φ is arbitrary,

$$1=\int\Psi'\big((u(x)+v(y)-c(x,y))/\varepsilon\big)\,d\beta(y). \tag{6}$$

We now prove a result analogous to Proposition 3.2. This will be the heart of the theoretical part in this document.

Proposition 4.1. *Let Ψ be the Legendre conjugate of the regularization function Φ. Let α,β be probability measures supported in bounded subsets in $X,Y\subset\mathbb{R}^d$. Let $c\colon X\times Y\to\mathbb{R}$ be a smooth, L-Lipschitz cost function. Set $K:=[-2(L|X|+\|c\|_\infty),2(L|X|+\|c\|_\infty)]$. As Ψ is convex, $\Psi''\geq 0$. Assume that Ψ'' is bounded from below on $\frac{1}{\varepsilon}K$ by a strictly positive constant $D_\varepsilon>0$ and that $s\in\mathbb{N}$ is such that*

$$\max_{r=3,\dots,s}\max_{t\in\frac{1}{\varepsilon}K}|\Psi^{(r)}(t)|\leq C_\varepsilon<\infty.$$

Let u,v be optimizers of $\mathrm{D}_\varepsilon(\alpha,\beta)$. Then u is L-Lipschitz, $u(x)\in\operatorname{ran}[c(x,\cdot)-v(\cdot)]$ and

$$\|u^{(s)}\|_\infty=\mathcal{O}\left(1+\left(\frac{C_\varepsilon}{\varepsilon D_\varepsilon}\right)^{s-1}\right)$$

with the constant depending only on X and on the supremum norm of the first s derivatives of the cost function c.

We will split the proof into several lemmas. Throughout this section we will work under the assumptions in the statement of Proposition 4.1 and we will use the constants L, C_ε, D_ε defined in the statement.

Lemma 4.2. *The optimal potential u is L-Lipschitz.*

This was proved in Marino and Gerolin (2020). Alternatively, it can be seen directly from Equation (6) by applying ∂_x, splitting the integral and taking absolute values.

Lemma 4.3. *If X and Y are compact sets of $\mathbb{R}^d$ and $c\in\mathcal{C}^\infty$, then*

$$u(x)\leq\max_y(c(x,y)-v(y))\quad\textit{and}\quad u(x)\geq\min_y(c(x,y)-v(y)).$$

for all x in X.

Proof. For u, v maximizers we know that the identity

$$1 = \int \Psi'\left((u(x) + v(y) - c(x,y))/\varepsilon\right)\, d\beta(y)$$

holds for every x in X. Since β is a probability measure, we have that for every x

$$\begin{aligned} 1 &\in \left[\min_y \Psi'\left((u(x) + v(y) - c(x,y))/\varepsilon\right), \max_y \Psi'\left((u(x) + v(y) - c(x,y)/\varepsilon)\right)\right] \\ &= \left[\Psi'\left(\frac{1}{\varepsilon}\min_y(u(x) + v(y) - c(x,y))\right), \Psi'\left(\frac{1}{\varepsilon}\max_y(u(x) + v(y) - c(x,y))\right)\right], \end{aligned}$$

where we used that Ψ' is increasing. We also have that Φ' is increasing and that, since Ψ is the Legendre conjugate of Φ, it holds that $(\Psi')^{-1} = \Phi'$. Hence,

$$\Phi'(1) \geq \left(\frac{1}{\varepsilon}\min_y(u(x) + v(y) - c(x,y))\right),$$

and

$$\Phi'(1) \leq \left(\frac{1}{\varepsilon}\max_y(u(x) + v(y) - c(x,y))\right).$$

Using the assumption $\Phi'(1) = 0$ in Definition 2.4, we conclude

$$u(x) \leq \max_y(c(x,y) - v(y)) \quad \text{and} \quad u(x) \geq \min_y(c(x,y) - v(y)).$$

□

Without loss of generality we can assume that $u(x_0) = 0$ for some $x_0 \in X$; else, replace u and v by $u - u(x_0)$ and $v + u(x_0)$. Thus, by the last two lemmas, $\forall x \in X, y \in Y$, $|u(x)| \leq L|x|$ and $|v(y)| \leq \|c\|_\infty + \|u\|_\infty$. In particular, if we define K as in the statement of Proposition 4.1 by $K := [-2(L|X| + \|c\|_\infty), 2(L|X| + \|c\|_\infty)]$, we have

$$\forall\, x \in X,\, y \in Y, \quad u(x) + v(y) - c(x,y) \in K.$$

We are now in the position to derive the desired bounds on the Sobolev norm under our assumptions. We set

$$E := \max_{i=1,\dots,s} \|c^{(i)}\|_\infty. \tag{7}$$

Lemma 4.4. *For every $m \leq s$ it holds that*

$$\|u^{(m)}\|_\infty = \mathcal{O}\left(1 + \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{m-1} \frac{1}{\varepsilon^{m-1}}\right),$$

where the implicit constant depends only on m and E.

Proof. We proceed by induction for $l \leq s$. For $l = 1$ we know from Lemma 4.2 that $\|u'\|_\infty \leq \|c'\|_\infty \leq E$.
For a bigger l, we use the iterated chain rule formula to derive that

$$0 = \partial_x^l \int \Psi'((u + v - c)/\varepsilon) d\beta = \sum_{\alpha \in T_l} \binom{l}{\alpha} \int \frac{\Psi^{(|\alpha|+1)}}{\varepsilon^{|\alpha|}} \prod_{m=1}^{l} \left(\frac{u^{(m)} - \partial_x^m c}{m!}\right)^{\alpha_m} d\beta,$$

with the multi-index set $T_l := \{\alpha \in \mathbb{N}^l \mid \sum_{s=1}^{l} s\alpha_s = l\}$. The argument of $(u+v-c)/\varepsilon$ is dropped for brevity whenever possible.
The α in T_l having $\alpha_l = 1$ gives us the $u^{(l)}$ we want to estimate, so

$$\begin{aligned}
|u^{(l)}(x)| &= \frac{1}{\int \Psi'' d\beta} \left| \int c^{(l)}\Psi'' + \sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} \frac{\Psi^{(|\alpha|+1)}}{\varepsilon^{|\alpha|-1}} \prod_{m=1}^{l-1} \left(\frac{u^{(m)} - \partial_x^m c}{m!} \right)^{\alpha_m} d\beta \right| \\
&\le E + \frac{1}{\int \Psi'' d\beta} \left[\sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} \frac{\int \Psi^{(|\alpha|+1)} d\beta}{\varepsilon^{|\alpha|-1}} \prod_{m=1}^{l-1} \left(\frac{\|u^{(m)}\|_\infty + E}{m!} \right)^{\alpha_m} \right] \\
&\le E + \frac{C_\varepsilon}{D_\varepsilon} \left[\sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} \frac{1}{\varepsilon^{|\alpha|-1}} \prod_{m=1}^{l-1} \frac{(\|u^{(m)}\|_\infty^{\alpha_m} \vee 1)}{m!^{\alpha_m}} (E+1)^{\alpha_m} \right].
\end{aligned}$$

In the following computations we will assume that $\varepsilon \le 1$; for $\varepsilon \ge 1$, one can just bound all terms of ε^{-l} by 1 and copy the same steps.
Suppose by inductive hypothesis that

$$\|u^{(m)}\|_\infty \le M_m \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{(m-1)} \frac{1}{\varepsilon^{m-1}}$$

for all $m \le l-1$, for some constant $M_m = M_m(E)$. Without loss of generality $M_{m-1} \le M_m$. In the following steps we use that for $\alpha \in T_l$, $\sum_m m\alpha_m = l$.

$$\begin{aligned}
\|u^{(l)}\|_\infty &\le E + \frac{C_\varepsilon}{D_\varepsilon} \left[\sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} \frac{1}{\varepsilon^{|\alpha|-1}} \prod_{m=1}^{l-1} \left(M_m \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{(m-1)} \frac{1}{\varepsilon^{m-1}} \right)^{\alpha_m} \frac{(E+1)^{\alpha_m}}{m!^{\alpha_m}} \right] \\
&\le E + \frac{C_\varepsilon}{D_\varepsilon} \left[\sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} \frac{1}{\varepsilon^{|\alpha|-1}} M_{l-1}^{|\alpha|} \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{\sum_{m=1}^{l-1}(m-1)\alpha_m} \frac{1}{\varepsilon^{\sum_{m=1}^{l-1}(m-1)\alpha_m}} \prod_{m=1}^{l-1} \frac{(E+1)^{\alpha_m}}{m!^{\alpha_m}} \right] \\
&\le E + \frac{C_\varepsilon}{D_\varepsilon} \left[\sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} \frac{1}{\varepsilon^{|\alpha|-1}} M_{l-1}^{|\alpha|} \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{l-|\alpha|} \frac{1}{\varepsilon^{l-|\alpha|}} \prod_{m=1}^{l-1} \frac{(E+1)^{\alpha_m}}{m!^{\alpha_m}} \right] \\
&\le E + \frac{1}{\varepsilon^{l-1}} \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{l-1} \left[\sum_{\substack{\alpha\in T_l\\ \alpha_l\neq 1}} \binom{l}{\alpha} M_{l-1}^{|\alpha|} \prod_{m=1}^{l-1} \frac{(E+1)^{\alpha_m}}{m!^{\alpha_m}} \right] \\
&= \mathcal{O}\left(1 + \left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{l-1} \frac{1}{\varepsilon^{l-1}} \right),
\end{aligned}$$

and the constant depends only on E and l. □

5 Sample Complexity for Generalized Regularization Functions

By applying Proposition 4.1 to $s = \lfloor \frac{d}{2} \rfloor + 1$, we have already proved that the optimal potentials are in a bounded ball with radius λ_ε in the RKHS $\mathbb{H}^s(\mathbb{R}^d)$ for $s = \lfloor \frac{d}{2} \rfloor + 1$, with λ_ε independent of the measures. We can now follow the reasoning in Genevay, Chizat, et al. (2019) to derive a quantitative bound on the sample complexity for a general regularization function.
We introduce the abbreviation $f_\varepsilon(u,v) = f_\varepsilon^{x,y}(u,v) = u(x) + v(y) - \varepsilon\, \Psi((u(x)+v(y)-c(x,y))/\varepsilon)$ and prove that f_ε is Lipschitz in (u,v) on a certain subset that contains the optimal potentials.

Lemma 5.1. *Let $\mathcal{A} = \{(u,v) \mid u \oplus v \leq 2L|X| + \|c\|_\infty\}$. We have:*

(i) the pairs of optimal potentials (u^*, v^*) such that $u^*(0) = 0$ belong to $\mathcal{A}$,

(ii) f_ε is B-Lipschitz in (u,v) on $\mathcal{A}$ with $B \leq 1 + \Psi'(\kappa\varepsilon^{-1})$, where $\kappa := 2L|X| + \|c\|_\infty$.

Proof. Let us prove that we can restrict ourselves to a subset on which f_ε is Lipschitz in (u,v).

$$\nabla f_\varepsilon(u,v) = 1 - \Psi'\left(\frac{u+v-c}{\varepsilon}\right).$$

To ensure that f_ε is Lipschitz, we need to ensure that the quantity inside Ψ' is upperbounded at optimality and then restrict the function to all (u,v) that satisfy that bound.
By Lemma 4.3 and Lemma 4.2 we know that if (u,v) are potentials such that $u(0) = 0$ then for every $x \in X, y \in Y$,

$$|u(x)| \leq L|x| \quad \text{and} \quad |v(y)| \leq \|c\|_\infty + \|u\|_\infty.$$

So at optimality for all $x \in X$, $y \in Y$,

$$|u(x) + v(y)| \leq 2L|X| + \|c\|_\infty.$$

Denoting $\mathcal{A} := \{(u,v) \in (\mathbb{H}^s(\mathbb{R}^d))^2 \mid u \oplus v \leq 2L|X| + \|c\|_\infty\}$, we have that for all (u,v) in $\mathcal{A}$,

$$|\nabla f_\varepsilon(u,v)| \leq 1 + \sup_{(u,v)\in\mathcal{A}} |\Psi'\left(\frac{u+v-c}{\varepsilon}\right)| \leq 1 + \Psi'(\kappa\varepsilon^{-1}).$$

for $\kappa := 2L|X| + \|c\|_\infty$. Where we used that Φ is restricted to $[0,\infty]$ (as in Definition 2.4), so $\Psi' \geq 0$ and we can get rid of the absolute value, and the fact that Ψ' is increasing. □

We will need one more auxiliary lemma before we can prove the main result.

Lemma 5.2. *Let $\mathcal{H}^s_\lambda := \{u \in \mathbb{H}^s(\mathbb{R}^d) \mid \|u\|_{\mathbb{H}^s(\mathbb{R}^d)} \leq \lambda\}$, then there exists λ such that:*

$$|\mathrm{D}_\varepsilon(\alpha,\beta) - \mathrm{D}_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| \leq 3 \sup_{(u,v)\in(\mathcal{H}^s_\lambda)^2} |\mathbb{E} f_\varepsilon^{XY}(u,v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v)|.$$

Proof. Applying the triangular inequality we have that

$$\begin{aligned}|\mathrm{D}_\varepsilon(\alpha,\beta) - \mathrm{D}_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| &= |\mathbb{E} f_\varepsilon^{XY}(u^*,v^*) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat{u},\hat{v})| \\ &\leq |\mathbb{E} f_\varepsilon^{XY}(u^*,v^*) - \mathbb{E} f_\varepsilon^{XY}(\hat{u},\hat{v})| + |\mathbb{E} f_\varepsilon^{XY}(\hat{u},\hat{v}) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat{u},\hat{v})|.\end{aligned}$$

From Proposition 4.1, we know that the all the dual potentials are bounded in $\mathbb{H}^s(\mathbb{R}^d)$ by a constant $\lambda = \lambda_\varepsilon$ which does not depend on the measures. Thus, the second term is bounded by

$\sup_{(u,v)\in(\mathcal{H}^s_\lambda)^2} |\mathbb{E}f_\varepsilon(u,v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon(u,v)|$.

The first quantity is non-negative since (u^*, v^*) is the maximizer of $\mathbb{E}f_\varepsilon(\cdot,\cdot)$ so we don't need the absolute value. We have the following bound

$$\begin{aligned}\mathbb{E}f_\varepsilon^{XY}(u^*,v^*) - \mathbb{E}f_\varepsilon^{XY}(\hat{u},\hat{v}) \leq & \left(\mathbb{E}f_\varepsilon^{XY}(u^*,v^*) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u^*,v^*)\right) \\ & + \left(\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u^*,v^*) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat{u},\hat{v})\right) \\ & + \left(\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\hat{u},\hat{v}) - \mathbb{E}f_\varepsilon^{XY}(\hat{u},\hat{v})\right).\end{aligned}$$

Now we can conclude since the first and last terms can be bounded by $\sup_{(u,v)\in(\mathcal{H}^s_\lambda)^2} |\mathbb{E}f_\varepsilon^{XY}(u,v) - \frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(u,v)|$ and the second term is non-positive since $(\hat{u},\hat{v})$ maximizes $\frac{1}{n}\sum_{i=1}^n f_\varepsilon^{X_iY_i}(\cdot,\cdot)$. □

To prove Theorem 5.4, we will use this standard result on RKHS's.

Proposition 5.3. *(Bartlett and Mendelson (2003)) Consider α a probability distribution, ℓ a B-lipschitz loss and $\mathcal{G}$ a given class of functions. Then*

$$\mathbb{E}_\alpha\left[\sup_{g\in\mathcal{G}} \mathbb{E}_\alpha \ell(g,X) - \frac{1}{n}\sum_{i=1}^n \ell(g,X_i)\right] \leq 2B\mathbb{E}_\alpha \mathcal{R}(\mathcal{G}(X_1^n))$$

where $\mathcal{R}(\mathcal{G}(X_1^n))$ is the Rademacher complexity of class $\mathcal{G}$ defined by
$\mathcal{R}(\mathcal{G}(X_1^n)) = \sup_{g\in\mathcal{G}} \mathbb{E}_\sigma \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)$, where $(\sigma_i)_i$ are iid Rademacher random variables. Besides, when $\mathcal{G}$ is a ball of radius λ in a RKHS with kernel k the Rademacher complexity is bounded by

$$\mathcal{R}(\mathcal{G}_\lambda(X_1^n)) \leq \frac{\lambda}{n}\sqrt{\sum_{i=1}^n k(X_i,X_i)}.$$

With all ingredients assembled, we can now prove Theorem 5.4.

Theorem 5.4. *Consider the regularized transport problem between two measures α and β on X and Y two bounded subsets of $\mathbb{R}^d$, with a $\mathcal{C}^\infty$, L-Lipschitz cost c and C^∞ transform Ψ. One has*

$$\mathbb{E}|\mathrm{D}_\varepsilon(\alpha,\beta) - \mathrm{D}_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| = \mathcal{O}\left(\frac{\left(1+\Psi'(\kappa\varepsilon^{-1})\right)\left(1+\left(\frac{1}{\varepsilon}\frac{C_\varepsilon}{D_\varepsilon}\right)^{\lfloor d/2\rfloor}\right)}{\sqrt{n}}\right)$$

where $\kappa := 2L|X| + \|c\|_\infty$, and the constant only depends on $|X|$, $|Y|$, d, and E (where C_ε, D_ε were defined in Proposition 4.1, and E was defined in (7) for $s = \lfloor d/2\rfloor + 1$).

Proof. We start by applying Lemma 5.2 to bound the right hand side. Then, since f_ε is Lipschitz and we are optimizing over $\mathbb{H}^s(\mathbb{R}^d)$ which is a RKHS, we can apply Proposition 5.3 to bound the supremum in Lemma 5.2. We get:

$$\mathbb{E}|\mathrm{D}_\varepsilon(\alpha,\beta) - \mathrm{D}_\varepsilon(\hat{\alpha}_n,\hat{\beta}_n)| \leq 3\frac{2B\lambda}{n}\mathbb{E}\sqrt{\sum_{i=1}^n k(X_i,X_i)} \leq 3\frac{2B\lambda}{n}\sqrt{n\max_{x\in X} k(x,x)}.$$

Where

- $B \leq 1 + \Psi'(\kappa\varepsilon^{-1})$, where $\kappa = 2L|X| + \|c\|_\infty$ (Lemma 5.1),
- $\lambda = \mathcal{O}\left(\max\left(1, \frac{1}{\varepsilon^{\lfloor d/2 \rfloor}}\left(\frac{C_\varepsilon}{D_\varepsilon}\right)^{\lfloor d/2 \rfloor}\right)\right)$ (Lemma 4.4).
- k is the kernel associated to $\mathbb{H}^s(\mathbb{R}^d)$ and thus $\max_{x\in X} k(x,x) = k(0,0) =: \mathfrak{K}$ which doesn't depend on n or ε.

Combining all these bounds, we get the desired convergence rate in $\frac{1}{\sqrt{n}}$. □

6 Computation of Sinkhorn Divergences

The regularized optimal transport problem is particularly interesting as it has favourable computational properties and, as we have seen, it does not suffer from the curse of dimensionality. In particular, entropic regularization, see Example 2.1, provides a multiplicative structure that can be exploited to construct an efficient algorithm. We recall the discrete, entropy-regularized optimal transport problem

$$\mathrm{OT}_{\mathrm{disc},\varepsilon}(a,b) = \min_{\mathrm{P}\in U(a,b)} \sum_{i,j} \mathrm{P}_{ij}\mathrm{C}_{ij} + \varepsilon \sum_{i,j} \mathrm{P}_{ij}(\log(\mathrm{P}_{ij}) - 1), \tag{8}$$

omitting constants that are not relevant for optimisation. Cuturi (2013) proved that the solution has a particular form.

Theorem 6.1. *The solution* P *to (8) is unique and has the form*

$$\mathrm{P}_{ij} = u_i \mathrm{K}_{ij} v_j, \quad \forall\, i \in \{1,\dots,n\}, j \in \{1,\dots,m\},$$

for unknown vectors $u \in \mathbb{R}^n_+, v \in \mathbb{R}^n_+$ *and* $\mathrm{K}_{ij} = \exp(-\mathrm{C}_{ij}/\varepsilon)$.

Remark 6.1. The vectors u and v are unique up to scaling and correspond to the dual solution.

Since P is an element of the space $U(a,b)$, we obtain

$$\mathrm{diag}(u)K\,\mathrm{diag}(v)\mathbb{1}_m = a, \qquad \mathrm{diag}(v)K^T\,\mathrm{diag}(u)\mathbb{1}_n = b.$$

We can use this two relations to define an algorithm that alternatingly updates u and v.

Algorithm 6.2 (Sinkhorn Algorithm)**.** Let l denote the iteration number. The update scheme

$$u^{(l+1)} := \frac{a}{\mathrm{K}v^{(l)}}, \qquad v^{(l+1)} := \frac{b}{\mathrm{K}^T u^{(l+1)}},$$

where $v^{(0)} = \mathbb{1}_m$ and the division is componentwise, is called *Sinkhorn algorithm.*

Sinkhorn (1964) proved the convergence of the iteration which gave the algorithm its name.
In recent years, Sinkhorn-type algorithms became an active field of research and many variations were proposed. Noteably, Muzellec et al. (2017) derived a variation for the Tsallis entropy.
Moreover, there exists a continuous analogue to Algorithm 6.2 which converges as well, as Rüschendorf (1995) proved. Marino and Gerolin (2020) applied this algorithm to the optimal transport problem with a general convex regularization function (1).

Algorithm 6.3 (Generalized Sinkhorn Algorithm)**.** In the setting of Definition 2.4, the *generalized Sinkhorn algorithm* is given by the iteration

$$\begin{aligned} u^{(l+1)}(x) &:= \underset{u \in C(X)}{\operatorname{argmax}} \left\{ \int_X u \, d\alpha - \varepsilon \int_{X \times Y} \Psi\left(\frac{u + v^{(l)} - c}{\varepsilon} \right) d(\alpha \otimes \beta) \right\}, \\ v^{(l+1)}(y) &:= \underset{v \in C(Y)}{\operatorname{argmax}} \left\{ \int_Y v \, d\beta - \varepsilon \int_{X \times Y} \Psi\left(\frac{u^{(l)} + v - c}{\varepsilon} \right) d(\alpha \otimes \beta) \right\}. \end{aligned}$$

Although this algorithm converges, finding the maximizers in closed form is only possible for specific regularizers. For this reason, we turn to another class of algorithms.
Ferradans et al. (2014) realized that the conditonal gradient algorithm, first proposed by Frank and Wolfe (1956), is suited to accommodate complex regularization functions. Indeed, if the objective is differentiable, we can find the steepest descent direction at the current location via interior-point methods, cf. Nesterov and Nemirovskii (1994), and then update the position towards the minimum.

Algorithm 6.4 (Frank-Wolfe Algorithm)**.** In the setting of Definition 2.4 with discrete probability measures α and β, we define

$$f(\mathrm{P}) := \sum_{i,j} \mathrm{P}_{ij} \mathrm{C}_{ij} + \varepsilon \, \Phi \left(\frac{\mathrm{P}_{ij}}{a_i b_j} \right),$$

and assume that f is differentiable. The *Frank-Wolfe algorithm* is given by the iteration

$$\begin{aligned} \tilde{\mathrm{P}}^{(l+1)} &\in \underset{\mathrm{P} \in U(a,b)}{\operatorname{argmin}} \sum_{i,j} \left(\nabla f(\mathrm{P}^{(l)}) \right)_{ij} \tilde{\mathrm{P}}_{ij} \\ \mathrm{P}^{(l+1)} &:= \mathrm{P}^{(l)} + \tau_l \big(\tilde{\mathrm{P}}^{(l+1)} - \mathrm{P}^{(l)} \big), \end{aligned}$$

where l denotes the iteration number and τ_l is found via line search.

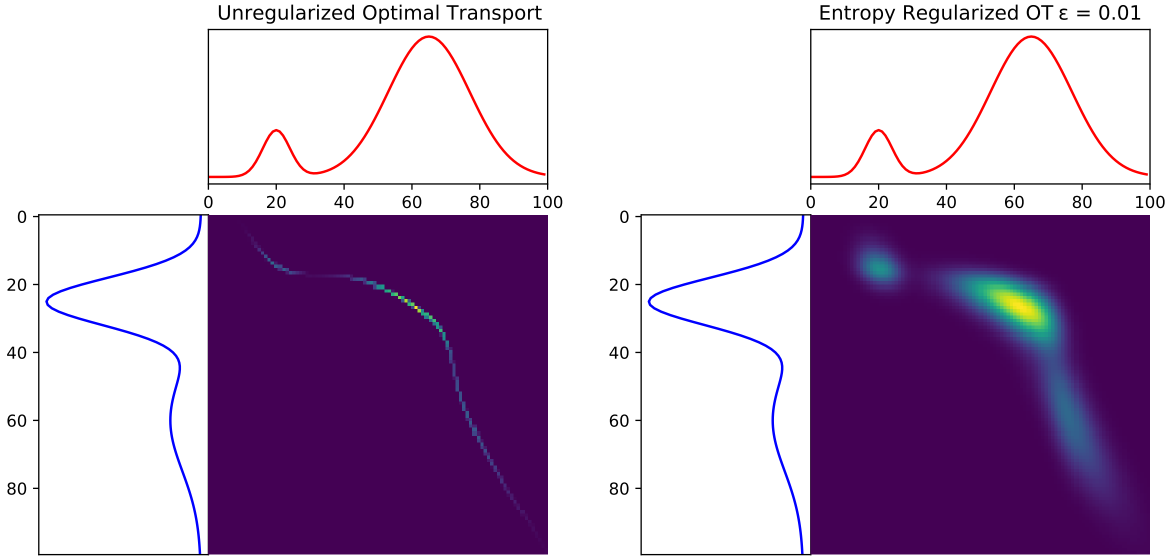
7 Experiments

The following numerical demonstrations were created using the Python Optimal Transport Library (Flamary et al. 2021).

7.1 Illustrating Optimal Transport

In this first section we illustrate the effect of regularization on the optimal transport plan of a 1D optimal transport problem. Figure 3 shows the optimal transport coupling between two 1D probability distributions, and also the optimal entropy regularized coupling (using the Sinkhorn algorithm) with $\varepsilon = 0.01$. Adding regularization broadens the coupling distribution and turns it away from a one-to-one mapping.
Figure 4 shows the optimal transport plan using Tsallis entropy regularization with different q. Specifically, the regularization term is ε times the relative Tsallis entropy with respect to the product distribution of the source and target distribution, where the relative Tsallis entropy between distributions α and β is given by

$$H_q(\alpha \| \beta) = \frac{1}{q-1} \int \left[\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta} \right)^{q-1} - 1 \right] d\alpha. \tag{9}$$



(a) unregularized

(b) entropy regularized ($\varepsilon = 0.01$)

Figure 3: Comparison of unregularized and entropy regularized 1D optimal transport plans.

This corresponds to the formalism of (2) together with (3). Numerically, we calculate this regularized optimal transport plan using the Frank-Wolfe algorithm (Algorithm 6.4). While it is not particularly fast or optimized for the given scenario, its generality allows for easy experimentation with different regularizers.

For each q in Figure 4, ε is chosen such that all three distributions are about equally "broad" to highlight the influence of q on the shape and the edges of the optimal transport plan. It is apparent that for suitably chosen ε, Tsallis entropy regularization achieves a regularization behavior similar in style to the Shannon entropy but with differences in the shape of the optimal transport plan, which depends on q. It is thus natural to expect an overall similar behavior of Tsallis entropy regularization in terms of usefulness in applications and also sample complexity.

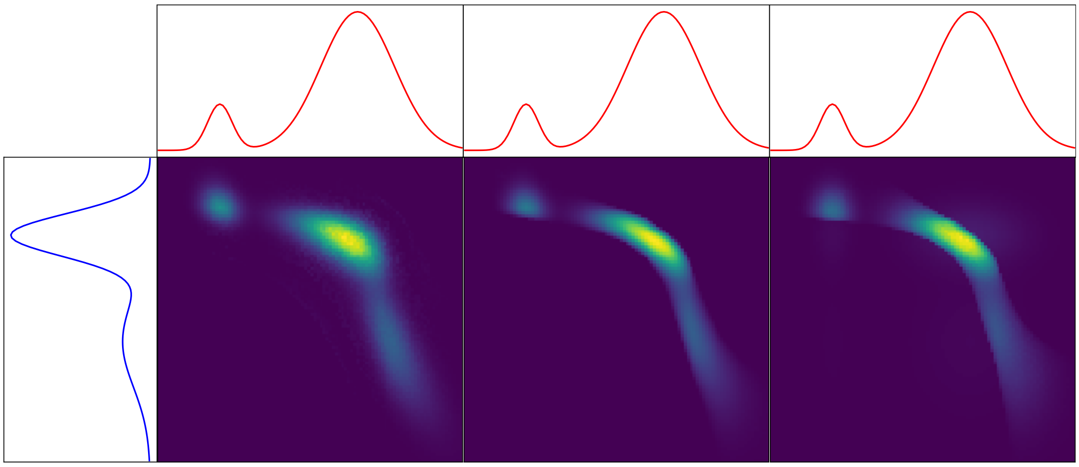


Figure 4: Optimal transport plans with relative Tsallis entropy for $q = (1.1, 2, 3)$ and $\varepsilon = (0.01, 0.001, 0.0005)$.

7.2 Demonstrating the Sample Complexity

7.2.1 Simulation Set-up

To demonstrate the sample complexity, ideally we would like to compare the optimal transport distance between two continuous distributions α, β to the optimal transport distance between two samples from them $\hat{\alpha}_n, \hat{\beta}_n$. However, we do not have a numerical method to calculate the optimal transport distance between two continuous distributions exactly. One easy case is when the optimal transport distance is zero i.e.,

$$\overline{\mathrm{OT}}_\varepsilon(\alpha,\beta) = \mathrm{OT}_\varepsilon(\alpha,\beta) - \frac{1}{2}\left(\mathrm{OT}_\varepsilon(\alpha,\alpha) + \mathrm{OT}_\varepsilon(\beta,\beta)\right) = 0 \quad \Longleftrightarrow \quad \alpha = \beta \,. \tag{10}$$

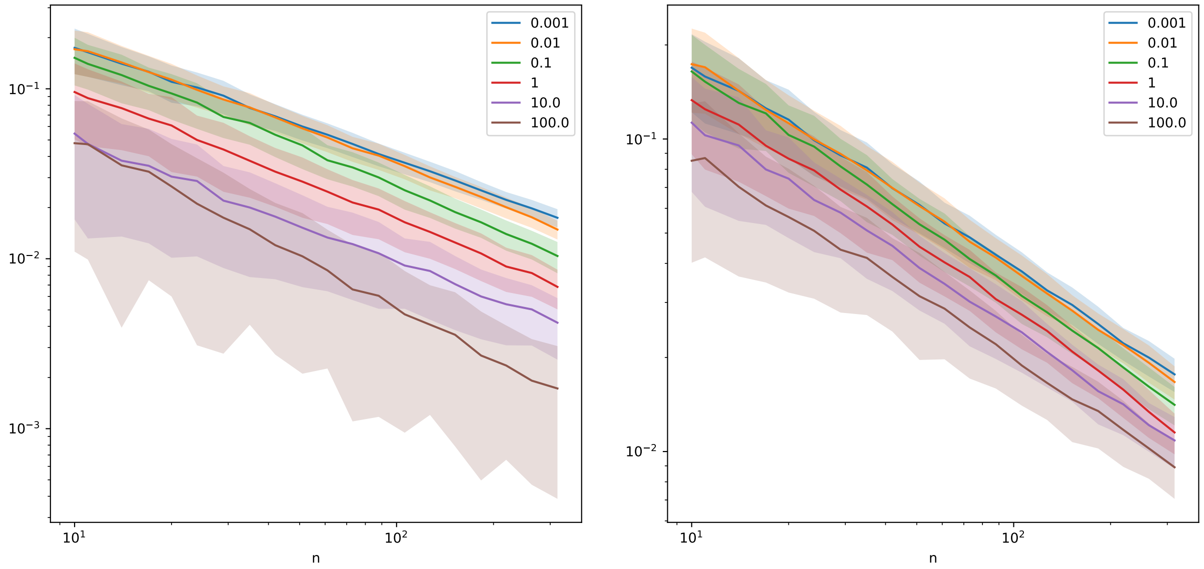
In this case, with $\hat{\alpha}_n$ and $\hat{\alpha}'_n$ two independent samples from α, we then have

$$\mathbb{E}|\overline{\mathrm{OT}}_\varepsilon(\hat{\alpha}_n,\hat{\alpha}'_n)| = \mathbb{E}|\overline{\mathrm{OT}}_\varepsilon(\hat{\alpha}_n,\hat{\alpha}'_n) - \overline{\mathrm{OT}}_\varepsilon(\alpha,\alpha)| \overset{?}{=} \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) . \tag{11}$$

So this tests the desired behavior and is straightforward to calculate numerically. The plots in this secion generally show $\mathbb{E}|\mathrm{OT}_\varepsilon(\hat{\alpha}_n,\hat{\alpha}'_n)|$ as a function of n, where n increases on a logarithmic scale from 10 to $10^{2.5}$, and each point is an average of 300 independent samples. The shaded area shows the range of one standard deviation. For α we choose the uniform distribution over the unit hypercube in d dimensions (d will vary). Different curves then stand for different values of ε, d or parameters of the regularization function.

7.2.2 Tsallis Relative Entropy Regularization

Figure 5 depicts the convergence behaviour in three dimensions for different ε and q.



(a) $d = 3$, $q = 2$, different ε

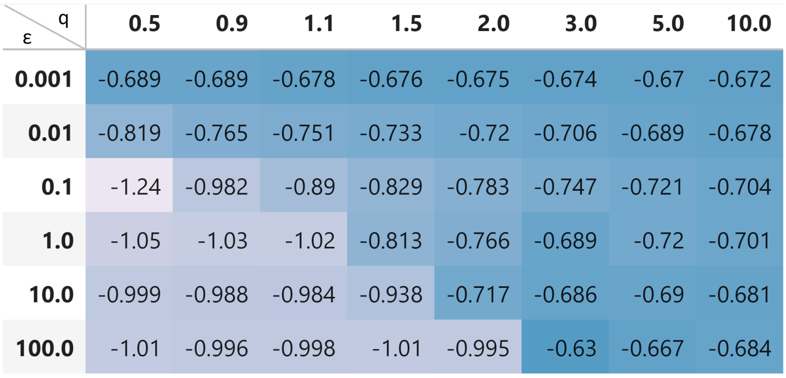
(b) $d = 3$, $q = 5$, different ε

Figure 5: Dependence of the convergence behavior on q and ε.

The plots hint at a linear relationship on a log-log scale. Thus, we fit functions of the form

$$\mathrm{OT}_\varepsilon(\hat{\alpha}_n,\hat{\alpha}'_n) \approx An^t \tag{12}$$

to investigate the respective exponents. Table 1 shows the exponent t for a range of different ε and q. We find that for all tested parameter combinations here the convergence behavior is better than $\mathcal{O}(1/\sqrt{n})$.

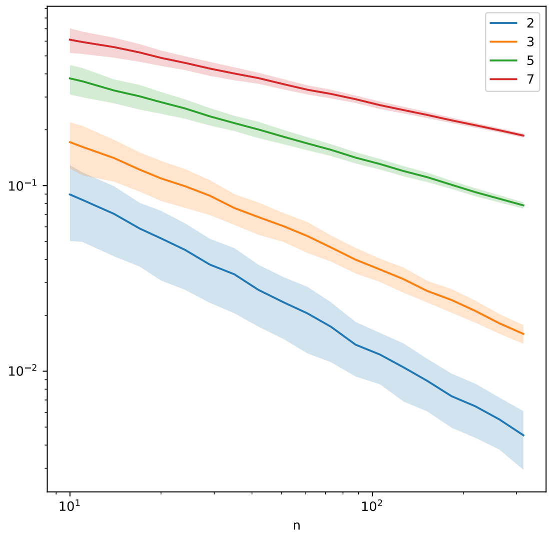


ε \ q	0.5	0.9	1.1	1.5	2.0	3.0	5.0	10.0
0.001	-0.689	-0.689	-0.678	-0.676	-0.675	-0.674	-0.67	-0.672
0.01	-0.819	-0.765	-0.751	-0.733	-0.72	-0.706	-0.689	-0.678
0.1	-1.24	-0.982	-0.89	-0.829	-0.783	-0.747	-0.721	-0.704
1.0	-1.05	-1.03	-1.02	-0.813	-0.766	-0.689	-0.72	-0.701
10.0	-0.999	-0.988	-0.984	-0.938	-0.717	-0.686	-0.69	-0.681
100.0	-1.01	-0.996	-0.998	-1.01	-0.995	-0.63	-0.667	-0.684

Table 1: Fit parameter t for the Tsallis relative entropy regularization (d = 3).

Larger dimension

Figure 6 shows the convergence behavior for fixed q and ε but different d. One finds that for large d, at least for the range of n considered, the behavior is actually worse than $1/\sqrt{n}$. This is found already using traditional entropy regularization (see plots in Genevay, Chizat, et al. (2019), although not explicitly mentioned in their text), and is not entirely surprising: If ε is very small, or alternatively the unregularized optimal transport distance very large, then the problem is essentially unregularized, and we expect the sample complexity to behave badly until n becomes large enough so that the unregularized distance is small and the regularization becomes dominant. This is consistent with the constant of the $1/\sqrt{n}$ upper bound increasing with decreasing ε. Since we are numerically limited to a quite small range of n where computation is still feasible, it is hard to demonstrate in practice that for very large n the convergence eventually



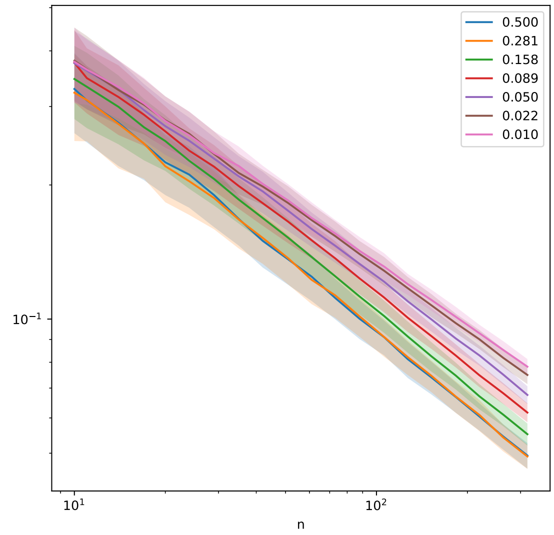
(a) Plot as a function of n

d	t
2	-0.870
3	-0.699
5	-0.463
7	-0.349

(b) Fit parameter t for different d

Figure 6: Dependence of the convergence behavior on d with $q = 3$ and $\varepsilon = 0.01$ fixed.

returns to $\mathcal{O}(1/\sqrt{n})$. However, for a fixed range of n we can try to find a "critical" range of ε where the sample complexity behavior shifts. This is attempted in Figure 7. One finds in this specific case that for $\varepsilon \lessapprox 0.25$ the exponent of the decay gradually shifts towards larger values.



(a) Plot as a function of n

ε	t
0.500	-0.554
0.281	-0.552
0.158	-0.546
0.089	-0.528
0.050	-0.506
0.022	-0.477
0.010	-0.464

(b) Fit parameter t for different ε

Figure 7: For $d = 5$ and $q = 3$, attempt to find the critical ε below which the convergence behavior becomes worse than $\mathcal{O}(1/\sqrt{n})$ in the range of n numerically accessible.

References

Bartlett, Peter L. and Shahar Mendelson (2003). "Rademacher and Gaussian complexities: risk bounds and structural results". In: *J. Mach. Learn. Res.* 3.3, pp. 463–482.

Brenier, Yann (1991). "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on Pure and Applied Mathematics* 44.4, pp. 375–417.

Cuturi, Marco (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc.

Ferradans, Sira, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol (2014). "Regularized discrete optimal transport". In: *SIAM J. Imaging Sci.* 7.3, pp. 1853–1882.

Flamary, Rémi, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer (2021). "POT: Python Optimal Transport". In: *Journal of Machine Learning Research* 22.78, pp. 1–8.

Frank, Marguerite and Philip Wolfe (1956). "An algorithm for quadratic programming". In: *Naval Research Logistics Quarterly* 3.1-2, pp. 95–110.

Genevay, Aude, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré (2019). "Sample Complexity of Sinkhorn Divergences". In: *Proceedings of Machine Learning Research*. Vol. 89. Proceedings of Machine Learning Research. PMLR, pp. 1574–1583.

Genevay, Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach (2016). "Stochastic Optimization for Large-scale Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc.

Kantorovich, Leonid (1942). “On the transfer of masses (in russian)”. In: *Doklady Akademii Nauk* 37.2, pp. 227–229.

Marino, Simone Di and Augusto Gerolin (2020). “Optimal Transport losses and Sinkhorn algorithm with general convex regularization”.

Monge, Gaspard (1781). “Memoire sur la theorie des deblais et des remblais”. In: *Histoire de l’Academie Royale des Sciences de Paris.*

Muzellec, Boris, Richard Nock, Giorgio Patrini, and Frank Nielsen (2017). “Tsallis Regularized Optimal Transport and Ecological Inference”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1.

Nesterov, Y. and A. Nemirovskii (1994). *Interior-point polynomial algorithms in convex programming.* Vol. 13. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, pp. ix + 405.

Peyré, Gabriel and Marco Cuturi (2019). “Computational Optimal Transport”. In: *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Rüschendorf, Ludger (1995). “Convergence of the iterative proportional fitting procedure”. In: *Ann. Stat.* 23.4, pp. 1160–1174.

Sinkhorn, R. (1964). “A relationship between arbitrary positive matrices and doubly stochastic matrices”. In: *Ann. Math. Stat.* 35, pp. 876–879.

Wilson, A. G. (1969). “The Use of Entropy Maximising Models, in the Theory of Trip Distribution, Mode Split and Route Split”. In: *Journal of Transport Economics and Policy* 3.1, pp. 108–126.