Minimax estimation of discontinuous optimal transport maps: The semi-discrete case

Aram-Alexandre Pooladian [*1] **Vincent Divol** [*2] **Jonathan Niles-Weed** [1 3]

Abstract

We consider the problem of estimating the optimal transport map between two probability distributions, P and Q in $\mathbb{R}^d$, on the basis of i.i.d. samples. All existing statistical analyses of this problem require the assumption that the transport map is Lipschitz, a strong requirement that, in particular, excludes any examples where the transport map is discontinuous. As a first step towards developing estimation procedures for discontinuous maps, we consider the important special case where the data distribution Q is a discrete measure supported on a finite number of points in $\mathbb{R}^d$. We study a computationally efficient estimator initially proposed by Pooladian & Niles-Weed (2021), based on entropic optimal transport, and show in the semi-discrete setting that it converges at the minimax-optimal rate $n^{-1/2}$, independent of dimension. Other standard map estimation techniques both lack finite-sample guarantees in this setting and provably suffer from the curse of dimensionality. We confirm these results in numerical experiments, and provide experiments for other settings, not covered by our theory, which indicate that the entropic estimator is a promising methodology for other discontinuous transport map estimation problems.

1. Introduction

The theory of optimal transport (OT) defines a natural geometry on the space of probability measures (Santambrogio, 2015; Villani, 2009) and has become ubiquitous in modern data-driven tasks. In this area, *optimal transport maps* are a central object of study: suppose P and Q are two probability distributions with finite second moments, with P having a density with respect to the Lebesegue measure on $\mathbb{R}^d$. Then, Brenier's theorem (see Section 2.1) states that there exists a convex function φ_0 whose gradient defines a unique *optimal transport map* between P and Q. This map is optimal in the sense that it minimizes the following objective function:

$$\nabla\varphi_0 \coloneqq \operatorname*{argmin}_{T\in\mathcal{T}(P,Q)} \int \tfrac{1}{2}\|x-T(x)\|^2 \,\mathrm{d}P(x)\,, \qquad (1)$$

where $\mathcal{T}(P,Q) \coloneqq \{T:\mathbb{R}^d\to\mathbb{R}^d \mid X\sim P,\ T(X)\sim Q\}$ is the set of transport maps between P and Q. The optimal value of the objective function in Equation (1) is called the (squared) 2-Wasserstein distance, written explicitly as

$$\mathrm{S}_0(P,Q) = \int \tfrac{1}{2}\|x-\nabla\varphi_0(x)\|^2 \,\mathrm{d}P(x)\,,$$

though a more general formulation is available (see Section 2.1). Computing or approximating $\mathrm{S}_0(P,Q)$ as well as $\nabla\varphi_0$ has found use in several academic communities, such as economics (Carlier et al., 2016; Chernozhukov et al., 2017; Gunsilius & Xu, 2021; Torous et al., 2021), computational biology (Bunne et al., 2021; 2022; Demetçi et al., 2022; Lübeck et al., 2022; Moriel et al., 2021; Schiebinger et al., 2019; Yang et al., 2020), and computer vision (Feydy et al., 2017; Solomon et al., 2015; 2016), among many others.

Practitioners seldom have access to P or Q, but instead have access to i.i.d. samples $X_1,\dots,X_n\sim P$ and $Y_1,\dots,Y_n\sim Q$. On the basis of these samples, practitioners face both computational and statistical challenges when estimating $\nabla\varphi_0$. From a theoretical perspective, the statistical task of estimating optimal transport maps has attracted much interest in the last few years (Deb et al., 2021; Divol et al., 2022; Ghosal & Sen, 2022; Hütter & Rigollet, 2021; Manole et al., 2021; Muzellec et al., 2021; Pooladian & Niles-Weed, 2021).

The first finite-sample analysis of this problem was performed by Hütter & Rigollet (2021), who proposed an estimator for $\nabla\varphi_0$ under the assumption that φ_0 is $s+1$-times continuously differentiable, for $s>1$. They showed that a wavelet-based estimator $\hat{\varphi}_{\mathrm{W}}$ satisfies

$$\mathbb{E}\|\nabla\hat{\varphi}_{\mathrm{W}}-\nabla\varphi_0\|^2_{L^2(P)} \lesssim n^{-\frac{2s}{2s+d-2}}\log^2(n)\,,$$

[*]Equal contribution [1]Center for Data Science, New York University, NY, USA [2]CEREMADE, Université Paris Dauphine-PSL, Paris, France [3]Courant Institute of Mathematical Sciences, New York University, NY, USA. Correspondence to: Aram-Alexandre Pooladian <aram-alexandre.pooladian@nyu.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

and that this rate is minimax optimal up to logarithmic factors. Their analysis requires that P and Q have bounded densities with compact support $\Omega \subseteq \mathbb{R}^d$, and that φ_0 be both strongly convex and smooth. Implementing the estimator $\hat{\varphi}_{\mathrm{W}}$ is computationally challenging even in moderate dimensions, and is practically infeasible for $d > 3$. Follow up work has proposed alternative estimators which improve upon $\hat{\varphi}_{\mathrm{W}}$ either in computational efficiency or in the generality in which they apply. Though these subsequent works go significantly beyond the setting considered by Hütter & Rigollet (2021), none has eliminated the crucial assumption that φ_0 is smooth, i.e., that the transport map $\nabla\varphi_0$ is Lipschitz.

We highlight two estimators proposed in this line of work that are particularly practical. Manole et al. (2021) study the 1-Nearest Neighbor estimator $\hat{T}_{\mathrm{1NN}}$. This estimator is obtained by solving the empirical optimal transport problem between the samples, which is then extended to a function defined on $\mathbb{R}^d$ using a projection scheme; see Section 4 for more details. Given n samples from the source and target measures in $\mathbb{R}^d$, $\hat{T}_{\mathrm{1NN}}$ has a runtime of $\mathcal{O}(n^3)$ via the Hungarian Algorithm (see Peyré & Cuturi, 2019, Chapter 3), and, for $d \geq 5$, achieves the rate

$$\mathbb{E}\|\hat{T}_{\mathrm{1NN}} - \nabla\varphi_0\|^2_{L^2(P)} \lesssim n^{-\frac{2}{d}} \tag{2}$$

whenever the optimal Brenier potential φ_0 is smooth and strongly convex, and under mild regularity conditions on P. In another work, Pooladian & Niles-Weed (2021) conducted a statistical analysis of an estimator originally proposed by Seguy et al. (2018) based on entropic optimal transport. The efficiency of Sinkhorn's algorithm for large-scale problems (Cuturi, 2013; Peyré & Cuturi, 2019) makes this estimator attractive from a computational perspective, and Pooladian & Niles-Weed (2021) also give statistical guarantees, though these fall short of being minimax-optimal.

Despite this progress, none of the aforementioned results can be applied in situations where $\nabla\varphi_0$ is not Lipschitz. And in practice, even requiring the *continuity* of the transport map can be far too stringent. It is indeed too much to hope for that an underlying data distribution (e.g. over the space of images) has one single connected component; this is supported by recent work that stipulates that the underlying data distribution is the union of *disjoint* manifolds of varying intrinsic dimension (Brown et al., 2022). In such a setting, the transport map $\nabla\varphi_0$ will not be continuous, demonstrating the need of considering the problem of the statistical estimation of *discontinuous* transport maps to get closer to real-world situations.

As a first step, we choose to focus on the case where the target distribution $Q = \sum_{j=1}^{J} q_j \delta_{y_j}$ is discrete while the source measure P has full support, often called the *semi-discrete* setting in the optimal transport literature. In this setting, the optimal transport map $\nabla\varphi_0$ is constant over regions known as Laguerre cells (each cell corresponding to a different atom of the discrete measure), while displaying discontinuities on their boundaries (see Section 2.1.1 for more details). Figure 1 provides such an example. Semi-discrete optimal transport therefore provides a natural class of discontinuous transport maps.

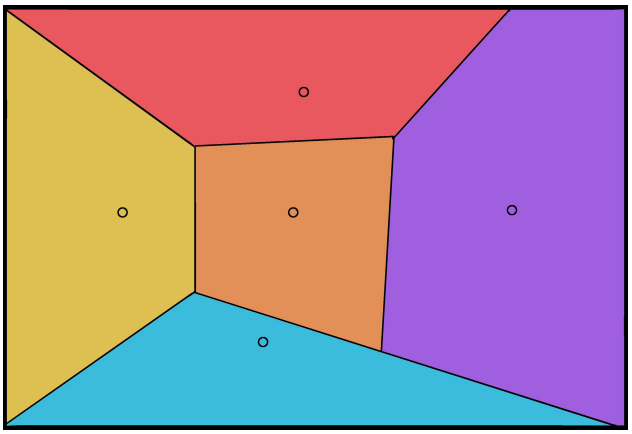

Figure 1. An illustration of a semi-discrete optimal transport map. The support of P, the whole rectangle, is partitioned into regions, each of which is transported to one of the atoms of the discrete target measure Q. The resulting map is discontinuous at the boundaries of each cell.

We focus on this setting for two reasons. First, it has garnered a lot of attention in recent years, in both computational and theoretical circles (see, e.g., Altschuler et al., 2022; Chen et al., 2022; Mérigot et al., 2021), due in particular to its connection with the quantization problem (Graf & Luschgy, 2007). Second, the semi-discrete setting is intriguing from a statistical perspective: existing results show that statistical estimation problems involving semi-discrete optimal transport can escape the curse of dimensionality (del Barrio & Loubes, 2019; del Barrio et al., 2022a; Forrow et al., 2019; Hundrieser et al., 2022). For example, Hundrieser et al. (2022, Theorem 3.2) show that if P_n and Q_n are empirical measures consisting of i.i.d. samples from P and Q, then the semi-discrete assumption implies

$$\mathbb{E}|S_0(P,Q) - S_0(P_n,Q_n)| \lesssim n^{-1/2}\,.$$

These results offer the tantalizing possibility that semi-discrete transport maps can be estimated at the rate $n^{-1/2}$, in sharp contrast to the dimension-dependent rates obtained in bounds such as (2). However, the optimal rates of estimation for semi-discrete transport maps are not known, and no estimators with finite-sample convergence guarantees exist.

MAIN CONTRIBUTIONS

We show that the computationally efficient estimator $\hat{T}_\varepsilon$ based on entropically regularized optimal transport, originally studied in (Pooladian & Niles-Weed, 2021; Seguy et al., 2018), provably estimates discontinuous semi-discrete optimal transport maps at the optimal rate. More precisely, our contributions are the following:

1. For Q discrete and P with full support on a compact, convex set, we show that $\hat{T}_\varepsilon$ achieves the following *dimension-independent* convergence rate to the optimal transport map (see Theorem 3.1)
$$\mathbb{E}\|\hat{T}_\varepsilon - \nabla\varphi_0\|^2_{L^2(P)} \lesssim n^{-1/2}\,, \tag{3}$$
when the regularization parameter $\varepsilon \asymp n^{-1/2}$. We further show (Proposition 4.1) that this rate is minimax optimal.
2. As a by-product of our analysis, we give new *parametric* rates of convergence to the entropic Brenier map T_ε, a result which improves exponentially on prior work in the dependence on ε (see Theorem 3.7 and Remark 3.8).
3. Our proof technique requires several new results, including a novel stability bound for the entropic Brenier maps (Proposition 3.9), and a new stability result for the entropic dual Brenier potentials in the semi-discrete case (Proposition 3.11).
4. We show that, unlike $\hat{T}_\varepsilon$, the 1-Nearest-Neighbor estimator is provably suboptimal in the semi-discrete setting (see Proposition 4.2) by exhibiting a discrete measure Q such that the risk suffers from the curse of dimensionality:
$$\mathbb{E}\|\hat{T}_{\text{1NN}} - \nabla\varphi_0\|^2_{L^2(P)} \gtrsim n^{-1/d}\,.$$
5. In Section 4, we verify our theoretical findings on synthetic experiments. We also show by simulation that the entropic estimator appears to perform well even outside the semi-discrete setting, suggesting it as a promising choice for estimating other types of discontinuous maps.

NOTATION

The Euclidean ball centered at a with radius $r > 0$ is written as $B(a; r)$. The symbols C and c denote positive constants whose value may change from line to line. Write $a \lesssim b$ and $a \asymp b$ if there exist constants $c, C > 0$ such that $a \le Cb$ and $cb \le a \le Cb$, respectively. For an integer $N \in \mathbb{N}$, we let $[N] := \{1, \dots, N\}$. For a function f and a probability measure ρ, we write $\|f\|^2_{L^2(\rho)} := \mathbb{E}_{X\sim\rho}\|f(X)\|^2$. Similarly, we write $\mathrm{Var}_\rho(f) := \mathbb{E}_{X\sim\rho}[(f(X) - \mathbb{E}_{X\sim\rho}[f(X)])^2]$ for the variance of f with respect to ρ.

2. Background on optimal transport

2.1. Optimal transport

We define $\mathcal{P}(\Omega)$ to be the space of probability measures whose support lies in a compact subset $\Omega \subseteq \mathbb{R}^d$. If a probability measure P has a density with respect to the Lebesgue measure on $\mathbb{R}^d$ with support $\Omega \subseteq \mathbb{R}^d$, then we write $P \in \mathcal{P}_{\text{ac}}(\Omega)$.

For two probability measures $P, Q \in \mathcal{P}(\Omega)$, we define the *(squared) 2-Wasserstein distance* to be (Kantorovitch, 1942)

$$\mathrm{S}_0(P, Q) := \min_{\pi\in\Gamma(P,Q)} \iint \tfrac{1}{2}\|x - y\|^2 \,\mathrm{d}\pi(x, y)\,, \tag{4}$$

where $\pi \in \Gamma(P, Q) \subseteq \mathcal{P}(\Omega \times \Omega)$ such that for any event A,

$$\pi(A \times \Omega) = P(A)\,, \quad \pi(\Omega \times A) = Q(A)\,.$$

We call $\Gamma(P, Q)$ the set of *couplings* between P and Q. In this work, we focus on the squared-Euclidean cost but Equation (4) is well-defined for convex, lower-semicontinuous costs; see (Santambrogio, 2015; Villani, 2009) for more information on optimal transport under general costs.

Equation (4) is a convex optimization problem on the space of joint measures, and a minimizer, denoted π_0, always exists; we call π_0 an *optimal plan* from P to Q. Moreover, Equation (4) possesses the following dual formulation,

$$\begin{aligned}\mathrm{S}_0(P, Q) = \tfrac{1}{2}M_2(P) + \tfrac{1}{2}M_2(Q)\\ - \inf_{(\varphi,\psi)\in\Phi} \int \varphi \,\mathrm{d}P + \int \psi \,\mathrm{d}Q\end{aligned} \tag{5}$$

where $M_2(P) := \int \|x\|^2 \,\mathrm{d}P(x)$ (similarly for $M_2(Q)$) and the functions $(\varphi, \psi) \in \Phi \subseteq L_1(P) \times L_1(Q)$ satisfy

$$\langle x, y\rangle \le \varphi(x) + \psi(y) \text{ for all } x, y \in \Omega\,,$$

As with the primal formulation, the infimum in Equation (5) is attained at functions (φ_0, ψ_0). These minimizers are called *(optimal) Brenier potentials*. In particular, at optimality, we have that these Brenier potentials are convex conjugates of one another, i.e. the Legendre transform of one of the potentials gives the other:

$$\varphi_0^*(y) := \sup_x\{\langle x, y\rangle - \varphi_0(x)\} = \psi_0(y)\,, \tag{6}$$

and vice-versa.

Apart from these two formulations of optimal transport under the squared-Euclidean cost, there exists a third, known as the Monge problem:

$$T_0 := \operatorname*{argmin}_{T\in\mathcal{T}(P,Q)} \int \tfrac{1}{2}\|x - T(x)\|^2 \,\mathrm{d}P(x)\,, \tag{7}$$

where $\mathcal{T}(P, Q)$ is the set of admissible transport maps, i.e. for $X \sim P$, $T(X) \sim Q$. This optimization problem is non-convex in T, and a solution is not always guaranteed to exist for arbitrary P and Q.

The following theorem unifies these three formulations of optimal transport under the squared-Euclidean cost:

Theorem 2.1 (Brenier's theorem; Brenier, 1991)**.** *Let $P \in \mathcal{P}_{ac}(\Omega)$ and let $Q \in \mathcal{P}(\Omega)$, then*

1. *the solution to Equation (7) exists and is of the form $T_0 = \nabla\varphi_0$, where φ_0 solves Equation (5)*
2. *π_0 is also uniquely defined as*

$$\mathrm{d}\pi_0(x, y) = \mathrm{d}P(x)\delta_{\{\nabla\varphi_0(x)\}}(y)\,.$$

When we want to place emphasis on the underlying measures, we will write $\varphi_0 = \varphi_0^{P\to Q}$, $\psi_0 = \psi_0^{P\to Q}$ and $T_0 = T_0^{P\to Q}$.

2.1.1. OT IN THE SEMI-DISCRETE CASE

In optimal transport, the semi-discrete setting refers to the case where P has as density with respect to the Lebesgue measure on $\mathbb{R}^d$, and Q is a discrete measure supported on points. The following theorem characterizes the optimal transport map in this situation, which exhibits a particular structure compared to the general results in the previous section. Let $[J] = \{1, \dots, J\}$.

Proposition 2.2 (Aurenhammer et al., 1998)**.** *If $P \in \mathcal{P}_{ac}(\Omega)$ and Q is a discrete measure supported on the points $y_1, \dots, y_J$, then the optimal transport map $\nabla\varphi_0$ is given by*

$$\nabla\varphi_0(x) \coloneqq \operatorname*{argmax}_{j\in[J]}\{\langle x, y_j\rangle - \psi_0(y_j)\}\,, \tag{8}$$

where ψ_0 is the dual to φ_0 in the sense of Equation (6).

Here, the optimal dual Brenier potential ψ_0 can be identified with a *vector* in $\mathbb{R}^J$, defined by the number of atoms, and the optimal Brenier potential is consequently given by

$$\varphi_0 \coloneqq \max_{j\in[J]}\{\langle x, y_j\rangle - \psi_0(y_j)\}\,.$$

Although φ_0 is not differentiable, only subdifferentiable, we still use the gradient notation as $\nabla\varphi_0$ is well-defined P-almost everywhere.

The map $\nabla\varphi_0$ partitions the space into J convex polytopes $L_j \coloneqq \nabla\varphi_0^{-1}(\{y_j\})$ called *Laguerre cells*; recall Figure 1. From this definition, it is clear that for a given $x \in L_j$, $x \mapsto \nabla\varphi_0(x) = y_j$ is the optimal transport mapping. The difficulty in finding this map lies in determining the cells L_j, or equivalently the dual variables $\psi_0(y_j)$.

2.2. Entropic optimal transport

Entropic regularization was introduced to both optimal transport and machine learning communities in the seminal paper by Cuturi (2013), allowing approximate optimal transport distances to be computed at unprecedented speeds. Entropic optimal transport (EOT) is defined as the following regularized version of Equation (4): for $\varepsilon > 0$

$$\begin{aligned}\mathrm{S}_\varepsilon(P, Q) \coloneqq \min_{\pi\in\Gamma(P,Q)} &\iint \tfrac{1}{2}\|x - y\|^2\,\mathrm{d}\pi(x, y)\\ &+ \varepsilon\mathrm{KL}(\pi\|P\otimes Q)\,,\end{aligned} \tag{9}$$

where $\mathrm{KL}(\mu\|\nu) = \int \log\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\mu$ when $\mu \in \mathcal{P}(\Omega)$ is absolutely continuous with respect to $\nu \in \mathcal{P}(\Omega)$. This speedup is due to the elegant connection of (9) to Sinkhorn's algorithm; we refer the interested reader to (Peyré & Cuturi, 2019, Chapter 4) for more information. The computational tractability of S_ε compared to S_0 when dealing with many samples lends itself to being a central object of study in its own right (see, e.g., Chizat et al., 2020; Genevay et al., 2019; Gonzalez-Sanz et al., 2022; Mena & Niles-Weed, 2019; Rigollet & Stromme, 2022).

Equation (9) admits the following dual formulation, which is now an unconstrained optimization problem (Genevay, 2019; Marino & Gerolin, 2020)

$$\begin{aligned}&\mathrm{S}_\varepsilon(P, Q) =\\ &\tfrac{1}{2}M_2(P) + \tfrac{1}{2}M_2(Q) - \inf_{\varphi,\psi}\Bigg(\int\varphi\,\mathrm{d}P + \int\psi\,\mathrm{d}Q\\ &+ \varepsilon\iint(e^{(\langle x,y\rangle-\varphi(x)-\psi(y))/\varepsilon} - 1)\,\mathrm{d}P(x)\,\mathrm{d}Q(y)\Bigg),\end{aligned} \tag{10}$$

where $(\varphi, \psi) \in L_1(P)\times L_1(Q)$. When P and Q have finite second moments, Equation (9) admits a *unique* minimizer, π_ε and we have the existence of minimizers to Equation (10), which we denote as $(\varphi_\varepsilon, \psi_\varepsilon)$. We call π_ε the *entropic optimal plan* and $(\varphi_\varepsilon, \psi_\varepsilon)$ are called *entropic Brenier potentials*. The following optimality relation further relates these primal and dual solutions (Csiszár, 1975):

$$\mathrm{d}\pi_\varepsilon(x, y) \coloneqq e^{(\langle x,y\rangle-\varphi_\varepsilon(x)-\psi_\varepsilon(y))/\varepsilon}\,\mathrm{d}P(x)\,\mathrm{d}Q(y)\,.$$

As a consequence, the following relationship holds at optimality:

$$\mathrm{S}_\varepsilon(P, Q) = \tfrac{1}{2}M_2(P) + \tfrac{1}{2}M_2(Q) - \int\varphi_\varepsilon\,\mathrm{d}P - \int\psi_\varepsilon\,\mathrm{d}Q\,,$$

and, moreover, we can define versions of φ_ε and ψ_ε such that the following relationships hold (see Mena & Niles-Weed, 2019; Nutz & Wiesel, 2022) over all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$, respectively:

$$\varphi_\varepsilon(x) = \varepsilon\log\int e^{(\langle x,y\rangle-\psi_\varepsilon(y))/\varepsilon}\,\mathrm{d}Q(y)\,, \tag{11}$$

$$\psi_\varepsilon(y) = \varepsilon\log\int e^{(\langle x,y\rangle-\varphi_\varepsilon(x))/\varepsilon}\,\mathrm{d}P(x)\,, \tag{12}$$

which are smoothed version of the Legendre transform, see Appendix A for details. In what follows, we always assume that we have selected φ_ε and ψ_ε so that these identities hold.

2.2.1. Entropic Brenier Map

If $(X, Y) \sim \pi_\varepsilon$, we may define the conditional probability π_ε^x of Y given that $X = x$, with density

$$\frac{\mathrm{d}\pi_\varepsilon^x}{\mathrm{d}Q}(y) \propto \exp\left((\langle x, y\rangle - \psi_\varepsilon(y))/\varepsilon\right) . \tag{13}$$

The barycentric projection of the optimal entropic coupling π_ε, or *entropic Brenier map*, is a central object of study in several works e.g. (del Barrio et al., 2022b; Goldfeld et al., 2022; Pooladian & Niles-Weed, 2021; Rigollet & Stromme, 2022), defined as

$$T_\varepsilon(x) = \int y \, \mathrm{d}\pi_\varepsilon^x(y) = \nabla\varphi_\varepsilon(x) , \tag{14}$$

where π_ε^x is as in Equation (13). Note that this quantity is well defined for all $x \in \mathbb{R}^d$ as long as the source and target measures have compact support; in particular, it applies to both discrete and continuous measures. The second equality follows from Equation (11) and the dominated convergence theorem. As in the unregularized case, we will write $\varphi_\varepsilon = \varphi_\varepsilon^{P\to Q}$, $\psi_\varepsilon = \psi_\varepsilon^{P\to Q}$ and $T_\varepsilon = T_\varepsilon^{P\to Q}$ when we want to emphasize on the dependency with respect to the underlying measures.

This particular barycentric projection was proposed as a tool for large-scale optimal transport by Seguy et al. (2018), but analyzed statistically for the first time by Pooladian & Niles-Weed (2021) as an estimator for the optimal transport map. We mention some of their results to highlight the differences with our new results for the semi-discrete setting in Section 3. First, they prove the following approximation result for T_ε.

Proposition 2.3 (Pooladian & Niles-Weed, 2021, Corollary 1)**.** *Let P, Q be compactly supported absolutely continuous measures on a compact set $\Omega \subseteq \mathbb{R}^d$ with densities p and q, that are bounded away from 0 and ∞. Assume that φ_0 is smooth and strongly convex, and that φ_0^* is at least $\mathcal{C}^3$. Then,*

$$\|T_\varepsilon - \nabla\varphi_0\|_{L^2(P)}^2 \lesssim \varepsilon^2 . \tag{15}$$

Their main statistical result is the following theorem:

Proposition 2.4 (Pooladian & Niles-Weed, 2021, Theorem 3)**.** *Suppose the same assumptions as Proposition 2.3, and let P_n and Q_n denote the empirical measures of P and Q constructed from i.i.d. samples. Let $\hat{T}_\varepsilon = T_\varepsilon^{P_n\to Q_n}$ denote the entropic Brenier map from P_n to Q_n and let $T_0 = \nabla\varphi_0$ be the optimal transport map from P to Q. Then, if $\varepsilon \asymp n^{-\frac{1}{d'+3}}$*

$$\mathbb{E}\|\hat{T}_\varepsilon - T_0\|_{L^2(P)}^2 \lesssim n^{-\frac{3}{2(d'+3)}} \log(n) , \tag{16}$$

where $d' = 2\lceil d/2\rceil$.

Note that in particular the the rate of convergence of the entropic estimator critically depends on the ambient dimension d in the continuous-to-continuous case.

2.2.2. Related work

Characterizing the convergence of entropic objects (e.g. potentials, cost, plans) to their unregularized counterparts in the $\varepsilon \to 0$ regime has been a topic of several works in recent years. Convergence of the costs S_ε to S_0 with precise rates was investigated in (Chizat et al., 2020; Conforti & Tamanini, 2021; Pal, 2019). The works (Bernton et al., 2022; Carlier et al., 2017; Ghosal et al., 2022; Léonard, 2012) study the convergence of the minimizers π_ε to π_0 under varying assumptions. Convergence of the potentials in a very general setting was established in (Nutz & Wiesel, 2022), though without a rate of convergence in ε. In the semi-discrete case, this gap was closed in (Altschuler et al., 2022) followed closely by (Delalande, 2022), which gave non-asymptotic rates. The Sinkhorn Divergence, a non-negative, symmetric version of S_ε, was introduced in (Genevay et al., 2018), was statistically analysed in (Goldfeld et al., 2022) and also in (del Barrio et al., 2022b; Gonzalez-Sanz et al., 2022), and was connected to the entropic Brenier map in (Pooladian et al., 2022). The recent pre-print by (Rigollet & Stromme, 2022) proved parametric rates of estimation between the empirical entropic Brenier map and its population counterpart, though with an exponentially poor dependence on the regularization parameter (see Remark 3.8). Using covariance inequalities, the entropic Brenier potentials were used give a new proof of Caffarelli's contraction theorem; see (Chewi & Pooladian, 2022); this approach was recently generalized in (Conforti, 2022). Entropic optimal transport has also come into contact with the area of deep generative modelling through the following works (De Bortoli et al., 2021; Finlay et al., 2020), among others.

3. Statistical performance of the entropic estimator in the semi-discrete setting

Let P_n and Q_n be the empirical measures associated with two n-samples from P and Q. We make the following regularity assumptions on P, already introduced by Delalande (2022).

(A) The measure P has a compact convex support $\Omega \subseteq B(0; R)$, with a density p satisfying $0 < p_{\min} \le p \le p_{\max} < \infty$ for positive constants $p_{\min}$, $p_{\max}$ and R.

For example, P can be the uniform distribution over Ω, or a truncated Gaussian distribution. Furthermore, we will need the following assumption on Q.

(B) The discrete probability measure $Q = \sum_{j=1}^{J} q_j \delta_{y_j}$ is such that $q_j \geq q_{\min} > 0$ and $y_j \in B(0; R)$ for all $j \in [J]$.

The goal of this section is to prove the following theorem:

Theorem 3.1. *Let P satisfy **(A)** and let Q satisfy **(B)**. Let $\hat{T}_\varepsilon = T_\varepsilon^{P_n \to Q_n}$. Then, for $\varepsilon \asymp n^{-1/2}$ and n large enough,*

$$\mathbb{E}\|\hat{T}_\varepsilon - T_0\|^2_{L^2(P)} \lesssim n^{-1/2} \,. \tag{17}$$

Remark 3.2. We remark that the hidden constants in Theorem 3.7 and related results depend on $J, p_{\min}, p_{\max}, q_{\min}$ and R.

Remark 3.3 (Fixing the support via rounding). At present, the entropic map need not necessarily map exactly to one of $\{y_1, \ldots, y_J\}$. In fact, $\hat{T}_\varepsilon : \mathbb{R}^d \to \mathrm{conv}(\{Y_1, \ldots, Y_n\})$, where $\mathrm{conv}(A)$ is the convex hull for some set A. In turn, the support of the entropic map does not in general match that of Q. However, this can be readily fixed with a rounding scheme. We can replace our estimator by $\bar{T}_\varepsilon$ which is obtained by mapping the output of $\hat{T}_\varepsilon$ to its nearest neighbor in the support of Q – this projection step is easy to compute, given that we essentially know the support of Q via samples. By viewing this as a projection onto an appropriate set (namely, the set of transport maps with codomain equal to the support of Q), and applying the triangle inequality, it holds that

$$\mathbb{E}\|\bar{T}_\varepsilon - T_0\|^2_{L^2(P)} \leq 2\mathbb{E}\|\hat{T}_\varepsilon - T_0\|^2_{L^2(P)}$$

but $\bar{T}_\varepsilon$ matches the support of Q.

Let $T_\varepsilon = T_\varepsilon^{P \to Q}$ denote the entropic Brenier map associated to P and Q. Our proof relies on the following bias-variance decomposition:

$$\mathbb{E}\|\hat{T}_\varepsilon - T_0\|^2_{L^2(P)} \lesssim \mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|^2_{L^2(P)} + \|T_\varepsilon - T_0\|^2_{L^2(P)}.$$

Following the next two results (Theorem 3.4 and Theorem 3.7) and the preceding decomposition, the proof of Theorem 3.1 is merely a balancing act in the regularization parameter ε.

Theorem 3.4. *Let P satisfy **(A)** and let Q satisfy **(B)**. Then, for ε small enough,*

$$\|T_\varepsilon - T_0\|^2_{L^2(P)} \lesssim \varepsilon \,. \tag{18}$$

The proof of Theorem 3.4 relies on the following qualitative picture: if a point x belongs to some Laguerre cell L_j, and is far away from the boundary of L_j, then the entropic optimal plan π_ε will send almost all of its mass towards the point $y_j = T_0(x)$, sending an exponentially small amount of mass to the other points y_j. Such a picture is correct as long as x is at distance at least ε from the boundary of the Laguerre cell L_j, incurring a total error of order ε. A rigorous proof of Theorem 3.4 can be found in Appendix B.

Note that this rate is slower than the rate appearing in Proposition 2.3 in the continuous-to-continuous case. The following example shows that the dependency in ε is optimal in Theorem 3.4, indicating that the presence of discontinuities necessarily affects the approximation properties of the entropic Brenier map.

Example 3.5. Let P be a probability measure on $\mathbb{R}$ having a symmetric bounded density p continuous at 0, and let $Q = \frac{1}{2}(\delta_{-1} + \delta_1)$. Following (Altschuler et al., 2022, Section 3), one can check that the entropic Brenier map in this setting is the following scaled sigmoidal function

$$T_\varepsilon(x) = \tanh(2x/\varepsilon) \,,$$

whereas the optimal transport map $T_0(x) = \mathrm{sign}(x)$. Then, performing a computation

$$\begin{aligned}\|T_\varepsilon - T_0\|^2_{L^2(P)} &= 2\int_0^\infty (1 - \tanh(2x/\varepsilon))^2 p(x)\,\mathrm{d}x \\ &= \varepsilon \int_0^\infty (1 - \tanh(u))^2 p(u\varepsilon/2)\,\mathrm{d}u \\ &= \varepsilon p(0)(\log(4) - 1) + o(\varepsilon) \,,\end{aligned}$$

where in the last step we invoked the dominated convergence theorem, and computed the limiting integral.

Remark 3.6. Assumption **(A)** can be relaxed for Theorem 3.4 to hold. More precisely, it can be replaced by Assumptions 2.2 and 2.9 of Altschuler et al. (2022), that hold for unbounded measures such as the normal distribution.

Finally, we present the sample-complexity result:

Theorem 3.7. *Let P satisfy **(A)** and let Q satisfy **(B)**. Then, for $0 < \varepsilon \leq 1$ such that $\log(1/\varepsilon) \lesssim n/\log(n)$*

$$\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|^2_{L^2(P)} \lesssim \varepsilon^{-1} n^{-1} \,. \tag{19}$$

Remark 3.8. In (Rigollet & Stromme, 2022), the authors show that if P and Q are merely compactly supported with $\mathrm{supp}(P), \mathrm{supp}(Q) \subseteq B(0; R)$, then

$$\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon\|^2_{L^2(P)} \lesssim e^{cR^2/\varepsilon} \varepsilon^{-1} n^{-1} \,, \tag{20}$$

where $c > 0$ is some absolute positive constant. Thus, under the additional structural assumptions of the semi-discrete formulation, we are able to significantly improve the rate of convergence between the empirical and population entropic Brenier maps.

The proof of Theorem 3.7 relies on a novel stability result, reminiscent of (Manole et al., 2021, Theorem 6), which is of independent interest. We provide the proof in Appendix C.

Proposition 3.9. *Let μ, ν, μ', ν' be four probability measures supported in $B(0;R)$. Then the entropic maps $T_\varepsilon^{\mu\to\nu}$ and $T_\varepsilon^{\mu'\to\nu'}$ satisfy*

$$\frac{\varepsilon}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu'\to\nu'}\|^2_{L^2(\mu)} \leq \int(\varphi_\varepsilon^{\mu'\to\nu'} - \varphi_\varepsilon^{\mu\to\nu})\,\mathrm{d}\mu + \int(\psi_\varepsilon^{\mu'\to\nu'} - \psi_\varepsilon^{\mu\to\nu})\,\mathrm{d}\nu + \varepsilon KL(\nu\|\nu')$$

Remark 3.10. The right side of the bound in Proposition 3.9 is equal to

$$S_\varepsilon(\mu,\nu) - S_\varepsilon(\mu',\nu') + \int f_\varepsilon^{\mu'\to\nu'}\,\mathrm{d}(\mu'-\mu) + \int g_\varepsilon^{\mu'\to\nu'}\,\mathrm{d}(\nu'-\nu) + \varepsilon\mathrm{KL}(\nu\|\nu')\,,$$

where $f_\varepsilon^{\mu'\to\nu'} = \frac{1}{2}\|\cdot\|^2 - \varphi_\varepsilon^{\mu'\to\nu'}$ and $g_\varepsilon^{\mu'\to\nu'} = \frac{1}{2}\|\cdot\|^2 - \psi_\varepsilon^{\mu'\to\nu'}$. Proposition 3.9 is therefore the entropic analogue of the stability bounds of Manole et al. (2021, Theorem 6) and Ghosal & Sen (2022, Lemma 5.1). Unlike those results, Proposition 3.9 allows both the source and target measure to be modified, and does not require any smoothness assumptions.

Proof sketch of Theorem 3.7

To prove Theorem 3.7, we first consider the *one-sample setting*, where we assume that we only have access to samples $Y_1,\dots,Y_n \sim Q$, but we have full access to P. We then consider the one-sample entropic estimator $T_\varepsilon^{P\to Q_n}$. We apply Proposition 3.9 with $\mu = \mu' := P$, $\nu := Q_n$ and $\nu' := Q$, yielding (see Corollary C.1 for details)

$$\frac{\varepsilon}{8R^2}\mathbb{E}\|T_\varepsilon^{P\to Q_n} - T_\varepsilon\|^2_{L^2(\mu)} \leq \mathbb{E}\Big(\int(\psi_\varepsilon - \psi_\varepsilon^{P\to Q_n})\,\mathrm{d}(Q_n - Q) + \varepsilon\mathrm{KL}(Q_n\|Q)\Big)$$

Let $\chi^2(P\|Q)$ denote the χ^2-divergence between probability measure. Young's inequality (see Lemma H.1) and the inequality $\mathrm{KL}(Q_n\|Q) \leq \chi^2(Q_n\|Q)$ yield the following bound:

$$\mathbb{E}\|T_\varepsilon^{P\to Q_n} - T_\varepsilon\|^2_{L^2(P)} \leq \frac{8R^2}{\varepsilon}\Big(\frac{\mathbb{E}[\mathrm{Var}_Q(\psi_\varepsilon^{P\to Q_n} - \psi_\varepsilon)]}{2} + \frac{\mathbb{E}[\chi^2(Q_n\|Q)]}{2}\Big) + 8R^2\mathbb{E}[\chi^2(Q_n\|Q)]\,.$$

To complete our proof sketch, we use a new stability result on the entropic dual Brenier potentials, catered for the semi-discrete setting.

Proposition 3.11. *Let μ be a measure that satisfies **(A)**. Let ν, ν' be two discrete probability measures supported on $\{y_1,\dots,y_J\}$, with $\nu' \geq \lambda\nu$ for some $\lambda > 0$. Then, for $0 < \varepsilon \leq 1$,*

$$\mathrm{Var}_\nu(\psi_\varepsilon^{\mu\to\nu'} - \psi_\varepsilon^{\mu\to\nu}) \leq \frac{C}{\lambda^2}\chi^2(\nu'\|\nu), \tag{21}$$

where C depends on R, $p_{\min}$ and $p_{\max}$.

Moreover, a computation provided in Lemma H.2 shows that $\mathbb{E}[\chi^2(Q_n\|Q)] = \frac{J-1}{n}$, which is enough to conclude the proof of the one-sample case, see Appendix E for details. The two-sample setting is tackled using similar reasoning, where we ultimately prove in Appendix F that the risk $\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon^{P\to Q_n}\|^2_{L^2(P)}$ is upper bounded by

$$\frac{8R^2}{\varepsilon}\mathbb{E}\int(\varphi_\varepsilon^{P\to Q_n} - \varphi_\varepsilon^{P_n\to Q_n})\,\mathrm{d}(P_n - P)\,.$$

Such a quantity can again be related to the estimation of the dual potentials $\psi_\varepsilon^{P\to Q_n}$ and $\psi_\varepsilon^{P_n\to Q_n}$. Using the same reasoning as before, we expect a parametric rate of convergence for this term as well. Merging the two results completes the proof of Theorem 3.7. We refer to Appendix F for full details.

4. Comparing against the 1NN estimator

4.1. Rate optimality of the entropic Brenier map

The upper bound of Theorem 3.7 shows that our estimator achieves the $n^{-1/2}$ rate. In fact, the following simple proposition tells us that this rate is optimal in the semi-discrete case.

Proposition 4.1. *Let P be the uniform distribution on $[-1/2,1/2]^d$ and for any $J \geq 2$, let $\mathcal{Q}_J$ denote the space of of probability measures with at most J atoms, supported on $[-1/2,1/2]^d$. Define the minimax rate of estimation*

$$\mathcal{R}_n(\mathcal{Q}_J) = \inf_{\hat{T}}\sup_{Q\in\mathcal{Q}_J}\mathbb{E}_{Q^n}[\|\hat{T} - T_0^{P\to Q}\|^2_{L_2(P)}]\,.$$

Then, it holds that $\mathcal{R}_n(\mathcal{Q}_J) \geq n^{-1/2}/64$.

Proof. Let e be a vector of the canonical basis of $\mathbb{R}^d$, scaled by $1/2$. Fix $0 < r < 1/2$ and let $Q_0 = \frac{1}{2}\delta_{-e} + \frac{1}{2}\delta_e$ and $Q_1 = (\frac{1}{2} - r)\delta_{-e} + (\frac{1}{2} + r)\delta_e$. A computation gives $\|T_0^{P\to Q_0} - T_0^{P\to Q_1}\|^2_{L_2(P)} = r$. Therefore, by Le Cam's lemma (see, e.g., Wainwright, 2019, Chapter 15),

$$\mathcal{R}_n(\mathcal{Q}_{J,R}) \geq \frac{r}{8}(1 - \mathrm{d}_{\mathrm{TV}}(Q_0^n, Q_1^n)). \tag{22}$$

Let $\mathrm{d}_{\mathrm{H}^2}(Q_0,Q_1)$ denote the (squared) Hellinger distance between measures. We have $\mathrm{d}_{\mathrm{TV}}(Q_0^n,Q_1^n)^2 \leq$

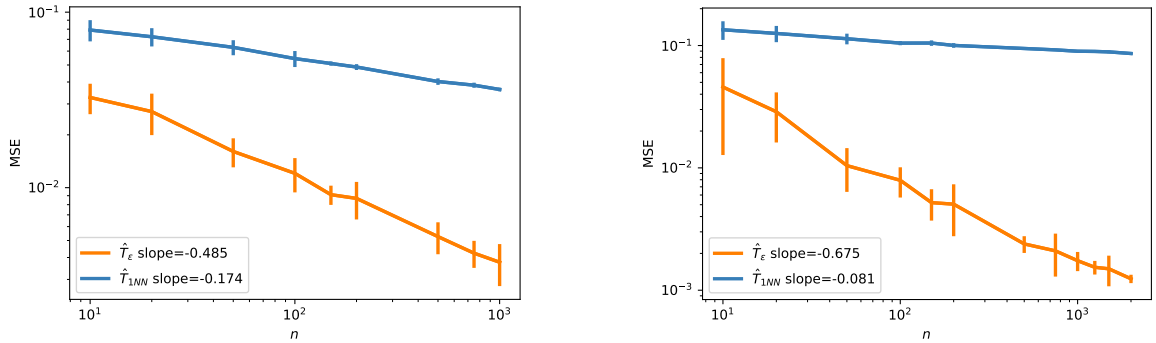

Figure 2. Left: $\hat{T}_\varepsilon$ versus $\hat{T}_{\text{1NN}}$ for $J = 2$ and $d = 10$. Right: $\hat{T}_\varepsilon$ versus $\hat{T}_{\text{1NN}}$ for $J = 10$ and $d = 50$.

$\mathrm{d}_{\mathrm{H}^2}(Q_0^n, Q_1^n) \leq n\mathrm{d}_{\mathrm{H}^2}(Q_0, Q_1)$. Furthermore, a computation gives

$$\begin{aligned}\mathrm{d}_{\mathrm{H}^2}(Q_0, Q_1) &= \left(\sqrt{\tfrac{1}{2} - r} - \sqrt{\tfrac{1}{2}}\right)^2 + \left(\sqrt{\tfrac{1}{2} + r} - \sqrt{\tfrac{1}{2}}\right)^2 \\ &= 2 - (\sqrt{1+2r} + \sqrt{1-2r}) \\ &\leq 4r^2.\end{aligned}$$

We obtain the conclusion by picking $r = n^{-1/2}/4$. □

4.2. The 1NN estimator is proveably suboptimal

The 1-Nearest-Neighbor estimator, henceforth denoted $\hat{T}_{\text{1NN}}$, was proposed by (Manole et al., 2021) as a computational surrogate for estimating optimal transport maps in the low smoothness regime. Written succinctly, their estimator is $\hat{T}_{\text{1NN}}(x) = \sum_{i=1}^n \mathbf{1}_{V_i}(x) Y_{\hat{\pi}(i)}$, where $(V_i)_{i=1}^n$ are Voronoi regions i.e.

$$V_i \coloneqq \{x \in \mathbb{R}^d \,:\, \|x - X_i\| \leq \|x - X_k\|, \forall\, k \neq i\},$$

and $\hat{\pi}$ is the optimal transport plan between the empirical measures P_n and Q_n, which amounts to a permutation. Computing the closest X_i to a new sample x has runtime $\mathcal{O}(n\log(n))$, though the complexity of this estimator is determined by computing the plan $\hat{\pi}$, which takes $\mathcal{O}(n^3)$ time via, e.g., the Hungarian Algorithm (see Peyré & Cuturi, 2019, Chapter 3).

When φ_0 is smooth and strongly convex, Manole et al. (2021) showed that, for $d \geq 5$,

$$\mathbb{E}\|\hat{T}_{\text{1NN}} - \nabla\varphi_0\|_{L^2(P)}^2 \lesssim n^{-2/d}.$$

In contrast to the rate optimality of the entropic Brenier map, we now show that $\hat{T}_{\text{1NN}}$ is proveably suboptimal in the semi-discrete setting. Not only does it fail to recover the minimax rate obtained by the entropic Brenier map, but its performance in fact degrades in comparison to the smooth case. A proof appears in Appendix G.

Proposition 4.2. *There exist a measure P satisfying **(A)** and a discrete measure Q satisfying **(B)** such that for $d \geq 3$*

$$\mathbb{E}\|\hat{T}_{1NN} - T_0^{P\to Q}\|_{L^2(P)}^2 \gtrsim n^{-1/d}.$$

4.3. Experiments

We briefly verify our theoretical findings on synthetic experiments. To create the following plots, we draw two sets of n i.i.d. points from P, $(X_1, \ldots, X_n)$ and $(X_1', \ldots, X_n')$, and create target points $Y_i = T_0(X_i')$, where T_0 is known to us in advance in order to generate the data. Our estimators are computed on the data $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, and we evaluate the Mean-Squared error criterion

$$\mathrm{MSE}(\hat{T}) = \|\hat{T} - T_0\|_{L^2(P)}^2$$

of a given map estimator $\hat{T}$ using Monte Carlo integration, using 50000 newly sampled points from P. We plot the means across 10 repeated trials, accompanied by their standard deviations.

4.3.1. Semi-discrete example #1

First consider $P = \mathrm{Unif}([0,1]^d)$ and create atoms $\{y_1, \ldots, y_J\}$ by partitioning the points along the first coordinate for all $j \in [J]$:

$$(y_j)[1] = \frac{(j - 1/2)}{J}, \quad (y_j)[2] = \cdots = (y_j)[d] = 0.5.$$

We choose uniform $q_j = 1/J$ for $j \in [J]$. In this case, it is easy to see that the optimal transport map $T_0(x)$ is uniquely defined by the first coordinate of x_1. Figure 2 illustrates the rate-optimal performance of the entropic Brenier map, and the proveably suboptimal performance of the 1-Nearest-Neighbor estimator.

4.3.2. Semi-discrete example #2

We now consider a synthetic experiment with far less symmetry. Let $P = \mathrm{Unif}([0,1]^d)$, and fix $J \in \mathbb{N}$. We randomly generate $y_1, \ldots, Y_J \in [0,1]^d$, and also randomly generate $\psi_0 \in \mathbb{R}^J$, and consider the optimal transport map $T_0(x) = \mathrm{argmin}_{j\in[J]}\{x^\top y_j - (\psi_0)_j\}$. We define $Q = (T_0)_\sharp P$, leading to the same setup as before, but with a less structured optimal transport map. We consider $J = 5$ and $d = 50$, and repeat the procedure of the preceding

section to generate our data, and the resulting estimator. Figure 3 contains plots the MSE as a function of n, where again we see a log-linear slope of around -0.5, which agrees with our theory.

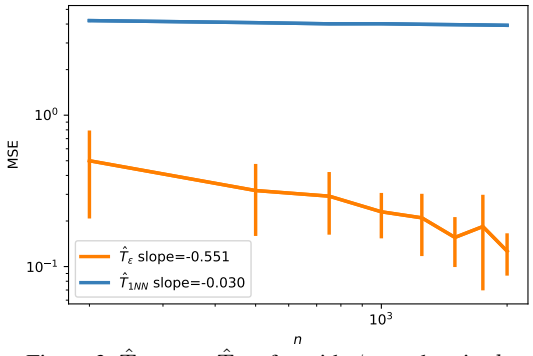

Figure 3. $\hat{T}_\varepsilon$ versus $\hat{T}_{1\text{NN}}$ for with ψ_0 random in $d = 50$

4.3.3. Discontinuous example

We turn our attention to a discontinuous transport map, where for $x \in \mathbb{R}^d$, all the coordinates are fixed except for the first one

$$T_0(x) = 2\text{sign}(x[1]) \otimes x[2] \otimes \cdots \otimes x[d]\,.$$

We choose $P = \text{Unif}([-1, 1]^d)$ to exhibit a discontinuity in the data. Focusing on $d = 10$, we see in Figure 4 that the entropic map estimator avoids the curse of dimensionality and enjoys a faster convergence rate, with better constants.

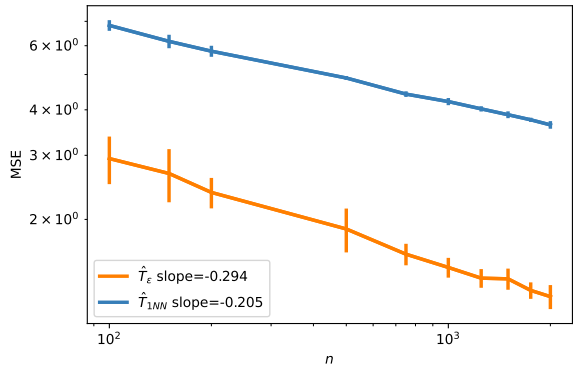

Figure 4. $\hat{T}_\varepsilon$ versus $\hat{T}_{1\text{NN}}$ for $d = 10$

5. Conclusion

Understanding optimal transport maps in the semi-discrete case is a natural stepping-stone to understanding the case for general discontinuous transport maps. In this work, we propose a tractable, minimax optimal estimator of the Brenier map in the semi-discrete setting, where the rate of estimation is dimension independent. To prove our result, we require several new results and techniques, and, as a by-product of our analysis, give the first parametric rates of estimation the entropic Brenier map, without exponential dependence in the regularization parameter. Our synthetic experiments indicate that the entropic Brenier map might be useful in estimating other variants of discontinuous transport maps, which constitutes an interesting direction for future research.

Acknowledgements

AAP would like to thank Tudor Manole for fruitful discussions, and gratefully thanks funding sources NSF Award 1922658, and Meta AI Research. JNW is supported by the Sloan Research Fellowship and NSF grant DMS-2210583. We thank the anonymous reviewer for suggesting the addition of a rounding scheme.

References

Altschuler, J. M., Niles-Weed, J., and Stromme, A. J. Asymptotics for semidiscrete entropic optimal transport. *SIAM Journal on Mathematical Analysis*, 54(2):1718–1741, 2022.

Aurenhammer, F., Hoffmann, F., and Aronov, B. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

Bernton, E., Ghosal, P., and Nutz, M. Entropic optimal transport: Geometry and large deviations. *Duke Mathematical Journal*, 171(16):3363–3400, 2022.

Bobkov, S. G. and Götze, F. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *J. Funct. Anal.*, 163(1):1–28, 1999. ISSN 0022-1236. doi: 10.1006/jfan.1998.3326.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Brown, B. C., Caterini, A. L., Ross, B. L., Cresswell, J. C., and Loaiza-Ganem, G. The union of manifolds hypothesis and its implications for deep generative modelling. *arXiv preprint arXiv:2207.02862*, 2022.

Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional Monge maps. *arXiv preprint arXiv:2206.14262*, 2022.

Carlier, G., Chernozhukov, V., and Galichon, A. Vector quantile regression: an optimal transport approach. *The Annals of Statistics*, 44(3):1165–1192, 2016.

Carlier, G., Duval, V., Peyré, G., and Schmitzer, B. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

Chen, R. T. Q., Amos, B., and Nickel, M. Semi-discrete normalizing flows through differentiable tessellation. In *Advances in Neural Information Processing Systems*, 2022.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223–256, 2017.

Chewi, S. and Pooladian, A.-A. An entropic generalization of Caffarelli's contraction theorem via covariance inequalities. *arXiv preprint arXiv:2203.04954*, 2022.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

Conforti, G. Weak semiconvexity estimates for Schrödinger potentials and logarithmic Sobolev inequality for Schrödinger bridges. *arXiv preprint arXiv:2301.00083*, 2022.

Conforti, G. and Tamanini, L. A formula for the time derivative of the entropic cost and applications. *Journal of Functional Analysis*, 280(11):108964, 2021.

Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3: 146–158, 1975.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Deb, N., Ghosal, P., and Sen, B. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

del Barrio, E. and Loubes, J.-M. Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.*, 47(2):926–951, 2019. ISSN 0091-1798. doi: 10.1214/18-AOP1275.

del Barrio, E., González-Sanz, A., and Loubes, J.-M. Central limit theorems for semidiscrete Wasserstein distances. *arXiv preprint arXiv:2202.06380*, 2022a.

del Barrio, E., Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. An improved central limit theorem and fast convergence rates for entropic transportation costs. *arXiv preprint arXiv:2204.09105*, 2022b.

Delalande, A. Nearly tight convergence bounds for semi-discrete entropic optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 1619–1642. PMLR, 2022.

Demetçi, P., Santorella, R., Sandstede, B., and Singh, R. Unsupervised integration of single-cell multi-omics datasets with disproportionate cell-type representation. In *International Conference on Research in Computational Molecular Biology*, pp. 3–19. Springer, 2022.

Divol, V., Niles-Weed, J., and Pooladian, A.-A. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 291–299. Springer, 2017.

Finlay, C., Gerolin, A., Oberman, A. M., and Pooladian, A.-A. Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv preprint arXiv:2006.06033*, 2020.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., and Weed, J. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2454–2465. PMLR, 2019.

Genevay, A. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Paris Sciences et Lettres (ComUE), 2019.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of Sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1574–1583. PMLR, 2019.

Ghosal, P. and Sen, B. Multivariate ranks and quantiles using optimal transport: consistency, rates and nonparametric testing. *Ann. Statist.*, 50(2):1012–1037, 2022. ISSN 0090-5364. doi: 10.1214/21-aos2136.

Ghosal, P., Nutz, M., and Bernton, E. Stability of entropic optimal transport and Schrödinger bridges. *Journal of Functional Analysis*, 283(9):109622, 2022.

Goldfeld, Z., Kato, K., Rioux, G., and Sadhu, R. Limit theorems for entropic optimal transport maps and the Sinkhorn divergence. *arXiv preprint arXiv:2207.08683*, 2022.

Gonzalez-Sanz, A., Loubes, J.-M., and Niles-Weed, J. Weak limits of entropy regularized optimal transport; potentials, plans and divergences. *arXiv preprint arXiv:2207.07427*, 2022.

Graf, S. and Luschgy, H. *Foundations of quantization for probability distributions*. Springer, 2007.

Gunsilius, F. and Xu, Y. Matching for causal effects via multimarginal optimal transport. *arXiv preprint arXiv:2112.04398*, 2021.

Hundrieser, S., Staudt, T., and Munk, A. Empirical optimal transport between different measures adapts to lower complexity. *arXiv preprint arXiv:2202.10434*, 2022.

Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2): 1166–1194, 2021.

Kantorovitch, L. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.

Léonard, C. From the Schrödinger problem to the Monge–Kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

Lübeck, F., Bunne, C., Gut, G., del Castillo, J. S., Pelkmans, L., and Alvarez-Melis, D. Neural unbalanced optimal transport via cycle-consistent semi-couplings. *arXiv preprint arXiv:2209.15621*, 2022.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

Marino, S. D. and Gerolin, A. An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):1–28, 2020.

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

Mérigot, Q., Santambrogio, F., and Sarrazin, C. Non-asymptotic convergence bounds for Wasserstein approximation using point clouds. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12810–12821. Curran Associates, Inc., 2021.

Moriel, N., Senel, E., Friedman, N., Rajewsky, N., Karaiskos, N., and Nitzan, M. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9):4177–4200, 2021.

Muzellec, B., Vacher, A., Bach, F., Vialard, F.-X., and Rudi, A. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. *arXiv preprint arXiv:2112.01907*, 2021.

Nutz, M. and Wiesel, J. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1-2):401–424, 2022.

Pal, S. On the difference between entropic cost and the optimal transport cost. *arXiv preprint arXiv:1905.12206*, 2019.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607, 2019.

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Pooladian, A.-A., Cuturi, M., and Niles-Weed, J. Debiaser beware: Pitfalls of centering regularized transport maps. *arXiv preprint arXiv:2202.08919*, 2022.

Rigollet, P. and Stromme, A. J. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.

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

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations*, 2018.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4):72:1–72:13, 2016.

Torous, W., Gunsilius, F., and Rigollet, P. An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*, 2021.

Vaart, A. W. and Wellner, J. A. Weak convergence and empirical processes with applications to statistics. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

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

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Yang, K. D., Damodaran, K., Venkatachalapathy, S., Soylemezoglu, A. C., Shivashankar, G., and Uhler, C. Predicting cell lineages using autoencoders and optimal transport. *PLoS computational biology*, 16(4):e1007828, 2020.

A. Reminders on semi-discrete entropic optimal transport

We recall in this section some known results on entropic optimal transport that will be needed later. Let $\mu, \nu \in \mathcal{P}(\Omega)$, where $\Omega \subset B(0; R)$ is a compact set.

Lemma A.1 (Genevay et al., 2019)**.** *The entropic potential* $(\varphi_\varepsilon^{\mu\to\nu}, \psi_\varepsilon^{\mu\to\nu})$ *have a bounded amplitude, in the sense that*

$$\max_{x\in\Omega} \varphi_\varepsilon^{\mu\to\nu} - \min_{x\in\Omega} \varphi_\varepsilon^{\mu\to\nu} \leq cR \tag{23}$$

for some absolute constant c, and similarly for $\psi_\varepsilon^{\mu\to\nu}$.

Assume now that $\nu = \sum_{j=1}^J \nu_j \delta_{y_j}$ is a discrete measure. In this situation, only the values of the dual potential $\psi_\varepsilon^{\mu\to\nu}$ on the points $y_1, \dots, y_J$ are relevant. We therefore consider $\psi_\varepsilon^{\mu\to\nu}$ as a vector in $\mathbb{R}^J$. The potentials $\varphi_\varepsilon^{\mu\to\nu}$ and $\psi_\varepsilon^{\mu\to\nu}$ are dual of one another, in the sense of the ε-Legendre transform. Given a finite measure ρ, the ε-Legendre transform of a function h with respect to ρ is given by

$$\Phi_\varepsilon^\rho(h)(x) = \varepsilon \log \int e^{(\langle x,y\rangle - h(x))/\varepsilon} \,\mathrm{d}\rho(x). \tag{24}$$

Relations (11) and (12) express that $\varphi_\varepsilon^{\mu\to\nu} = \Phi_\varepsilon^\nu(\psi_\varepsilon^{\mu\to\nu})$ and vice-versa. In the semi-discrete setting, it is also convenient to introduce the ε-Legendre transform with respect to the counting measure σ on $\{y_1, \dots, y_J\}$. For a vector $\psi \in \mathbb{R}^J$, we have

$$\Phi_\varepsilon(\psi)(x) \coloneqq \Phi_\varepsilon^\sigma(\psi)(x) = \varepsilon \log \sum e^{(\langle x,y_j\rangle - \psi(y_j))/\varepsilon}. \tag{25}$$

The Φ_ε transform and the Φ_ε^ν transform are linked through the relation

$$\Phi_\varepsilon^\nu(\psi) = \Phi_\varepsilon(\tilde\psi) \qquad \text{where} \qquad \tilde\psi(y_j) = \psi(y_j) - \varepsilon \log \nu_j\,, \tag{26}$$

where we call $\tilde\psi$ a *shifted* potential. With this notation, the optimality condition on the potentials can be rephrased. Let

$$F_\varepsilon^{\mu\to\nu} : \psi \in \mathbb{R}^J \to \int \Phi_\varepsilon(\psi)\,\mathrm{d}\mu + \int \psi\,\mathrm{d}\nu\,. \tag{27}$$

Then, the function $F_\varepsilon^{\mu\to\nu}$ is minimized at $\tilde\psi_\varepsilon^{\mu\to\nu}$. For $\psi \in \mathbb{R}^J$ and $x \in \mathbb{R}^d$, we introduce the probability measure supported on $\{y_1, \dots, y_J\}$ given by

$$\forall i \in [J], \quad \pi_\varepsilon^x[\psi](y_i) = \frac{e^{(\langle x,y_i\rangle - \psi(y_i))/\varepsilon}}{\sum_{j=1}^J e^{(\langle x,y_j\rangle - \psi(y_j))/\varepsilon}} = e^{(\langle x,y_i\rangle - \Phi_\varepsilon(\psi)(x) - \psi(y_i))/\varepsilon}. \tag{28}$$

A computation gives $\nabla F_\varepsilon^{\mu\to\nu}(\psi) = \int \pi_\varepsilon^x[\psi]\,\mathrm{d}\mu(x) - \nu$, so that at optimality, we have

$$\int \pi_\varepsilon^x[\tilde\psi_\varepsilon^{\mu\to\nu}]\,\mathrm{d}\mu(x) = \nu. \tag{29}$$

In this case, $\pi_\varepsilon^x = \pi_\varepsilon^x[\tilde\psi_\varepsilon^{\mu\to\nu}]$ is the conditional distribution of the second marginal of π_ε given that the first is equal to x, as in Section 2.2.1. More generally, for any potential ψ, the first order condition implies that ψ is equal to $\tilde\psi_\varepsilon^{\mu\to\nu_\psi}$, the optimal dual potential between μ an $\nu_\psi = \int \pi_\varepsilon^x[\psi]\,\mathrm{d}\mu(x)$.

B. Bound on the approximation error

Proof of Theorem 3.4. Let $i, j \in [J]$. We define the jth slack at $x \in L_i$ by

$$\frac{1}{2}\Delta_{ij}(x) = -\langle x, y_j\rangle + \varphi_0(x) + \psi_0(y_j). \tag{30}$$

As φ_0 is the Legendre transform of ψ_0, we have $\Delta_{ij}(x) \geq 0$. If the cells L_i and L_j have a nonempty intersection, the set $H_{ij}(t) = \{x \in L_i : \Delta_{ij}(x) = t\}$ represents the trace on L_i of the hyperplane spanned by the boundary between L_i and L_j, shifted by t. It is stated in (Altschuler et al., 2022) that for every nonnegative measurable function $f : \mathbb{R} \to \mathbb{R}_+$,

$$\int_{L_i} f(\Delta_{ij}(x))p(x)\,\mathrm{d}x = \frac{1}{2\|y_i - y_j\|}\int_0^\infty f(t)h_{ij}(t)\,\mathrm{d}t, \tag{31}$$

where $h_{ij}(t) = \int_{H_{ij}(t)} p(x)\,\mathrm{d}\mathcal{H}_{d-1}(x)$ and $\mathcal{H}_{d-1}$ is the $(d-1)$-dimensional Hausdorff measure. In particular, $w_{ij} = h_{ij}(0)$ is the (weighted) surface of the boundary between the ith and j^{th} Laguerre cells (should it exist). Given $x \in L_i$, let $s(x) = \min_{j\neq i} \frac{1}{2}\Delta_{ij}(x)$. When the point x is sufficiently inside its Laguerre cell, the conditional probability π_ε^x becomes extremely concentrated around the point y_i, as the next lemma shows. Note that $\pi_0^x = \delta_{y_i}$ when $x \in L_i$.

Lemma B.1. *Let $x \in L_i$. For ε small enough, it holds that for every $j \in [J]$, $|\pi_\varepsilon^x(y_j) - \pi_0^x(y_j)| \leq ce^{-s(x)/\varepsilon}$, where c depends on J, the distances $\|y_i - y_j\|$ and on the quantities w_{ij}.*

Such a result was already stated in (Delalande, 2022, Corollary 2.2), although while requiring that the source measure P has a Hölder continuous density. Only assumption **(A)** is needed here.

Proof. According to (Altschuler et al., 2022, Proposition 4.6), for ε small enough,

$$\varepsilon^{-1}\|\tilde{\psi}_\varepsilon - \psi_0\|_\infty \leq C, \tag{32}$$

where $\tilde{\psi}_\varepsilon$ is the shifted version of ψ_ε (see (25)) and C depends on the distances $\|y_i - y_j\|$ and on the w_{ij}s. Following (Delalande, 2022, Proof of Corollary 2.2) and (28), we have for $j \neq i$

$$|\pi_\varepsilon^x(y_j) - \pi_0^x(y_j)| = \pi_\varepsilon^x(y_j) = \frac{e^{(\langle x, y_j\rangle - \tilde{\psi}_\varepsilon(y_j))/\varepsilon}}{\sum_{j'=1}^J e^{(\langle x, y_{j'}\rangle - \tilde{\psi}_\varepsilon(y_{j'}))/\varepsilon}} \leq e^{2C}\frac{e^{(\langle x, y_j\rangle - \psi_0(y_j))/\varepsilon}}{\sum_{j'=1}^J e^{(\langle x, y_{j'}\rangle - \psi_0(y_{j'}))/\varepsilon}} \leq e^{2C}e^{-s(x)/\varepsilon}.$$

A similar computation yields that $|\pi_\varepsilon^x(y_i) - \pi_0^x(y_i)| = |\pi_\varepsilon^x(y_i) - 1| \leq Je^{2C}e^{-s(x)/\varepsilon}$. □

We can bound for any $x \in L_i$,

$$\|T_\varepsilon(x) - T_0(x)\| = \|\sum_{j=1}^J y_j(\pi_\varepsilon^x(y_j) - \pi_0^x(y_j))\| \leq c\sum_{j=1}^J \|y_j\|e^{-s(x)/\varepsilon}. \tag{33}$$

Therefore, letting C' denote a constant, which may depend on J, whose value may change from line to line, we obtain

$$\|T_\varepsilon - T_0\|_{L_2(P)}^2 = \sum_{i=1}^J \int_{L_i} \|T_\varepsilon(x) - T_0(x)\|^2\,\mathrm{d}P(x) \leq C'\sum_{i=1}^J \int_{L_i} \sum_{j=1}^J e^{-2s(x)/\varepsilon}\,\mathrm{d}P(x) \tag{34}$$

$$\leq C'\sum_{i\neq j}\int_{L_i} e^{-\Delta_{ij}(x)/\varepsilon}\,\mathrm{d}P(x) \leq C'\sum_{i\neq j}\frac{1}{2\|y_i - y_j\|}\int_0^\infty e^{-t/\varepsilon}h_{ij}(t)\,\mathrm{d}t\,, \tag{35}$$

where in the second equality, we used the definition of $s(x)$. Assumption **(A)** ensures that the functions h_{ij}s are bounded, which implies that the right-hand side in (35) is of order ε. □

C. Stability of entropic transport plans

Proof of Proposition 3.9. Note that we may assume without loss of generality that $\nu \ll \nu'$ and that $\mathrm{KL}(\nu\|\nu') < \infty$, for otherwise the bound is vacuous. For notational convenience, we omit the dependence on ε in the subscripts.

Write $\pi^{\mu,\nu} = \gamma^{\mu,\nu}(x,y)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$ for the entropic optimal plan between μ and ν, where $\gamma^{\mu,\nu} = \exp\left(\frac{1}{\varepsilon}(\langle x, y\rangle - \varphi^{\mu\to\nu}(x) - \psi^{\mu\to\nu}(y))\right)$, and analogously define $\gamma^{\mu',\nu'} = \exp\left(\frac{1}{\varepsilon}(\langle x, y\rangle - \varphi^{\mu'\to\nu'}(x) - \psi^{\mu'\to\nu'}(y))\right)$.

Consider the measure $\gamma^{\mu',\nu'}(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y)$. The first-order optimality condition for $(\varphi^{\mu'\to\nu'}, \psi^{\mu'\to\nu'})$ implies that

$$\int \gamma^{\mu',\nu'}(y)\,\mathrm{d}\nu'(y) = 1 \quad \forall x \in \Omega\,, \tag{36}$$

so that $\gamma^{\mu',\nu'}(x,y)\,\mathrm{d}\nu'(y)$ is a probability measure. Let us write $\mathrm{d}\pi^x(y) = \gamma^{\mu,\nu}(x,y)\,\mathrm{d}\nu(y)$ and $\mathrm{d}\rho^x(y) = \gamma^{\mu',\nu'}(x,y)\,\mathrm{d}\nu'(y)$.

We make the following observations: first, $T^{\mu\to\nu}(x) = \int y \,\mathrm{d}\pi^x(y)$ and $T^{\mu'\to\nu'}(x) = \int y \,\mathrm{d}\rho^x(y)$. Second, the support of ρ^x lies inside $B(0;R)$; since any Lipschitz function f on $B(0;R)$ satisfies $\sup_x f(x) - \inf_x f(x) \le 2R$, Hoeffding's lemma (see Boucheron et al., 2013, Lemma 2.2) implies that if f is Lipschitz and $\int f \,\mathrm{d}\rho^x = 0$, then

$$\int e^{tf} \,\mathrm{d}\rho^x \le e^{2R^2t^2} \quad \forall t \in \mathbb{R}\,.$$

This implies (Bobkov & Götze, 1999, Theorem 3.1) that

$$W_1(\pi^x, \rho^x)^2 \le 8R^2 \mathrm{KL}(\pi^x \| \rho^x)\,. \tag{37}$$

Third, Jensen's inequality implies that for any coupling γ between π^x and ρ^x,

$$\int \|y - y'\| \,\mathrm{d}\gamma(y,y') \ge \left\| \int (y - y') \,\mathrm{d}\gamma(y,y') \right\| = \|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\|\,, \tag{38}$$

so that in particular, $\|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\| \le W_1(\pi^x, \rho^x)$. Combining these facts, we obtain

$$\frac{1}{8R^2}\|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\|^2 \le \mathrm{KL}(\pi^x\|\rho^x) = \int \log\left(\frac{\gamma^{\mu,\nu}}{\gamma^{\mu',\nu'}}(x,y)\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\right)\gamma^{\mu,\nu}(x,y)\,\mathrm{d}\nu(y)\,. \tag{39}$$

Integrating both sides of this equation with respect to μ yields

$$\frac{1}{8R^2}\|T^{\mu\to\nu}(x) - T^{\mu'\to\nu'}(x)\|^2_{L^2(\mu)} \le \int \log\left(\frac{\gamma^{\mu,\nu}}{\gamma^{\mu',\nu'}}(x,y)\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\right)\,\mathrm{d}\pi^{\mu,\nu}(x,y)\,. \tag{40}$$

Expanding the definition of $\gamma^{\mu,\nu}$ and $\gamma^{\mu',\nu'}$ and using that $\int \log \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\,\mathrm{d}\pi^{\mu,\nu}(x,y) = \int \log \frac{\mathrm{d}\nu}{\mathrm{d}\nu'}(y)\,\mathrm{d}\nu(y) = \mathrm{KL}(\nu\|\nu')$ yields the claim. □

We now record two corollaries of this bound, which apply when either the source or the target measures of the entropic maps agree.

Corollary C.1. *For any μ, ν, ν' supported in $B(0;R)$,*

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu\to\nu'}\|^2_{L^2(\mu)} \le \varepsilon^{-1}\int(\psi_\varepsilon^{\mu\to\nu'} - \psi_\varepsilon^{\mu\to\nu})\,\mathrm{d}(\nu - \nu') + \mathit{KL}(\nu\|\nu') \tag{41}$$

Proof. We apply Proposition 3.9 with $\mu = \mu'$, which yields (once again omitting the dependency in ε)

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu\to\nu'}\|^2_{L^2(\mu)} \le \varepsilon^{-1}\left(\int(\varphi^{\mu\to\nu'} - \varphi^{\mu\to\nu})\,\mathrm{d}\mu + \int(\psi^{\mu\to\nu'} - \psi^{\mu\to\nu})\,\mathrm{d}\nu\right) + \mathrm{KL}(\nu\|\nu')\,. \tag{42}$$

By definition, $(\varphi^{\mu\to\nu'}, \psi^{\mu\to\nu'})$ minimizes the expression $\int\varphi\,\mathrm{d}\mu + \int\psi\,\mathrm{d}\nu' + \varepsilon\iint e^{(\langle x,y\rangle - \varphi(x) - \psi(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y) - \varepsilon$, so, recalling that $\iint e^{(\langle x,y\rangle - \varphi^{\mu\to\nu'}(x) - \psi^{\mu\to\nu'}(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y) = 1$, we have in particular

$$\begin{aligned}\int\varphi^{\mu\to\nu'}\,\mathrm{d}\mu + \int\psi^{\mu\to\nu'}\,\mathrm{d}\nu' &\le \int\varphi^{\mu\to\nu}\,\mathrm{d}\mu + \int\psi^{\mu\to\nu}\,\mathrm{d}\nu' + \varepsilon\iint e^{(\langle x,y\rangle - \varphi^{\mu\to\nu}(x) - \psi^{\mu\to\nu}(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y) - \varepsilon \\ &= \int\varphi^{\mu\to\nu}\,\mathrm{d}\mu + \int\psi^{\mu\to\nu}\,\mathrm{d}\nu'\,,\end{aligned}$$

where we have used that the first-order optimality condition for $(\varphi^{\mu\to\nu}, \psi^{\mu\to\nu})$ implies that $\iint e^{(\langle x,y\rangle - \varphi^{\mu\to\nu}(x) - \psi^{\mu\to\nu}(y))/\varepsilon}\,\mathrm{d}\mu(x)\,\mathrm{d}\nu'(y) = 1$ as well (see (11)). This implies

$$\int(\varphi^{\mu\to\nu'} - \varphi^{\mu\to\nu})\,\mathrm{d}\mu \le -\int(\psi^{\mu\to\nu'} - \psi^{\mu\to\nu})\,\mathrm{d}\nu'\,. \tag{43}$$

Applying this inequality to (42) yields

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu\to\nu'}\|^2_{L^2(\mu)} \le \varepsilon^{-1}\int(\psi^{\mu\to\nu'} - \psi^{\mu\to\nu})\,\mathrm{d}(\nu - \nu') + \mathrm{KL}(\nu\|\nu').$$ □

Corollary C.2. *For any μ, μ', ν supported in $B(0;R)$,*

$$\frac{1}{8R^2}\|T_\varepsilon^{\mu\to\nu} - T_\varepsilon^{\mu'\to\nu}\|_{L^2(\mu)}^2 \le \varepsilon^{-1}\int(\varphi_\varepsilon^{\mu'\to\nu} - \varphi_\varepsilon^{\mu\to\nu})\,\mathrm{d}(\mu-\mu')\,. \tag{44}$$

Proof. We apply Proposition 3.9 with $\nu = \nu'$, yielding (dropping the dependency on ε)

$$\frac{1}{8R^2}\|T^{\mu\to\nu} - T^{\mu'\to\nu}\|_{L^2(\mu)}^2 \le \varepsilon^{-1}\left(\int(\varphi^{\mu'\to\nu} - \varphi^{\mu\to\nu})\,\mathrm{d}\mu + \int(\psi^{\mu'\to\nu} - \psi^{\mu\to\nu})\,\mathrm{d}\nu\right)\,. \tag{45}$$

An argument analogous to the one used in the proof of Corollary C.1 gives the inequality

$$\int\varphi^{\mu'\to\nu}\,\mathrm{d}\mu' + \int\psi^{\mu'\to\nu}\,\mathrm{d}\nu \le \int\varphi^{\mu\to\nu}\,\mathrm{d}\mu' + \int\psi^{\mu\to\nu}\,\mathrm{d}\nu\,, \tag{46}$$

or, equivalently,

$$\int(\psi^{\mu'\to\nu} - \psi^{\mu\to\nu})\,\mathrm{d}\nu \le -\int(\varphi^{\mu'\to\nu} - \varphi^{\mu\to\nu})\,\mathrm{d}\mu'\,, \tag{47}$$

and combining this inequality with (45) proves the claim. □

D. Strong convexity of the entropic semi-dual problem

Proposition D.1 (Strong convexity of $F_\varepsilon^{\mu\to\nu}$)**.** *Let $\nu = \sum_{j=1}^J \nu_j\delta_{y_j}$ be a measure supported on $\{y_1,\dots,y_J\} \subseteq B(0;R)$ and let μ supported on a compact convex set $\Omega \subseteq B(0;R)$ with a density p satisfying $p_{\min} \le p \le p_{\max}$ for some $p_{\max} \ge p_{\min} > 0$. For $\psi \in \mathbb{R}^J$, define $\nu_\psi = \int \pi_\varepsilon^x[\psi]\,\mathrm{d}\mu(x)$ and assume that $\nu_\psi \ge \lambda\nu$ for some $0 < \lambda \le 1$. Then, we have for $\varepsilon \in (0,1)$*

$$F_\varepsilon^{\mu\to\nu}(\psi) - \min_\psi F_\varepsilon^{\mu\to\nu} \ge C\lambda\cdot\mathrm{Var}_\nu(\psi - \psi_\varepsilon^{\mu\to\nu}), \tag{48}$$

where $C = \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)^{-1}\frac{p_{\min}}{p_{\max}}$.

Proof. As μ and ε are fixed, we will simply write ψ_ν instead of $\psi_\varepsilon^{\mu\to\nu}$, and write similarly $F_\nu = F_\varepsilon^{\mu\to\nu}$. Recall the definition (25) of the shifted potential $\tilde\psi_\nu(y_j) = \psi_\nu(y_j) - \varepsilon\log\nu_j$. According to (Delalande, 2022, Theorem 3.2), the functional F_ν is minimized at the vector $\tilde\psi_\nu$, with

$$\forall v \in \mathbb{R}^J, \quad \mathrm{Var}_\nu(v) \le \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)v^\top\nabla^2 F_\nu(\tilde\psi_\nu)v. \tag{49}$$

For $t \in [0,1]$, let $\psi_t = \tilde\psi_\nu + t(\psi - \tilde\psi_\nu)$ and let $\nu_t = \int\pi_\varepsilon^x[\psi_t]\,\mathrm{d}\mu(x)$. The potential ψ_t is the (shifted) entropic Brenier potential between μ and ν_t, so that it minimizes the functional F_{ν_t} (see Appendix A). Also, note that $\nabla^2 F_\nu$ does not depend on ν, so that

$$v^\top\nabla^2 F_\nu(\psi_t)v = v^\top\nabla^2 F_{\nu_t}(\psi_t)v \ge \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)^{-1}\mathrm{Var}_{\nu_t}(v). \tag{50}$$

Let $v = \psi - \psi_\varepsilon^{\mu\to\nu}$. A Taylor expansion of F_ν gives

$$F_\nu(\psi) - F_\nu(\tilde\psi_\nu) = \int_0^1 v^\top\nabla^2 F_\nu(\psi_t)v\,\mathrm{d}t \ge \left(e^{2R^2}\frac{p_{\max}}{p_{\min}} + \varepsilon\right)^{-1}\int_0^1\mathrm{Var}_{\nu_t}(v)\,\mathrm{d}t. \tag{51}$$

Lemma D.2. *Write $\nu_t = \sum_{j=1}^J \nu_{t,j}\delta_{y_j}$. Then, for all $t \in [0,1]$ and $j \in [J]$, we have $\nu_{t,j} \ge \frac{p_{\min}}{p_{\max}}\nu_{0,j}^{1-t}\nu_{1,j}^t$.*

This lemma is enough to conclude the proof. Indeed, $\nu_1 = \nu_\psi \ge \lambda\nu$, so that it implies that $\mathrm{Var}_{\nu_t}(v) \ge \frac{p_{\min}}{p_{\max}}\lambda\mathrm{Var}_\nu(v)$. □

Proof of Lemma D.2. According to (Delalande, 2022, Proof of Proposition 4.1),

$$\Phi_\varepsilon(\psi_t)(tx + (1-t)y) \le t\Phi_\varepsilon(\tilde\psi_\varepsilon^{\mu\to\nu})(x) + (1-t)\Phi_\varepsilon(\psi)(y). \tag{52}$$

Therefore, if we let $h_t(x) = e^{(\langle x, y_j\rangle - \psi_t(y_j) - \Phi_\varepsilon(\psi_t)(x))/\varepsilon}$, then we have $h_t(tx + (1-t)y) \geq h_0(x)^t h_1(y)^{1-t}$. By the Prékopa-Leindler inequality,

$$\nu_{t,j} = \int h_t(x)\,\mathrm{d}\mu(x) \geq p_{\min} \int_{\mathcal{X}} h_t(x)\,\mathrm{d}x \geq p_{\min} \left(\int_{\mathcal{X}} h_0(x)\,\mathrm{d}x\right)^t \left(\int_{\mathcal{X}} h_1(x)\,\mathrm{d}x\right)^{1-t} \geq \frac{p_{\min}}{p_{\max}} \nu_{0,j}^t \nu_{1,j}^{1-t}.$$

□

Proof of Proposition 3.11. As in the previous proof, we drop the ε and μ dependency in our notation. Write $\nu_k = \sum_{j=1}^J \nu_{k,j}\delta_{y_j}$ for $k = 0, 1$, and define as before the shifted potentials $\tilde{\psi}_{\nu_k}(y_j) = \psi_{\nu_1}(y_j) - \varepsilon \log \nu_{k,j}$. Let $\theta > 0$ be a parameter to fix. According to Proposition D.1, Lemma H.1, and using the inequality $F_{\nu_1}(\tilde{\psi}_{\nu_1}) \leq F_{\nu_1}(\tilde{\psi}_{\nu_0})$, we have

$$\begin{aligned} C\lambda \mathrm{Var}_{\nu_0}(\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0}) \leq F_{\nu_0}(\tilde{\psi}_{\nu_1}) - F_{\nu_0}(\tilde{\psi}_{\nu_0}) &\leq F_{\nu_0}(\tilde{\psi}_{\nu_1}) - F_{\nu_1}(\tilde{\psi}_{\nu_1}) + F_{\nu_1}(\tilde{\psi}_{\nu_0}) - F_{\nu_0}(\tilde{\psi}_{\nu_0}) \\ &= \int (\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0})(\,\mathrm{d}\nu_0 - \,\mathrm{d}\nu_1) \\ &\leq \frac{\theta}{2}\mathrm{Var}_{\nu_0}(\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0}) + \frac{1}{2\theta}\chi^2(\nu_1\|\nu_0). \end{aligned}$$

We pick $\theta = C\lambda$ to conclude that

$$\mathrm{Var}_{\nu_0}(\tilde{\psi}_{\nu_1} - \tilde{\psi}_{\nu_0}) \leq \frac{1}{(C\lambda)^2}\chi^2(\nu_1\|\nu_0). \tag{53}$$

Therefore, using the inequality $|\log(a/b)| \leq |a-b|/\min\{a,b\}$ for $a, b > 0$,

$$\begin{aligned} \mathrm{Var}_{\nu_0}(\psi_1 - \psi_0) &\leq 2\mathrm{Var}_{\nu_0}(\tilde{\psi}_1 - \tilde{\psi}_0) + 2\sum_{j=1}^J \nu_{0,j}\left(\log\left(\frac{\nu_{1,j}}{\nu_{0,j}}\right)\right)^2 \\ &\leq \frac{2}{(C\lambda)^2}\chi^2(\nu_1\|\nu_0) + 2\sum_{j=1}^J \nu_{0,j}\left(\frac{\nu_{1,j} - \nu_{0,j}}{\min\{\nu_{0,j}, \nu_{1,j}\}}\right)^2 \\ &\leq \frac{2}{(C\lambda)^2}\chi^2(\nu_1\|\nu_0) + \frac{2}{\lambda^2}\sum_{j=1}^J \frac{1}{\nu_{0,j}}(\nu_{1,j} - \nu_{0,j})^2 \leq \left(\frac{2}{(C\lambda)^2} + \frac{2}{\lambda^2}\right)\chi^2(\nu_1\|\nu_0). \end{aligned}$$

□

E. Control of the fluctuations in the one-sample case

Lemma E.1 (Sample complexity in the one-sample case)**.** *Assume that P satisfy **(A)** and that Q satisfy **(B)**. Then, it holds that $\mathbb{E}\|T_\varepsilon^{P\to Q_n} - T_\varepsilon\|_{L^2(P)}^2 \lesssim \varepsilon^{-1}n^{-1}$.*

Proof. To ease notation, we write $T_{\varepsilon,n} = T_\varepsilon^{P\to Q_n}$ and $\psi_{\varepsilon,n} = \psi_\varepsilon^{P\to Q_n}$. As explained in Section 3, the stability result Proposition 3.9 implies that

$$\mathbb{E}\|T_{\varepsilon,n} - T_\varepsilon\|_{L^2(P)}^2 \leq \frac{8R^2}{\varepsilon}\left(\frac{\mathbb{E}[\mathrm{Var}_Q(\psi_{\varepsilon,n} - \psi_\varepsilon)]}{2} + \frac{\mathbb{E}[\chi^2(Q_n\|Q)]}{2}\right) + 8R^2\mathbb{E}[\chi^2(Q_n\|Q)]\,. \tag{54}$$

Write $Q = \sum_{j=1}^J q_j\delta_{y_j}$ and $Q_n = \sum_{j=1}^J \hat{q}_j\delta_{y_j}$, and introduce the event $E = \{\forall j \in [J],\ \hat{q}_j \geq q_j/2\}$. If E is satisfied, we have $Q_n \geq Q/2$, so that Proposition 3.11 yields

$$\mathrm{Var}_Q(\psi_{\varepsilon,n} - \psi_\varepsilon) \leq C\chi^2(Q_n\|Q). \tag{55}$$

If E is not satisfied, we use the fact that the entropic potentials have a bounded amplitude (see Lemma A.1), to obtain that

$$\mathrm{Var}_Q(\psi_{\varepsilon,n} - \psi_\varepsilon) \leq C'. \tag{56}$$

Lemma E.2. *Let E be the event that $Q_n \geq Q/2$. Then $\mathbb{P}(E^c) \leq Je^{-cq_{\min}n}$ for some $c > 0$.*

Proof. By (Vershynin, 2018, Exercise 2.3.2), we have $\mathbb{P}(E^c) \leq \sum_{j=1}^J \mathbb{P}(\hat{q}_j < q_j/2) \leq Je^{-cq_{\min}n}$ for some $c > 0$. □

We obtain

$$\mathbb{E}\|\hat{T}_{\varepsilon,n} - T_\varepsilon\|^2_{L^2(P)} \lesssim \frac{R^2}{\varepsilon}\mathbb{E}[\chi^2(Q_n\|Q)] + \frac{R^2}{\varepsilon}Je^{-cq_{\min}n} \lesssim \varepsilon^{-1}n^{-1} \tag{57}$$

by Lemma H.2. □

F. Control of the fluctuations in the two-sample case

The goal of this section is to prove Theorem 3.7. We will actually prove a more general result, and show that *for any discrete measure* $\nu = \sum_{j=1}^J \nu_j\delta_{y_j}$ supported on $\{y_1,\dots,y_J\}$ with $\nu_j \geq \nu_{\min} > 0$ for all $j \in [J]$, we have for $\log(1/\varepsilon) \lesssim n/\log(n)$,

$$\mathbb{E}\|T_\varepsilon^{P_n\to\nu} - T_\varepsilon^{P\to\nu}\|^2_{L_2(P)} \lesssim \varepsilon^{-1}n^{-1}. \tag{58}$$

Theorem 3.7 follows from (58) by conditioning on Q_n. Let E be the event that $Q_n \geq Q/2$. Then, by Lemma E.2,

$$\begin{aligned}\mathbb{E}\|\hat{T}_\varepsilon - T_\varepsilon^{P\to Q_n}\|^2_{L_2(P)} &\leq \mathbb{E}\left[\mathbb{E}[\|\hat{T}_\varepsilon - T_\varepsilon^{P\to Q_n}\|^2_{L_2(P)}|Q_n]\mathbb{1}\{E\}\right] + R^2\mathbb{P}(E^c)\\ &\leq C\varepsilon^{-1}n^{-1} + R^2Je^{-cq_{\min}n} \lesssim \varepsilon^{-1}n^{-1}.\end{aligned}$$

We obtain Theorem 3.7 by combining this bound with Lemma E.1.

To prove (58), we first use Corollary C.2 which yields

$$\begin{aligned}\mathbb{E}\|T_\varepsilon^{P_n\to\nu} - T_\varepsilon^{P\to\nu}\|^2_{L_2(P)} &\leq 8R^2\varepsilon^{-1}\mathbb{E}\int(\varphi_\varepsilon^{P_n\to\nu} - \varphi_\varepsilon^{P\to\nu})\,\mathrm{d}(P_n - P)\\ &= 8R^2\varepsilon^{-1}\mathbb{E}\int(\Phi_\varepsilon(\tilde{\psi}_\varepsilon^{P_n\to\nu}) - \Phi_\varepsilon(\tilde{\psi}_\varepsilon^{P\to\nu}))\,\mathrm{d}(P_n - P),\end{aligned} \tag{59}$$

where we recall that for a potential ψ, the shifted potential $\tilde{\psi}$ is given by $\tilde{\psi}_j = \psi_j - \varepsilon\log\nu_j$. The remainder of the proof consists in bounding this integral by using localization arguments and standard bounds on suprema of empirical processes. Our first goal is to show that the potential $\psi_\varepsilon^{P_n\to\nu}$ is close to to the potential $\psi_\varepsilon^{P\to\nu}$ for the ∞-norm. It will be convenient to work with the "L_∞-variance"

$$\mathrm{Var}_\infty(\psi) = \inf_{c\in\mathbb{R}}\max_{j\in[J]}|\psi(y_j) - c|^2 = \left(\frac{\max\psi - \min\psi}{2}\right)^2. \tag{60}$$

As the measure ν is lower bounded, it holds that

$$\mathrm{Var}_\nu(\psi) \geq \nu_{\min}\mathrm{Var}_\infty(\psi). \tag{61}$$

Lemma F.1 (Supremum of ε-Legendre transforms)**.** *Let ψ_0 be a fixed potential and let $\tau > 0$. Then, for all $j \in [J]$,*

$$\mathbb{E}\left[\sup_{\mathrm{Var}_\infty(\psi-\psi_0)\leq\tau^2}\left|\int(\pi_\varepsilon^x(\psi)_j - \pi_\varepsilon^x(\psi_0)_j)\,\mathrm{d}(P - P_n)(x)\right|\right] \leq C\sqrt{\frac{J\max\{\log(\tau/\varepsilon),1\}}{n}} \tag{62}$$

$$\mathbb{E}\left[\sup_{\mathrm{Var}_\infty(\psi-\psi_0)\leq\tau^2}\left|\int(\Phi_\varepsilon(\psi)(x) - \Phi_\varepsilon(\psi_0))(x)\,\mathrm{d}(P - P_n)(x)\right|\right] \leq C\tau\sqrt{\frac{J}{n}} \tag{63}$$

for some absolute constant C.

Proof. For a metric space (A,d) and $u > 0$, we let $N(u,A,d)$ be the covering number of A at scale u, that is the smallest number of balls of radius u needed to cover A. Let B be the L_∞-ball of radius τ in $\mathbb{R}^J$, centered at ψ_0, and let $\|\cdot\|_\infty$ denote the ∞-norm. For $0 < u \leq \tau$, we have $\log N(u,B,\|\cdot\|_\infty) \leq J\log(\tau/u)$.

We start with the second inequality. Note that $\psi \mapsto \Phi_\varepsilon(\psi)$ is 1-Lipschitz continuous, and that the functional Φ_ε satisfies $\Phi_\varepsilon(\psi + c) = \Phi_\varepsilon(\psi) + c$ for all $c \in \mathbb{R}$. Then the set $\{\psi :\ \mathrm{Var}_\infty(\psi - \psi_0) \leq \tau^2\}$ is equal to the set $\{\psi + c :\ \psi \in B,\ c \in \mathbb{R}\}$.

As $\int c\,\mathrm{d}(P-P_n)=0$, we can therefore restrict the supremum to vectors $\psi\in B$. Furthermore, an envelope function of the class $\{\Phi_\varepsilon(\psi)-\Phi_\varepsilon(\psi_0):\ \psi\in B\}$ is the constant function equal to τ. Therefore, by Lemma H.3, we obtain

$$\mathbb{E}\left[\sup_{\|\psi-\psi_0\|_\infty\le\tau}\left|\int(\Phi_\varepsilon(\psi)-\Phi_\varepsilon(\psi_0))(\,\mathrm{d}P-\,\mathrm{d}P_n)\right|\right]\le\frac{c_0}{\sqrt{n}}\int_0^{c_1\tau}\sqrt{J\log 2N(u,\{\Phi_\varepsilon(\psi):\ \psi\in B\},\|\cdot\|_\infty)}\,\mathrm{d}u$$
$$\le\sqrt{\frac{c_3J\tau}{n}}.\qquad\square$$

We repeat the same argument for the first inequality. The functional π_ε^x is invariant by translation: $\pi_\varepsilon^x(\psi+c)=\pi_\varepsilon^x(\psi)$ for all $c\in\mathbb{R}$. This implies that

$$\sup_{\mathrm{Var}_\infty(\psi-\psi_0)\le\tau^2}\left|\int(\Phi_\varepsilon(\psi)(x)-\Phi_\varepsilon(\psi_0))(x)\,\mathrm{d}(P-P_n)(x)\right|=\sup_{\|\psi-\psi_0\|_\infty\le\tau}\left|\int(\Phi_\varepsilon(\psi)(x)-\Phi_\varepsilon(\psi_0))(x)\,\mathrm{d}(P-P_n)(x)\right|.$$

As the function $\psi\mapsto\pi_\varepsilon^x(\psi)_j$ is ε^{-1}-Lipschitz continuous for every $x\in\mathbb{R}^d$, we have for $0<u\le\tau/\varepsilon$,

$$\log N(u,\{x\mapsto\pi_\varepsilon^x(\psi)_j:\ \psi\in B\},\|\cdot\|_\infty)\le J\log(\tau/(u\varepsilon))\,.$$

Remarking furthermore that $0\le\pi_\varepsilon^x(\psi)_j\le 1$ (so that the class of functions $\{x\mapsto\pi_\varepsilon^x(\psi)_j:\ \psi\in B\}$ admits the constant function 1 as an envelope function), we obtain the following control using Lemma H.3:

$$\mathbb{E}\left[\sup_{\|\psi-\psi_0\|_\infty\le\tau}\left|\int(\pi_\varepsilon^x(\psi)_j-\pi_\varepsilon^x(\psi_0)_j)(\,\mathrm{d}P-\,\mathrm{d}P_n)(x)\right|\right]\le\frac{c_0}{\sqrt{n}}\int_0^{c_1}\sqrt{J\log 2N(u,\{x\mapsto\pi_\varepsilon^x(\psi)_j:\ \psi\in B\},\|\cdot\|_\infty)}\,\mathrm{d}u$$
$$\le\sqrt{\frac{c_2J\max\{\log(\tau/\varepsilon),1\}}{n}},$$

where c_0, c_1 and c_2 are absolute constants, and the last line follows from arguing whether $c_1<\tau/\varepsilon$ or not.

Proposition F.2. *Assume that P satisfies **(A)** and let $\nu=\sum_{j=1}^J\nu_j\delta_{y_j}$ be a measure supported on $\{y_1,\dots,y_J\}\subset B(0;R)$, with $\nu_j\ge q_{\min}$ for all $j\in[J]$. Then, for all $0<\varepsilon\le 1$ with $\log(1/\varepsilon)\lesssim n/\log(n)$, it holds that*

$$\mathbb{E}\mathrm{Var}_\infty(\tilde\psi_\varepsilon^{P_n\to\nu}-\tilde\psi_\varepsilon^{P\to\nu})\lesssim n^{-1}.\tag{64}$$

Proof. To alleviate notation, we will write $\psi_n=\psi_\varepsilon^{P_n\to\nu}$ and $\psi_0=\psi_\varepsilon^{P\to\nu}$. Similarly, we write $F_n=F_\varepsilon^{P_n\to\nu}$ and $F_0=F_\varepsilon^{P\to\nu}$. Let $\nu_n=\int\pi_\varepsilon^x(\psi_\varepsilon^{P_n\to\nu})\,\mathrm{d}P(x)$. Under the event $E=\{\nu_n\ge\nu/2\}$, we have according to Proposition D.1 and the fact that $\tilde\psi_n$ minimizes F_n,

$$\begin{aligned}C\nu_{\min}\mathrm{Var}_\infty(\tilde\psi_n-\tilde\psi_0)\le C\mathrm{Var}_\nu(\tilde\psi_n-\tilde\psi_0)\le F_0(\tilde\psi_n)-F_0(\tilde\psi_0)&\le F_0(\tilde\psi_n)-F_n(\tilde\psi_n)+F_n(\tilde\psi_0)-F_0(\tilde\psi_0)\\&=\int(\Phi_\varepsilon(\tilde\psi_n)-\Phi_\varepsilon(\tilde\psi_0))\,\mathrm{d}(P-P_n)\end{aligned}\tag{65}$$

Let us bound $\mathbb{P}(E^c)$. As $\tilde\psi_n$ is the minimum of F_n, we have $\nu=\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}P_n(x)$ (see Appendix A). Therefore, we may write $\nu_{n,j}=\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}P_n(x)+\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}(P-P_n)(x)=\nu_j+Z_j$, where

$$Z_j=\int\pi_\varepsilon^x(\tilde\psi_n)_j\,\mathrm{d}(P-P_n)(x)=\int(\pi_\varepsilon^x(\tilde\psi_n)_j-\pi_\varepsilon^x(\tilde\psi_0)_j)\,\mathrm{d}(P-P_n)(x).$$

Note that $\mathrm{Var}_\infty(\tilde\psi_n-\tilde\psi_0)\lesssim R^2$ (see Lemma A.1), so that by Lemma F.1 and Lemma H.3,

$$\mathbb{P}(E^c)\le\sum_{j=1}^J\mathbb{P}(|Z_j|>\nu_j/2)\le J\exp\left(-c\frac{\sqrt{n}q_{\min}}{(\sqrt{J\log(1/\varepsilon)}+\log n}\right)\lesssim n^{-1},\tag{66}$$

under the condition $\log(1/\varepsilon)\lesssim n/\log(n)$.

For $k \geq 0$, let $a_k = 2^k/\sqrt{n}$ and fix some $p > 2$. Let $B_a = \sup_{\mathrm{Var}_\infty(\psi - \tilde{\psi}_0) \leq a^2} \left| \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\tilde{\psi}_0)) \, \mathrm{d}(P - P_n) \right|$. Assume that E is satisfied and that $\mathrm{Var}_\infty(\tilde{\psi}_0 - \tilde{\psi}_n) \in [a^2, b^2]$. Then, according to (65), it holds that $B_b \geq ca^2$. Using Markov's inequality, Lemma F.1 and Lemma H.3, we bound

$$\begin{aligned}
\mathbb{E}\mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0) &\leq a_0^2 + \sum_{k \geq 0} \mathbb{P}(\mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0) \in [a_k^2, a_{k+1}^2] \text{ and } E) a_{k+1}^2 + C\mathbb{P}(E^c) \\
&\lesssim n^{-1} + \sum_{k \geq 0} \mathbb{P}\left(B_{a_{k+1}} \geq c a_k^2\right) a_{k+1}^2 \lesssim n^{-1} + \sum_{k \geq 0} \frac{\mathbb{E}[B_{a_{k+1}}^p]}{a_k^{2p}} a_{k+1}^2 \\
&\lesssim n^{-1} + \sum_{k \geq 0} \frac{(2^k/n)^p}{(4^k/n)^p} \frac{4^{k+1}}{n} \lesssim n^{-1} + \sum_{k \geq 0} \frac{2^{2k-pk}}{n} \lesssim n^{-1}.
\end{aligned}$$

□

Proposition F.3. *Under the same assumptions than Proposition F.2, it holds that*

$$\mathbb{E}\|T_\varepsilon^{P_n \to \nu} - T_\varepsilon^{P \to \nu}\|_\infty^2 \lesssim \varepsilon^{-1} n^{-1}. \tag{67}$$

Proof. Let $Z = \mathrm{Var}_\infty(\tilde{\psi}_n - \tilde{\psi}_0)$. Let once again $a_k = 2^k/\sqrt{n}$ for $k \geq 1$, with $a_0 = 0$. Fix some $p > 2$, with $q = \frac{p}{p-1}$. For $a > 0$, let $D_a = \sup_{\mathrm{Var}_\infty(\psi - \tilde{\psi}_0) \leq a^2} \left| \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\tilde{\psi}_0)) \, \mathrm{d}(P - P_n) \right|$. By Hölder inequality and Markov inequality, we obtain,

$$\begin{aligned}
\mathbb{E} \int (\Phi_\varepsilon(\tilde{\psi}_n) - \Phi_\varepsilon(\tilde{\psi}_0)) \, \mathrm{d}(P - P_n) &\leq \sum_{k \geq 0} \mathbb{E}\left[\mathbb{1}\{Z \in [a_k^2, a_{k+1}^2]\} \sup_{\mathrm{Var}_\infty(\psi - \tilde{\psi}_0) \leq a_{k+1}^2} \int (\Phi_\varepsilon(\psi) - \Phi_\varepsilon(\tilde{\psi}_0)) \, \mathrm{d}(P - P_n) \right] \\
&\leq \mathbb{E}[D_{a_1}] + \sum_{k \geq 1} \left(\mathbb{P}(Z \geq a_k^2)\right)^{1/q} \mathbb{E}\left[D_{a_{k+1}}^p\right]^{1/p} \\
&\lesssim n^{-1} + \sum_{k \geq 0} \left(\frac{\mathbb{E}[Z]}{a_k^2}\right)^{1/q} \frac{2^k}{n} \lesssim \sum_{k \geq 0} \frac{2^{k(1-2/q)}}{n} \lesssim n^{-1},
\end{aligned}$$

where we use Proposition F.2, Lemma F.1 and Lemma H.3 at the last line. Equation (59) then gives the conclusion. □

G. A lower bound for the performance of the 1NN estimator

In this section, we prove Proposition 4.2. We let P be the Lebesgue measure on $\Omega = [0,1]^d$, and let $y_0 = (0, 1/2, \ldots, 1/2)$ and $y_1 = (1, 1/2, \ldots, 1/2)$. We denote by P_n an empirical measure consisting of i.i.d. samples from P. As in Appendix F, we work in a general setting of a generic discrete target measure ν, which may either be fixed or may be a random measure independent of P_n. We let $\nu = \sum_{j=0,1} \nu_j \delta_{y_j}$ for $\nu_0, \nu_1 \geq \frac{1}{4}$; this latter condition will hold with overwhelming probability if ν is an empirical measure Q_n corresponding to n i.i.d. samples from $Q = \frac{1}{2}\delta_{y_0} + \frac{1}{2}\delta_{y_1}$. Following Manole et al. (2021), we define the one-nearest neighbor estimator $\hat{T}_{\mathrm{1NN}}$ in this general context by

$$\hat{T}_{\mathrm{1NN}}(x) = \sum_{i=1}^n \sum_{j=0,1} \mathbf{1}_{V_i}(x)(n\hat{\pi}(X_i, y_j)),$$

where $\hat{\pi}$ is the empirical optimal coupling between P_n and ν.

We first examine the structure of the Brenier map $T_0 = \nabla\varphi_0$. The considerations in Section 2.1.1 imply that

$$T_0(x) = \begin{cases} y_0 & \langle e_1, x \rangle \leq \nu_0 \\ y_1 & \langle e_1, x \rangle > \nu_0, \end{cases}$$

where e_1 is the first elementary basis vector. The potential φ_0 is not differentiable on the separating hyperplane $\langle e_1, x \rangle = \nu_0$, which has measure 0 under P, but we may arbitrarily assign points on this hyperplane to y_0.

Similar arguments imply that the empirical transport plan $\hat{\pi}$ between P_n and ν has the following property: there exists a (random) threshold $\tau \in (0,1)$ such that

$$\hat{\pi}(x, y_0) = \begin{cases} 1 & \langle e_1, x\rangle < \tau \\ 0 & \langle e_1, x\rangle > \tau\,. \end{cases}$$

The set $\langle e_1, x\rangle = \tau$ may not have measure 0 under P_n, and $\hat{\pi}(x, y_0)$ may take values strictly between 0 and 1 on this set.

The following lemma shows that τ is close to ν_0 with high probability.

Lemma G.1. *For any $t \geq 0$,*

$$\mathbb{P}\left\{\tau \geq \nu_0 + t\right\} \leq e^{-2nt^2}\,.$$

Proof. If $\tau \geq \nu_0 + t$, this implies that $P_n(\{x : \langle e_1, x\rangle < \nu_0 + t\}) \leq \nu_0$. On the other hand, $nP_n(\{x : \langle e_1, x\rangle < \nu_0 + t\}$ is a $\mathrm{Bin}(n, \nu_0 + t)$ random variable. The result then follows from Hoeffding's inequality (Boucheron et al., 2013, Theorem 2.8). □

Let us write H for the halfspace $\{x : \langle e_1, x\rangle \leq \nu_0\}$, and $\hat{H}$ for the halfspace $\{x : \langle e_1, x\rangle \leq \tau\}$. Let x be any point in Ω such that $x \in H$. We are interested in the event that there exists an element $X_i \in \{X_1, \ldots, X_n\}$ such that a) $x \in V_i$ and b) $X_i \in \hat{H}^c$. Call this event $\mathcal{E}(x)$. On this event, $\hat{T}_{\mathrm{1NN}}(x) = y_1$ and $T_0(x) = y_0$, so $\|\hat{T}_{\mathrm{1NN}}(x) - T_0(x)\|^2 = 1$.

We therefore obtain

$$\begin{aligned}
\mathbb{E}\|\hat{T}_{\mathrm{1NN}} - T_0\|^2_{L^2(P)} &= \mathbb{E}\int \|\hat{T}_{\mathrm{1NN}}(x) - T_0(x)\|^2 \,\mathrm{d}P(x) \\
&\geq \mathbb{E}\int_H \|\hat{T}_{\mathrm{1NN}}(x) - T_0(x)\|^2 \mathbb{1}\{\mathcal{E}(x)\}\,\mathrm{d}P(x) \\
&\gtrsim \mathbb{E}\int_H \mathbb{1}\{\mathcal{E}(x)\}\,\mathrm{d}P(x) \\
&= \int_H \mathbb{P}\left\{\mathcal{E}(x)\right\}\,\mathrm{d}P(x)\,,
\end{aligned}$$

where the final equality follows from the Fubini–Tonelli theorem.

We now lower bound the probability of $\mathcal{E}(x)$. Let us write $\mathcal{A}_t$ for the event that $\tau < \nu_0 + t$, for $t > 0$ to be specified, and write H_t for the halfspace $\{x : \langle e_1, x\rangle \leq \nu_0 + t\}$. Given any $x \in H$, write $\Delta = d(x, H_t^c)$, and let B be a ball of radius 2Δ around x, intersected with Ω.

Denote by $\mathcal{F}(x)$ the event that there are no samples in $V = B \cap H_t$ but there is at least one point in $B \cap H_t^c$. Then $\mathcal{F}(x) \cap \mathcal{A}_t \subseteq \mathcal{E}(x)$, since on $\mathcal{F}(x)$ the nearest neighbor to x must be a sample in H_t^c, and on $\mathcal{A}_t$ we have $H_t^c \subseteq \hat{H}^c$.

Lemma G.2.

$$\mathbb{P}\left\{\mathcal{F}(x) \cap \mathcal{A}_t\right\} \geq (1 - \mathrm{vol}(V))^n - (1 - \mathrm{vol}(B))^n - e^{-2nt^2}\,.$$

Proof. We first compute $\mathbb{P}\left\{\mathcal{F}(x)\right\}$. The probability that there are no samples in V is $(1 - \mathrm{vol}(V))^n$, and this event may be written as the disjoint union of $\mathcal{F}(x)$ and the event that all of B is empty. The latter event has probability $(1 - \mathrm{vol}(B))^n$. Therefore

$$(1 - \mathrm{vol}(V))^n = \mathbb{P}\left\{\mathcal{F}(x)\right\} + (1 - \mathrm{vol}(B))^n\,.$$

Since $\mathbb{P}\left\{\mathcal{A}_t^c\right\} \leq e^{-2nt^2}$, the claim follows. □

We need the following lemma.

Lemma G.3. *Assume that $\Delta > 0$ and that $d(x, \partial\Omega) \geq 2\Delta$. There exist positive constants $c_{d,0} < 1$ and $c_{d,1}$ such that*

$$\mathrm{vol}(V) \leq c_{d,0}\,\mathrm{vol}(B) \tag{68}$$

and

$$\mathrm{vol}(B) \geq c_{d,1}\Delta^d \tag{69}$$

Proof. This is immediate from a scaling argument: since $d(x, \partial\Omega) \geq 2\Delta$, the set B is a Euclidean ball of radius 2Δ, and the set V is a Euclidean ball of radius 2Δ minus a spherical dome cut off by a hyperlane at distance Δ from the center. When $\Delta = 1$, it is clear that the claimed inequalities hold, and the general case is obtained by dilation. □

We assume in what follows that $d(x, \partial\Omega) \geq 2\Delta$. The inequalities $(1+x)^n \geq 1 + nx$ and $e^x \leq 1 + x + x^2$, valid for all $x \in [-1, 0]$ and $n \geq 1$, imply that for any $\delta > 0$ there exists a constant $c_{d,\delta} > 0$ such that if $\Delta \leq c_{d,\delta} n^{-1/d}$, then we will have

$$(1 - \mathrm{vol}(V))^n \geq 1 - n c_{d,0} \,\mathrm{vol}(B) \tag{70}$$

$$(1 - \mathrm{vol}(B))^n \leq e^{-n\,\mathrm{vol}(B)} \leq 1 - (1-\delta) n \,\mathrm{vol}(B) \tag{71}$$

Choosing δ sufficiently small, we obtain the existence of a small $c_{d,3} > 0$ such that if $\Delta \leq c_{d,3} n^{-1/d}$, then

$$(1 - \mathrm{vol}(V))^n - (1 - \mathrm{vol}(B))^n \geq C_d n \Delta^d \,.$$

Define $\Delta_n = c_{d,4} n^{-1/d}$. Putting it all together, consider the set

$$S = \{x \in H \cap \Omega : \Delta_n/2 \leq d(x, H_t^c) \leq \Delta_n, d(x, \partial\Omega) \geq 2\Delta_n\} \,.$$

The above considerations imply that $\mathbb{P}\{\mathcal{E}(x)\} \geq C_d n (\Delta_n/2)^d - e^{-2nt^2} \geq C_d' - e^{-2nt^2}$ for all $x \in S$. Choosing t to be a sufficiently large constant multiple of $n^{-1/2}$, we obtain

$$\int_H \mathbb{P}\{\mathcal{E}(x)\} \,\mathrm{d}P(x) \geq \int_S \mathbb{P}\{\mathcal{E}(x)\} \,\mathrm{d}P(x) \gtrsim_d \mathrm{vol}(S) \,.$$

Since $t \asymp n^{-1/2}$, we will have that $t \ll \Delta_n$ for n sufficiently large (as $d \geq 3$). Therefore, for n large enough, the set S contains the set

$$S' = \{x \in \Omega : \nu_0 - \Delta_n + t \leq \langle e_1, x\rangle \leq \nu_0 - \Delta_n/2 + t, 2\Delta_n \leq \langle e_j, x\rangle \leq 1 - 2\Delta_n \quad \forall j = 2, \ldots, d\} \,.$$

Since $\mathrm{vol}(S') \gtrsim_d \Delta_n \gtrsim n^{-1/d}$, the claim follows.

H. Auxiliary lemmas

Lemma H.1 (Young's inequality)**.** *Let Q_0, Q_1 be probability measures with $Q_1 \ll Q_0$ and let f be a function. Then, for $\theta > 0$,*

$$\int f(\,\mathrm{d}Q_0 - \,\mathrm{d}Q_1) \leq \frac{\theta \mathrm{Var}_{Q_0}(f)}{2} + \frac{\chi^2(Q_1 \| Q_0)}{2\theta}. \tag{72}$$

Proof. Recall Young's inequality: for $a, b \in \mathbb{R}$, $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$. As the left-hand side is invariant by translation, we may assume without loss of generality that $\int f \,\mathrm{d}Q_0 = 0$, so that $\mathrm{Var}_{Q_0}(f) = \int f^2 \,\mathrm{d}Q_0$. We write

$$\begin{aligned}\int f(\,\mathrm{d}Q_0 - \,\mathrm{d}Q_1) &= \int (\sqrt{\theta} f) \frac{\left(1 - \frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}\right)}{\sqrt{\theta}} \,\mathrm{d}Q_0 \leq \frac{\theta}{2} \int f^2 \,\mathrm{d}Q_0 + \frac{1}{2\theta} \int \left(1 - \frac{\mathrm{d}Q_1}{\mathrm{d}Q_0}\right)^2 \,\mathrm{d}Q_0 \\ &= \frac{\theta \mathrm{Var}_{Q_0}(f)}{2} + \frac{\chi^2(Q_1 \| Q_0)}{2\theta}.\end{aligned}$$

□

Lemma H.2 (Expectation of empirical χ^2-divergence)**.** *Let $Q = \sum_{j=1}^J q_j \delta_{y_j}$ be a discrete measure supported on J atoms, and let Q_n denote its empirical measure, consisting of n i.i.d. samples. Then,*

$$\mathbb{E}[\chi^2(Q_n \| Q)] = \frac{J-1}{n} \,. \tag{73}$$

Proof. We can write $Q_n = \sum_{j=1}^{J} \hat{q}_j \delta_{y_j}$, where $\hat{q}_j$ is a binomial random variable with parameters n and q_j. We obtain

$$\chi^2(Q_n \| Q) = \sum_{j=1}^{J} \frac{(\hat{q}_j - q_j)^2}{q_j} .$$

Taking expectations, our bound reads

$$\mathbb{E}[\chi^2(Q_n \| Q)] = \sum_{j=1}^{J} \frac{\mathrm{Var}(\hat{q}_j)}{q_j} = \sum_{j=1}^{J} \frac{q_j(1-q_j)}{nq_j} = \frac{J-1}{n}.$$

□

Lemma H.3 (Control of suprema of empirical processes)**.** *Let $X_1, \dots, X_n$ be an i.i.d. sample from some probability measure P on $\mathbb{R}^d$, with P_n the associated empirical measure. Consider $\mathcal{F}$ a class of functions $\mathbb{R}^d \to \mathbb{R}$ with $\|f\|_\infty \leq A$ for all $f \in \mathcal{F}$. For $u > 0$, let $N(u)$ be the u-covering numbers of $\mathcal{F}$, that is the minimal number of balls of radius u for the $\|\cdot\|_\infty$-metric required to cover $\mathcal{F}$. Then,*

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}(P_n - P) \right| \right] \leq \frac{C_0}{\sqrt{n}} \int_0^{C_1 A} \sqrt{\log 2N(u)} \, \mathrm{d}u =: \frac{I}{\sqrt{n}} \tag{74}$$

for two positive absolute constants C_0 and C_1. Furthermore, for all $t > 0$,

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}(P_n - P) \right| > t \right) \leq \exp\left(-\frac{C_2 \sqrt{n} t}{I + A \log n} \right), \tag{75}$$

for some positive absolute constant C_2. Eventually, for all $p \geq 2$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \int f \, \mathrm{d}(P_n - P) \right|^p \right]^{1/p} \leq C_p \frac{I + A}{\sqrt{n}}. \tag{76}$$

Proof. See (Vaart & Wellner, 1996, Theorem 2.14.2 and Theorem 2.14.5). □