Estimation Rates for Optimal Transport Maps to the Barycenter in Multidimensional Spaces

Anonymous authors
Paper under double-blind review

Abstract

The optimal transport barycenter problem seeks transport maps that align multiple distributions with their barycenter, a central object in applications such as image interpolation and fair regression. Although several algorithms estimate these maps in continuous multidimensional spaces, their statistical estimation rates remain largely unknown. We establish, to our knowledge, the first convergence rate for optimal transport map estimation to the barycenter under the quadratic cost in multidimensional Euclidean spaces. Crucially, the rate is governed solely by the approximation error of the estimator's class, which may be non-convexly parametrized, as with deep neural networks, and requires no regularity of the underlying transport maps. The resulting rate interpolates between the fast rate under the Poincaré inequality and the slow rate in general, two extremes previously established only in isolation for transport map estimation between two distributions.

1 Introduction

The optimal transport barycenter serves as a summary of multiple distributions in statistics and machine learning, with applications including smooth shape interpolation (Solomon et al., 2015), image interpolation (Lacombe et al., 2023; Simon & Aberdam, 2020; Rabin et al., 2012), clustering (Ho et al., 2017; del Barrio et al., 2019), and fair regression (Chzhen et al., 2020; Chzhen & Schreuder, 2022). For probability measures ν and ν' on a measurable space $(\Omega, \mathcal{F})$ and a cost function $c : \Omega \times \Omega \to \mathbb{R}_+$, the Monge formulation of the optimal transport cost is

$$\mathbb{M}_c(\nu, \nu') = \inf\left\{ \int c(x, T(x))\nu(dx) : T \in L^1(\nu), T\sharp\nu = \nu' \right\},$$

where $L^1(\nu)$ is the space of ν-integrable functions, and $T\sharp\nu$ is the pushforward of ν under T. We refer to any T attaining the above infimum as an optimal transport map from ν to ν'. Given probability measures $\nu^{(1)}, \dots, \nu^{(K)}$ and positive weights $\lambda^{(1)}, \dots, \lambda^{(K)}$ summing to one, the barycenter with respect to the optimal transport cost $\mathbb{M}_c$ is a probability measure $\bar{\nu}$ attaining the following infimum:

$$\inf\left\{ \sum_{i=1}^{K} \lambda^{(i)} \mathbb{M}_c(\nu^{(i)}, \nu) : \nu \in \mathcal{P}(\Omega) \right\}, \tag{1}$$

where $\mathcal{P}(\Omega)$ is the set of all Borel probability measures on Ω. The existence and uniqueness of $\bar{\nu}$ under suitable regularity conditions are well established (Agueh & Carlier, 2011).

In many of these applications, the object of interest is not the barycenter itself but the transport maps that align each distribution with it. In image interpolation, for example, the transport maps construct an interpolated image directly from two given images instead of sampling from an interpolated distribution. In fair regression, they are used to construct a regression function that aligns the output distributions of different demographic groups. Estimating these maps is therefore often the primary goal, and the barycenter serves only as an intermediate object for aligning the distributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Several algorithms have been proposed for estimating the transport maps to the barycenter. Korotin et al. (2021) estimate them via potential minimization, Fan et al. (2021) develop a scalable method based on input-convex neural networks, Korotin et al. (2022) use an iterative scheme, Kim et al. (2025) formulate a minimax optimization problem, and Visentin & Cheridito (2025) employ normalizing flows.

In contrast, the statistical error of estimating transport maps to the barycenter remains largely unexplored. Fukuchi (2025) proposed an estimator of these maps in the context of fair regression and established its convergence rate. Because of this application, however, their estimator and analysis are restricted to probability measures on the real line. Portales et al. (2025) derived the sample complexity of this problem in multidimensional spaces, but their estimator and analysis are limited to discretely supported probability measures.

Our contributions This paper constructs estimators of the transport maps to the barycenter and derives their convergence rates for $\Omega = \mathbb{R}^p$ and the quadratic cost $c(x, y) = \|x-y\|^2/2$, where $\|\cdot\|$ is the Euclidean norm. Our contributions are as follows.

- To our knowledge, we establish the first convergence rate for estimating transport maps to the barycenter in multidimensional Euclidean spaces (Thm. 2). Unlike existing results (Fukuchi, 2025; Portales et al., 2025), which are limited to one-dimensional or discrete probability measures, our analysis covers general probability measures on $\mathbb{R}^p$.
- Our convergence rate interpolates between the fast rate under the Poincaré inequality and the slow rate without it. These two regimes were previously established only separately for transport map estimation between two distributions: the fast rate by Hütter & Rigollet (2021); Divol et al. (2025) and the slow rate by Li et al. (2025). We achieve this interpolation through a generalized Poincaré inequality of order $\alpha \in [0, 1]$: $\alpha = 1$ recovers the usual Poincaré inequality, $\alpha = 0$ imposes no assumption on the underlying measures, and the rate improves continuously as α increases.
- Our convergence rate is fully characterized by the potential approximation error of the estimator's class with respect to the underlying transport maps. Crucially, our result yields a non-trivial rate whenever a sequence of complexity-controlled function classes approaches the underlying transport maps, without requiring any regularity of the maps themselves. This extends the scope beyond the settings considered in existing work on transport map estimation between two measures (e.g., Hütter & Rigollet, 2021; Divol et al., 2025).
- Our estimator minimizes an empirical potential over a class of transport-map tuples that may be non-convexly parametrized. It therefore accommodates the deep neural network parametrizations widely used in practice.

2 Related Work

Optimal transport map estimation between two probability measures Estimation of the optimal transport map between two probability measures has been studied extensively. Hütter & Rigollet (2021) established minimax rates for optimal transport map estimation when the transport map belongs to a Hölder class with upper- and lower-bounded gradient norms and compact support. Considerable effort has been devoted to establishing convergence rates in more general settings (DEB et al., 2021; Vacher & Vialard, 2023; Manole et al., 2024; Pooladian et al., 2023; Muzellec et al., 2021; Gunsilius, 2022; Divol et al., 2025; Li et al., 2025). Convergence rates have also been obtained for entropically regularized optimal transport map estimation (Pooladian & Niles-Weed, 2024; Rigollet & Stromme, 2025). These techniques do not directly apply to transport maps to the barycenter, because the barycenter is unknown and is determined jointly by all input distributions.

Estimation of optimal transport maps to the barycenter Only a few studies have analyzed the estimation error of the optimal transport map to the barycenter (Fukuchi, 2025; Portales et al., 2025). As discussed in the introduction, these analyses are limited to one-dimensional or discrete probability measures. Our work fills this gap by establishing the convergence rate in multidimensional Euclidean spaces.

3 Problem Formulation and Preliminaries

Notation For a positive integer m, let $[m] = \{1, \dots, m\}$. We write $\mathbb{1}$ for the indicator function, and $a \vee b$ and $a \wedge b$ for the maximum and the minimum of $a, b \in \mathbb{R}$, respectively. The Euclidean inner product and norm on $\mathbb{R}^p$ are denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, respectively, and B_R denotes the closed ball of radius $R > 0$ centered at the origin. We denote by $\mathcal{P}_2(\Omega)$ the set of probability measures on Ω with finite second moments, and by $C_b^2(\Omega)$ the set of continuous functions on Ω with at most quadratic growth. For a probability measure ν, we denote the expectation and the variance under ν by $\mathbb{E}_\nu$ and $\mathbb{V}_\nu$, and the norm of $L^q(\nu)$ by $\| \cdot \|_{L^q(\nu)}$ for $q \in [1, \infty]$. For a set $\Omega' \subseteq \Omega$, $\| \cdot \|_{L^\infty(\Omega')}$ denotes the supremum norm over Ω'. For a vector-valued function f, we write $\| \|f\| \|_{L^q(\nu)}$ and $\| \|f\| \|_{L^\infty(\Omega')}$ for the corresponding norms of $x \mapsto \|f(x)\|$. For a function $f : \mathbb{R}^p \to \mathbb{R}$, we denote its gradient and its convex conjugate by ∇f and f^*, respectively. For a set A equipped with a norm $\| \cdot \|$, $N(\epsilon, A, \| \cdot \|)$ denotes the ϵ-covering number of A. Rigorous definitions are given in Appendix A.

Problem formulation We consider estimating the transport maps to the barycenter on $\Omega = \mathbb{R}^p$ with the quadratic cost $c(x, y) = \|x - y\|^2/2$. Let $\mathcal{Q} \subseteq \mathcal{P}_2(\Omega)$ be a class of candidate probability measures. Given a K-tuple of probability measures $\nu^{(\cdot)} = (\nu^{(1)}, \dots, \nu^{(K)}) \in \mathcal{Q}^K$, let $\bar{\nu}^{(\cdot)}$ be the barycenter of $\nu^{(\cdot)}$ with respect to $\mathbb{M}_c$ as defined in Eq. (1), and let $T_\nu^{(i)} : \Omega \to \Omega$ be the optimal transport map from $\nu^{(i)}$ to the barycenter $\bar{\nu}^{(\cdot)}$. For each $i \in [K]$, the learner observes n_i i.i.d. samples from $\nu^{(i)}$. We denote the jth sample from $\nu^{(i)}$ by $X_j^{(i)} \in \Omega$ and write $X^{(i)} = (X_1^{(i)}, \dots, X_{n_i}^{(i)})$ and $X^{(\cdot)} = (X^{(1)}, \dots, X^{(K)})$. Given predetermined weights $\lambda^{(\cdot)} = (\lambda^{(1)}, \dots, \lambda^{(K)})$, the goal is to estimate $T_\nu^{(\cdot)} = (T_\nu^{(1)}, \dots, T_\nu^{(K)})$ from $X^{(\cdot)}$. We denote the estimated transport maps by $T_n^{(\cdot)} = (T_n^{(1)}, \dots, T_n^{(K)})$. We measure the estimation error by the weighted squared deviation

$$\left\| \left\| T_n^{(\cdot)} - T_\nu^{(\cdot)} \right\| \right\|^2_{L^2(\nu^{(\cdot)})} = \sum_{i=1}^K \lambda^{(i)} \left\| \left\| T_n^{(i)} - T_\nu^{(i)} \right\| \right\|^2_{L^2(\nu^{(i)})}.$$

Let $\mathcal{T}_\nu(\mathcal{Q}) = \{T_\nu^{(\cdot)} : \nu^{(\cdot)} \in \mathcal{Q}^K\}$ denote the set of transport maps induced by $\mathcal{Q}$, abbreviated to $\mathcal{T}_\nu$ when $\mathcal{Q}$ is clear from the context. We denote the expectation over the samples by $\mathbb{E}_{n\nu}$.

Characterization of transport maps to the barycenter Agueh & Carlier (2011) characterized the optimal transport maps to the barycenter under the quadratic cost. Specifically, the gradients of the convex conjugates of the minimizing potentials are the optimal transport maps to the barycenter[1].

Theorem 1 (Agueh & Carlier (2011)). *Given $\nu^{(\cdot)} \in \mathcal{P}_2(\Omega)^K$ with $\lambda^{(i)} > 0$ for all $i \in [K]$, assume that $\nu^{(i)}$ vanishes on small sets for all $i \in [K]$. Let $\phi_\nu^{(\cdot)}$ be a K-tuple that solves*

$$\inf \left\{ \sum_{i=1}^K \lambda^{(i)} \int \left(\phi^{(i)}\right)^* d\nu^{(i)} : \phi^{(i)} \in C_b^2(\Omega) \text{ convex and l.s.c.}, \sum_{i=1}^K \lambda^{(i)} \phi^{(i)} = \frac{1}{2}\|\cdot\|^2 + C \right\}, \quad (2)$$

for some constant $C \in \mathbb{R}$. Then, $\nabla(\phi_\nu^{(i)})^*$ is the unique optimal transport map from $\nu^{(i)}$ to the barycenter $\bar{\nu}^{(\cdot)}$ up to $\nu^{(i)}$-null sets.

Following Korotin et al. (2021), we call potentials satisfying $\sum_{i=1}^K \lambda^{(i)} \phi^{(i)} = \frac{1}{2}\|\cdot\|^2 + C$ *congruent*.

4 Main Result

This section establishes the convergence rate for estimating transport maps to the barycenter. The rate is governed by two factors: the Poincaré inequality satisfied by the underlying probability measures and the approximation error of the transport maps. We first introduce these two notions and then present the main result.

[1]Since Thm. 1 is stated differently from the original result of Agueh & Carlier (2011), we provide a proof in Appendix C.

Poincaré inequality Existing convergence rates for transport map estimation between two measures fall into two regimes: a fast rate under the Poincaré inequality (e.g., Hütter & Rigollet, 2021; Divol et al., 2025) and a slow rate without it (Li et al., 2025). We derive our rate under an assumption that interpolates between these two regimes.

Definition 1 (Poincaré inequality of order α)**.** *Let $\alpha \in [0,1]$. A probability measure ν satisfies the Poincaré inequality of order α with constant $C_\alpha < \infty$ if for every locally Lipschitz $f \in L^2(\nu)$,*

$$\mathbb{V}_\nu[f] \le C_\alpha \|\|\nabla f\|\|_{L^2(\nu)}^{2\alpha} \|f\|_{L^\infty(\nu)}^{2(1-\alpha)}.$$

We assume that every probability measure in $\mathcal{Q}$ satisfies the Poincaré inequality of order α with a common constant C_α. The case $\alpha = 1$ recovers the usual Poincaré inequality, whereas $\alpha = 0$ imposes no assumption, since $\mathbb{V}_\nu[f] \le \|f\|_{L^\infty(\nu)}^2$ for every $f \in L^2(\nu)$.

Potential approximation error of transport maps Our rate is characterized by how well a class of transport maps with specific properties approximates the underlying maps in terms of the potential. Building on Thm. 1, we define the potential functions of a tuple $T^{(\cdot)}$ and of the underlying maps $T_\nu^{(\cdot)}$ as

$$\Phi\left(T^{(\cdot)}\right) = \sum_{i=1}^K \lambda^{(i)} \int \phi^*\left[T^{(i)}\right] d\nu^{(i)}, \text{ and } \Phi_\nu = \sum_{i=1}^K \lambda^{(i)} \int \left(\phi_\nu^{(i)}\right)^* d\nu^{(i)},$$

respectively, where each $\phi_\nu^{(i)} : \Omega \to \mathbb{R}$ is a convex function with $\nabla(\phi_\nu^{(i)})^* = T_\nu^{(i)}$, the tuple $\phi_\nu^{(\cdot)}$ is congruent with $C = 0$ in Thm. 1, and

$$\phi[T](x) = \left\langle \int_0^1 T^{-1}(tx)dt, x \right\rangle. \tag{3}$$

The approximation error of $T_\nu^{(\cdot)}$ by a class $\mathcal{T}$ is defined as

$$A\left(T_\nu^{(\cdot)}; \mathcal{T}\right) = \inf\left\{\Phi\left(T^{(\cdot)}\right) - \Phi_\nu : T^{(\cdot)} \in \mathcal{T}\right\}.$$

We consider classes with the following properties.

Definition 2. *Given $R, L, D \ge 1$, a class of K-tuples of continuous bijective gradient maps $\mathcal{T}$ is an (R, L, D)-*class*, denoted by $\mathcal{T}_{R,L,D}$, if it satisfies the following conditions:*

a. (congruent) for all $T^{(\cdot)} \in \mathcal{T}$ and all $x \in \Omega$,

$$\sum_{i=1}^K \lambda^{(i)} \left(T^{(i)}\right)^{-1}(x) = x;$$

b. (R-bounded zero) for all $T^{(\cdot)} \in \mathcal{T}$ and all $i \in [K]$,

$$\left\| \left(T^{(i)}\right)^{-1}(0) \right\| \le R;$$

c. (monotone and L-Lipschitz) for all $T^{(\cdot)} \in \mathcal{T}$, all $i \in [K]$, and all $x, y \in \Omega$,

$$0 \le \left\langle T^{(i)}(x) - T^{(i)}(y), x - y \right\rangle \le L\|x - y\|^2;$$

d. (D-bounded covering entropy[2]) there exists $\beta \in [1, \infty)$ such that

$$\limsup_{\epsilon \to 0} \frac{\ln N(\epsilon, \mathcal{T}_{R,L,D}, \|\|\cdot\|\|_{\bar{L}^\infty(B_R)})}{\ln^\beta(1/\epsilon)} \le D,$$

where $\|\|T^{(\cdot)}\|\|_{\bar{L}^\infty(B_R)}^2 = \sum_{i=1}^K \lambda^{(i)} \|\|T^{(i)}\|\|_{L^\infty(B_R)}^2$.

[2]The logarithm of the covering number is referred to as the covering entropy.

The first condition mirrors the congruency of the potentials in Thm. 1. The second and third conditions ensure the stability of the estimation procedure. The fourth condition controls the complexity of the class. Roughly speaking, it requires that the class admit a D-dimensional parametrization. Indeed, if there is a Lipschitz map from a d-dimensional parameter space onto the class, then Def. 2 (d) is satisfied with $D = d$. Many common neural network architectures are known to have a finite covering entropy, including ReLU networks (Schmidt-Hieber, 2020; Ou & Bölcskei, 2026), recurrent neural networks (Chen et al., 2020; F. Pour & Ashtiani, 2023), and transformers (Takakura & Suzuki, 2023; Havrilla & Liao, 2024).

Convergence rate of transport map estimation to the barycenter We now state our main result.

Theorem 2. *Given constants $\alpha \in [0,1]$ and $C_\alpha > 0$, suppose that every probability measure in $\mathcal{Q}$ vanishes on small sets and satisfies the Poincaré inequality of order α with the constant C_α. Let $\tilde{n} = \min_{i\in[K]} n_i/\lambda^{(i)}$. Let $R, L, D \geq 1$, possibly depending on $\tilde{n}$. Assume that $LR = o(\tilde{n})$ and $\Xi_R = o(\frac{R^2 D}{\tilde{n}})$, where*

$$\Xi_R^{(i)} = \mathbb{E}_{\nu^{(i)}}\left[\left\|X^{(i)}\right\|^2 \mathbb{1}\left\{X^{(i)} \notin B_R\right\}\right], \text{ and } \Xi_R = \sum_{i=1}^K \lambda^{(i)} \Xi_R^{(i)}. \tag{4}$$

Then, there exist an estimator $T_n^{(\cdot)}$ and $\beta \in [1,\infty)$ such that for any $\nu^{(\cdot)} \in \mathcal{Q}^K$, if $A\left(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}\right) = 0$,

$$\mathbb{E}_{n\nu^{(\cdot)}}\left[\left\|T_n^{(\cdot)} - T_\nu^{(\cdot)}\right\|^2_{L^2(\nu^{(\cdot)})}\right] \leq$$

$$2L(1+o(1))\left(2C_\alpha \frac{14+\sqrt{33}}{12}\left(LR^2\left(1 \vee \left(\frac{(LR^2)^{2(1-\alpha)}}{L^{1-\alpha}R^2}\right)^2\right)\right)\frac{D\ln^\beta(\tilde{n}/LR)}{\tilde{n}}\right)^{\frac{1}{2-\alpha}},$$

and if $A\left(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}\right) = \omega(((LR^2(1 \vee (\frac{(LR^2)^{2(1-\alpha)}}{L^{1-\alpha}R^2})^2))\frac{D\ln^\beta(\tilde{n}/LR)}{\tilde{n}})^{\frac{1}{2-\alpha}})$,

$$\mathbb{E}_{n\nu^{(\cdot)}}\left[\left\|T_n^{(\cdot)} - T_\nu^{(\cdot)}\right\|^2_{L^2(\nu^{(\cdot)})}\right] \leq 2L(1+o(1))A\left(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}\right).$$

In Thm. 2, $o(\cdot)$ denotes the little-o notation as $\tilde{n}$ increases. We highlight some implications:

1. Thm. 2 establishes a convergence rate for estimating transport maps to the barycenter in multidimensional spaces, governed mainly by the approximation error $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$. When $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) = 0$, the rate is of order $((LR^2(1 \vee (\frac{(LR^2)^{2(1-\alpha)}}{L^{1-\alpha}R^2})^2))D\ln^\beta(\tilde{n}/LR)/\tilde{n})^{1/(2-\alpha)}$. When $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) = O(D^{-\gamma})$ for some $\gamma > 0$, the rate with the optimal D is of order $((LR^2(1 \vee (\frac{(LR^2)^{2(1-\alpha)}}{L^{1-\alpha}R^2})^2))\ln^\beta(\tilde{n}/LR)/\tilde{n})^{\gamma/(\gamma(2-\alpha)+1)}$.
2. Thm. 2 imposes no condition on the underlying transport maps $T_\nu^{(\cdot)}$ themselves. All conditions in Def. 2 are imposed on the estimator's class $\mathcal{T}_{R,L,D}$. In particular, $T_\nu^{(\cdot)}$ need not be Lipschitz or strongly monotone, and it influences the rate only through the potential approximation error $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$.
3. The exponent $1/(2-\alpha)$ interpolates between the two regimes known for transport map estimation between two distributions. Under the usual Poincaré inequality ($\alpha = 1$), the statistical term is of order $D\ln\tilde{n}/\tilde{n}$, matching the fast rate of Hütter & Rigollet (2021); Divol et al. (2025). Without any assumption ($\alpha = 0$), it is of order $\sqrt{D\ln\tilde{n}/\tilde{n}}$, matching the slow rate of Li et al. (2025). For $\alpha \in (0,1)$, the rate improves continuously as α increases.
4. Thm. 2 does not require compactly supported measures. The tails of $\nu^{(\cdot)}$ affect the rate only through Ξ_R. If all the supports are contained in B_{R_0}, then $\Xi_R = 0$ for $R \geq R_0$, and

the constraint involving $L\Xi_R$ becomes vacuous. For light-tailed probability measures, such as sub-Gaussian ones, Ξ_R decays exponentially in R^2, and thus choosing R of order $\sqrt{\ln \tilde{n}}$ makes $L\Xi_R$ negligible at the cost of polylogarithmic factors.

Remark 1. *To attain the optimal rate in Thm. 2, we need to select the parameters (R, L, D) optimally. The optimal parameters depend on the unknown distributions. In practice, (R, L, D) can be selected in a data-driven manner, for example by cross-validation.*

5 Estimator and Analysis

In this section, we introduce the estimator and show that it achieves the convergence rate in Thm. 2. The estimator solves the empirical counterpart of the potential minimization in Thm. 1. Specifically, it is a tuple of transport maps attaining the infimum

$$\inf \left\{ \Phi_n\left(T^{(\cdot)}\right) \coloneqq \sum_{i=1}^{K} \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \phi^*\left[T^{(i)}\right]\left(X_j^{(i)}\right) : T^{(\cdot)} \in \mathcal{T}_{R,L,D} \right\}. \tag{5}$$

We assume that a minimizer of Eq. (5) exists and can be chosen measurably in the samples, and we define $T_n^{(\cdot)}$ to be such a minimizer.

Overview of the analysis The estimator in Eq. (5) is sensible only if a small potential error implies a small estimation error. We establish this in Section 5.1 through a stability bound relating the two errors. This bound extends the stability of optimal transport maps between two measures and relies on the monotonicity and Lipschitzness of the class. The key difference is that the dependence on the unknown barycenter $\bar{\nu}$ must be eliminated, which the construction of the potential in Eq. (3) achieves.

Given this stability, a bound on the deviation between the empirical and population potentials yields the convergence rate of the transport map error. We obtain such a bound from concentration inequalities for empirical processes. The main obstacle is that the potential is unbounded, whereas standard concentration inequalities require boundedness. We overcome this by truncation, a technique also used for transport map estimation between two measures. Section 5.2 presents the details.

5.1 Stability

We show that the stability bound holds for any transport maps satisfying Def. 2 (a) and Def. 2 (c).

Theorem 3 (Stability)**.** *Let $\nu^{(\cdot)} \in \mathcal{P}_2(\Omega)^K$ vanish on small sets. Let $T^{(\cdot)}$ be a K-tuple of continuous bijective gradient maps from Ω to Ω such that congruency in Def. 2 (a) and monotonicity and L-Lipschitzness in Def. 2 (c) hold for some $L > 0$. Then, we have*

$$\left\| \left\| T^{(\cdot)} - T_\nu^{(\cdot)} \right\| \right\|^2_{L^2(\nu^{(\cdot)})} \leq 2L\left(\Phi\left(T^{(\cdot)}\right) - \Phi_\nu\right).$$

By Thm. 3, the transport map error is bounded by the potential error up to the factor $2L$ for L-Lipschitz maps. Notably, the bound holds even if $T_\nu^{(\cdot)}$ is not Lipschitz. This theorem is an extension of the result of Fukuchi (2025) to the multidimensional case. The proof of Thm. 3 is deferred to Appendix E.

5.2 Potential Error Analysis

This subsection bounds the potential error $\Phi(T_n^{(\cdot)}) - \Phi_\nu$ as follows.

Theorem 4. *Suppose that $\nu^{(\cdot)} \in \mathcal{P}_2(\Omega)^K$ vanishes on small sets and satisfies the Poincaré inequality of order α with constant C_α. Let $T_n^{(\cdot)}$ be the estimator defined by Eq. (5). Given $\epsilon > 0$, let $H_\epsilon = \ln N(\epsilon, \mathcal{T}_{R,L,D}, \|\cdot\|_{\bar{L}^\infty(B_R)})$. Let $C_{L,R} = \frac{C_\alpha (4LR^2)^{2(1-\alpha)}}{(2L)^{1-\alpha} R^2}$. Define*

$$E_{\tilde{n},\epsilon,L,R} = 4LR^2(1 \vee C_{L,R})^2 \frac{1 + H_\epsilon}{\tilde{n}}.$$

For $\gamma_{\tilde{n}} \in (0,1)$, assume $\epsilon \le \gamma_{\tilde{n}} \frac{LR(1+H_\epsilon)}{\tilde{n}}$ and $\Xi_R \le \gamma_{\tilde{n}} \frac{R^2(1+H_\epsilon)}{\tilde{n}}$. Then,

1. if $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) = 0$ and $E_{\tilde{n},\epsilon,L,R} \le 1/5$,

$$\mathbb{E}_{n\nu^{(\cdot)}}\left[\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu\right] \le \left(\frac{14+\sqrt{33}}{12}\left(1+2\gamma_{\tilde{n}}^{1/4}\right)^2 E_{\tilde{n},\epsilon,L,R}\right)^{\frac{1}{2-\alpha}},$$

2. if $E_{\tilde{n},\epsilon,L,R} \le \gamma_{\tilde{n}} A^{2-\alpha}(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$,

$$\mathbb{E}_{n\nu^{(\cdot)}}\left[\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu\right] \le \left(1 + 3\sqrt{\gamma_{\tilde{n}}} + 20\gamma_{\tilde{n}} A^{1-\alpha}(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})\right) A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}).$$

Since $\gamma_{\tilde{n}}$ is intended to be $o(1)$ as $\tilde{n}$ increases, Thm. 4 makes the leading constants in the potential error bounds explicit.

Let $T_{\mathcal{T}}^{(\cdot)} \in \mathcal{T}_{R,L,D}$ be a minimizer in the definition of $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$[3]. We decompose the potential error into three terms:

$$\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu = A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) + \left(\Phi_n\left(T_n^{(\cdot)}\right) - \Phi_n\left(T_{\mathcal{T}}^{(\cdot)}\right)\right) + \left[\left(\Phi\left(T_n^{(\cdot)}\right) - \Phi\left(T_{\mathcal{T}}^{(\cdot)}\right)\right) - \left(\Phi_n\left(T_n^{(\cdot)}\right) - \Phi_n\left(T_{\mathcal{T}}^{(\cdot)}\right)\right)\right]. \quad (6)$$

The second term of Eq. (6) is non-positive because $T_n^{(\cdot)}$ minimizes Φ_n over $\mathcal{T}_{R,L,D}$. Hence,

$$\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu \le A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) + \left[\left(\Phi\left(T_n^{(\cdot)}\right) - \Phi\left(T_{\mathcal{T}}^{(\cdot)}\right)\right) - \left(\Phi_n\left(T_n^{(\cdot)}\right) - \Phi_n\left(T_{\mathcal{T}}^{(\cdot)}\right)\right)\right] \quad (7)$$

The first term on the right-hand side of Eq. (7) is the approximation error incurred by restricting the estimator to $\mathcal{T}_{R,L,D}$. To interpret the second term, which we denote by Δ_n, we rewrite it as

$$\Delta_n = \sum_{i=1}^{K} \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \bar{g}_n^{(i)}\left(X_j^{(i)}\right),$$

where

$$g_n^{(i)} = \phi^*\left[T_{\mathcal{T}}^{(i)}\right] - \phi^*\left[T_n^{(i)}\right], \text{ and } \bar{g}_n^{(i)} = g_n^{(i)} - \int g_n^{(i)} d\nu^{(i)}.$$

Thus, Δ_n is a centered multi-sample empirical process evaluated at the data-dependent functions $g_n^{(\cdot)}$.

A concentration bound on Δ_n would yield the desired bound on the potential error, but the unboundedness of $g_n^{(i)}$ precludes standard concentration inequalities. To overcome this, we truncate $g_n^{(i)}$ by clipping its values to $[-M, M]$ and decompose $g_n^{(i)}$ as $g_n^{(i)} = g_{M,n}^{(i)} + g_{M^c,n}^{(i)}$, where

$$g_{M,n}^{(i)}(x) = \left(g_n^{(i)}(x) \wedge M\right) \vee (-M), \text{ and } g_{M^c,n}^{(i)}(x) = g_n^{(i)}(x) - g_{M,n}^{(i)}(x). \quad (8)$$

This induces the decomposition $\Delta_n = \Delta_{M,n} + \Delta_{M^c,n}$, where

$$\Delta_{M,n} = \sum_{i=1}^{K} \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \bar{g}_{M,n}^{(i)}\left(X_j^{(i)}\right), \text{ and } \Delta_{M^c,n} = \sum_{i=1}^{K} \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \bar{g}_{M^c,n}^{(i)}\left(X_j^{(i)}\right). \quad (9)$$

Since $\Delta_{M,n}$ involves only bounded functions, standard concentration inequalities apply to it.

[3] If the infimum in $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$ is not attained, the same argument applies to $T_{\mathcal{T}}^{(\cdot)} \in \mathcal{T}_{R,L,D}$ with $\Phi(T_{\mathcal{T}}^{(\cdot)}) - \Phi_\nu \le A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) + \eta$ for an arbitrary $\eta > 0$, and letting $\eta \to 0$ yields the same bound.

Lemma 1. *Let $\Delta_{M,n}$ be as defined in Eq. (9) with $M = 4LR^2$. Given $\epsilon > 0$, let $H_\epsilon = \ln N(\epsilon, \mathcal{T}_{R,L,D}, \|\cdot\|_{\bar{L}^\infty(B_R)})$. For any $t > 0$, with probability at least $1 - e^{-t}$,*

$$\Delta_{M,n} \le 4R\epsilon + 8L(\Xi_{R,n} + \Xi_R) + \sqrt{\frac{2(H_\epsilon + t)}{\tilde{n}}}\left(\left\|\bar{g}^{(i)}_{M,n}\right\|_{L^2(\nu^{(i)})} + \left(4R^2\epsilon^2 + 64L^2R^2\Xi_R\right)^{1/2}\right) + \frac{2M(H_\epsilon + t)}{3\tilde{n}},$$

where $\Xi_{R,n} = \sum_{i=1}^K \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \|X_j^{(i)}\|^2 \mathbb{1}\{X_j^{(i)} \notin B_R\}$.

The remainder $\Delta_{M^c,n}$ is controlled in expectation through the tail quantity Ξ_R in Eq. (4).

Lemma 2. *Let $\Delta_{M^c,n}$ and Ξ_R be as defined in Eq. (9) and Eq. (4), respectively, with $M = 4LR^2$. Then, we have*

$$\mathbb{E}_{n\nu}[\Delta_{M^c,n}] \le \frac{8L}{3}\Xi_R.$$

The proofs of Lem. 1 and Lem. 2 are provided in Appendix F and Appendix G, respectively.

Combining these two lemmas with Eq. (7) yields Thm. 4, whose proof we now sketch.

Proof sketch of Thm. 4. By Eq. (7), the potential error is at most $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D}) + \Delta_{M,n} + \Delta_{M^c,n}$. The truncated term $\Delta_{M,n}$ is controlled by Lem. 1. The key observation is that its variance factor is again bounded by the potential error: the Poincaré inequality bounds the variance by the L^2 distance between transport maps, which Thm. 3 converts back into the potential error. The remaining term $\Delta_{M^c,n}$ is bounded in expectation by Lem. 2. Taking expectations yields a self-bounding inequality for $\mathbb{E}_{n\nu^{(\cdot)}}[\Phi(T_n^{(\cdot)}) - \Phi_\nu]$. The stated conditions ensure that all additional terms are at most a $\gamma_{\tilde{n}}$-fraction of $A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$, and solving the inequality gives the claim. □

The full proof is given in Appendix H.

6 Conclusion

We established, to our knowledge, the first convergence rate for estimating optimal transport maps to the barycenter in multidimensional Euclidean spaces. The rate interpolates between the fast and slow regimes previously known only for two-distribution transport map estimation, and it depends on the underlying maps only through their approximability by the estimator's class, which may be non-convexly parametrized, for example by deep neural networks. Our guarantee is an oracle upper bound that assumes exact optimization. Establishing matching lower bounds and developing adaptive, computationally tractable estimators are left for future work.

AI use statement

In this work, we used generative AI tools to assist in the writing of proofs. We did not use generative AI tools to develop theoretical models or conceptual frameworks, to formulate mathematical claims, to provide critical ingredients for proving mathematical claims, or to assist with translation. Other uses, such as proposing or refining hypotheses, interpreting results, generating synthetic data sets, designing or providing feedback on research methodology or experiments, implementing methods, cleaning and reformatting data sets, and supporting qualitative and thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools to draft parts of the paper and to edit the paper, including the statements of mathematical claims, to improve readability. We have reviewed all AI-assisted work. All claims and proofs were first formulated and derived by the author, and generative AI tools were then used in collaboration with the author only to refine their presentation. After this refinement, the author revised all AI-drafted or AI-edited passages to ensure that they accurately reflect the technical content and do not introduce unsupported claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, January 2011. doi: 10.1137/100805741. URL https://epubs.siam.org/doi/10.1137/100805741.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1233–1243. PMLR, June 2020. URL https://proceedings.mlr.press/v108/chen20d.html.

Evgenii Chzhen and Nicolas Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, August 2022. doi: 10.1214/22-AOS2198.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331. Curran Associates, Inc., 2020.

NABARUN DEB, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29736–29753. Curran Associates, Inc., 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/f862d13454fd267baa5fedfffb200567-Abstract.html.

E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and A. Mayo-Íscar. Robust clustering tools based on optimal transportation. *Statistics and Computing*, 29(1):139–160, January 2019. doi: 10.1007/s11222-018-9800-z. URL https://doi.org/10.1007/s11222-018-9800-z.

Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *The Annals of Statistics*, 53(3):963–988, June 2025. doi: 10.1214/24-AOS2482. URL https://projecteuclid.org/journals/annals-of-statistics/volume-53/issue-3/Optimal-transport-map-estimation-in-general-function-spaces/10.1214/24-AOS2482.full.

Alireza F. Pour and Hassan Ashtiani. On the role of noise in the sample complexity of learning recurrent neural networks: Exponential gaps for long sequences. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80257–80269. Curran Associates, Inc., 2023. doi: 10.52202/075280-3517. URL https://papers.nips.cc/paper_files/paper/2023/hash/fe03053bd2cf5b5c56de1e463bc53e1a-Abstract-Conference.html.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable computations of wasserstein barycenter via input convex neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1571–1581. PMLR, July 2021. URL https://proceedings.mlr.press/v139/fan21d.html.

Kazuto Fukuchi. Meta optimality for demographic parity constrained regression via post-processing. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 18024–18046. PMLR, October 2025. URL https://proceedings.mlr.press/v267/fukuchi25a.html.

Florian F. Gunsilius. ON THE CONVERGENCE RATE OF POTENTIALS OF BRENIER MAPS. *Econometric Theory*, 38(2):381–417, April 2022. doi: 10.1017/S0266466621000037. URL https://www.cambridge.org/core/journals/econometric-theory/article/on-the-convergence-rate-of-potentials-of-brenier-maps/FF998CC16CF76AE13674854A37E3ADB4.

Alex Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. *Advances in Neural Information Processing Systems*, 37:42162–42210, December 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4a68bb0e075a3215d4ff37382ea598dc-Abstract-Conference.html.

Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1501–1509. PMLR, July 2017. URL https://proceedings.mlr.press/v70/ho17a.html.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, April 2021. doi: 10.1214/20-AOS1997. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-2/Minimax-estimation-of-smooth-optimal-transport-maps/10.1214/20-AOS1997.full.

Kaheon Kim, Rentian Yao, Changbo Zhu, and Xiaohui Chen. Optimal transport barycenter via nonconvex-concave minimax optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 30879–30899. PMLR, October 2025. URL https://proceedings.mlr.press/v267/kim25al.html.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3tFAs5E-Pe.

Alexander Korotin, Vage Egiazarian, Lingxiao Li, and Evgeny Burnaev. Wasserstein iterative networks for barycenter estimation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15672–15686, December 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/6489f2c6ac6420124fcef2a489615a97-Abstract-Conference.html.

Julien Lacombe, Julie Digne, Nicolas Courty, and Nicolas Bonneel. Learning to generate wasserstein barycenters. *Journal of Mathematical Imaging and Vision*, 65(2):354–370, April 2023. doi: 10.1007/s10851-022-01121-y. URL https://doi.org/10.1007/s10851-022-01121-y.

Shubo Li, Yizhe Ding, Lingzhou Xue, and Runze Li. Stability and oracle inequalities for optimal transport maps between general distributions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025. URL https://openreview.net/forum?id=aAW7ibNYOT.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, June 2024. doi: 10.1214/24-AOS2379. URL https://projecteuclid.org/journals/annals-of-statistics/volume-52/issue-3/Plugin-estimation-of-smooth-optimal-transport-maps/10.1214/24-AOS2379.full.

Boris Muzellec, Adrien Vacher, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. arXiv preprint arXiv:2112.01907 [stat], December 2021. URL http://arxiv.org/abs/2112.01907.

Weigutian Ou and Helmut Bölcskei. Covering numbers for deep ReLU networks with applications to function approximation and nonparametric regression. arXiv preprint arXiv:2410.06378, March 2026. URL http://arxiv.org/abs/2410.06378.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. arXiv preprint arXiv:2109.12004 [math], May 2024. URL http://arxiv.org/abs/2109.12004.

Aram-Alexandre Pooladian, Vincent Divol, and Jonathan Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28128–28150. PMLR, July 2023. URL https://proceedings.mlr.press/v202/pooladian23b.html.

Léo Portales, Edouard Pauwels, and Elsa Cazelles. Sample complexity of optimal transport barycenters with discrete support. arXiv preprint arXiv:2505.21274 [math], November 2025. URL http://arxiv.org/abs/2505.21274.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In Alfred M. Bruckstein, Bart M. ter Haar Romeny, Alexander M. Bronstein, and Michael M. Bronstein (eds.), *Scale Space and Variational Methods in Computer Vision*, pp. 435–446, Berlin, Heidelberg, 2012. Springer. ISBN 978-3-642-24785-9. doi: 10.1007/978-3-642-24785-9_37.

Philippe Rigollet and Austin J. Stromme. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 53(1), February 2025. doi: 10.1214/24-AOS2455. URL https://projecteuclid.org/journals/annals-of-statistics/volume-53/issue-1/On-the-sample-complexity-of-entropic-optimal-transport/10.1214/24-AOS2455.full.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, August 2020. doi: 10.1214/19-AOS1875. URL https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-4/Nonparametric-regression-using-deep-neural-networks-with-ReLU-activation-function/10.1214/19-AOS1875.full.

Dror Simon and Aviad Aberdam. Barycenters of natural images - constrained wasserstein barycenters for image morphing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7907–7916. IEEE Computer Society, June 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00793. URL https://www.computer.org/csdl/proceedings-article/cvpr/2020/716800h907/1m3o6YLlYAg.

Justin Solomon, Fernando de Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015. doi: 10.1145/2766963. URL https://doi.org/10.1145/2766963.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 33416–33447. PMLR, July 2023. URL https://proceedings.mlr.press/v202/takakura23a.html.

Adrien Vacher and François-Xavier Vialard. Semi-dual unbalanced quadratic optimal transport: fast statistical rates and convergent algorithm. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 34734–34758. PMLR, July 2023. URL https://proceedings.mlr.press/v202/vacher23a.html.

Cédric Villani. *Optimal Transport*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9. URL http://link.springer.com/10.1007/978-3-540-71050-9.

Gabriele Visentin and Patrick Cheridito. Computing optimal transport maps and wasserstein barycenters using conditional normalizing flows. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 61564–61581. PMLR, October 2025. URL https://proceedings.mlr.press/v267/visentin25a.html.

A Notation

For $\Omega \subseteq \mathbb{R}^p$, let $\mathcal{P}(\Omega)$ denote the set of all Radon probability measures on Ω. The set of all probability measures with finite second moments is defined as

$$\mathcal{P}_2(\Omega) = \left\{ \nu \in \mathcal{P}(\Omega) : \int \|x\|^2 \nu(dx) < \infty \right\}.$$

For a probability measure ν on $(\Omega, \mathcal{F})$, we denote the expectation and the variance under ν by $\mathbb{E}_\nu$ and $\mathbb{V}_\nu$, respectively. We denote the probability over the samples $X^{(\cdot)}$ by $\mathbb{P}_{n\nu}$, and the corresponding expectation by $\mathbb{E}_{n\nu}$.

For a topological space Ω, let $C(\Omega)$ denote the set of all continuous functions on Ω, and let $C_b(\Omega) \subseteq C(\Omega)$ denote the set of all bounded continuous functions on Ω. For $\Omega \subseteq \mathbb{R}^p$, the set of all continuous functions with at most quadratic growth is defined as

$$C_b^2(\Omega) = \left\{ f \in C(\Omega) : \sup_{x \in \Omega} \frac{|f(x)|}{1 + \|x\|^2} < \infty \right\}.$$

For a probability measure ν on $(\Omega, \mathcal{F})$ and $q \in [1, \infty)$, let $L^q(\nu)$ denote the space of measurable functions $f : \Omega \to \mathbb{R}$ with $\|f\|_{L^q(\nu)} = (\int |f|^q d\nu)^{1/q} < \infty$. Let $L^\infty(\nu)$ denote the space of measurable functions $f : \Omega \to \mathbb{R}$ whose ν-essential supremum $\|f\|_{L^\infty(\nu)} = \inf\{M \in \mathbb{R} : \int \mathbb{1}\{|f| > M\} d\nu = 0\}$ is finite. For a set $\Omega' \subseteq \Omega$ and a function $f : \Omega \to \mathbb{R}$, let $\|f\|_{L^\infty(\Omega')} = \sup_{x \in \Omega'} |f(x)|$. For a vector-valued function $f : \Omega \to \mathbb{R}^p$, we define $\|\|f\|\|_{L^q(\nu)}$ and $\|\|f\|\|_{L^\infty(\Omega')}$ as the $L_q(\nu)$-norm and $L^\infty(\Omega')$-norm of the function $x \to \|f(x)\|$. For a function $f : \mathbb{R}^p \to \mathbb{R}$, its convex conjugate f^* is defined, for $y \in \mathbb{R}^p$, by

$$f^*(y) = \sup\{\langle x, y \rangle - f(x) : x \in \mathbb{R}^p\}.$$

For a set A equipped with a norm $\|\cdot\|$ and $\epsilon > 0$, the ϵ-covering number $N(\epsilon, A, \|\cdot\|)$ is the smallest integer N for which there exist $a_1, \ldots, a_N \in A$ such that for every $a \in A$, $\|a - a_j\| \leq \epsilon$ holds for some $j \in [N]$. Such a set $\{a_1, \ldots, a_N\}$ is called an ϵ-net of A.

B Additional Preliminaries: Kantorovich Problem and Its Barycenter

Given two probability measures ν and ν' on a measurable space $(\Omega, \mathcal{F})$ and a cost function $c : \Omega \times \Omega \to \mathbb{R}_+$, the Kantorovich formulation of the optimal transport cost between ν and ν' is

$$\mathbb{K}_c(\nu, \nu') = \inf\left\{ \int c d\pi : \pi \in \Pi(\nu, \nu') \right\}, \tag{10}$$

where $\Pi(\nu, \nu')$ is the set of couplings of ν and ν', that is, the set of all probability measures π on $\Omega \times \Omega$ such that $\pi(\cdot \times \Omega) = \nu(\cdot)$ and $\pi(\Omega \times \cdot) = \nu'(\cdot)$. A coupling $\pi \in \Pi(\nu, \nu')$ attaining the infimum in Eq. (10) is called an optimal transport plan between ν and ν'. Under mild regularity conditions, an optimal transport plan exists.

The Kantorovich problem admits a dual formulation:

Theorem 5 (Villani (2009))**.** *Given two Borel probability measures ν and ν' on a Polish space Ω and a cost function $c : \Omega \times \Omega \to \mathbb{R}_+$, the following duality holds:*

$$\mathbb{K}_c(\nu, \nu') = \sup\left\{ \int f^c d\nu + \int f d\nu' : f \in C_{b,c}(\Omega) \right\},$$

where $f^c(x) = \inf\{c(x, y) - f(y) : y \in \Omega\}$ and $C_{b,c}(\Omega) \subseteq C_b(\Omega)$ is the set of bounded continuous functions f for which there exists a function $g : \Omega \to \mathbb{R}$ such that $f(y) = \inf\{c(x, y) - g(x) : x \in \Omega\}$ for all $y \in \Omega$.

Given a K-tuple of probability measures $\nu^{(\cdot)} = (\nu^{(1)}, ..., \nu^{(K)})$ on $(\Omega, \mathcal{F})$, the barycenter of $\nu^{(\cdot)}$ with respect to the optimal transport cost $\mathbb{K}_c$ is a probability measure that attains the following infimum:

$$\inf\left\{ \sum_{i=1}^K \lambda^{(i)} \mathbb{K}_c\Big(\nu^{(i)}, \nu\Big) : \nu \in \mathcal{P}(\Omega) \right\}, \tag{11}$$

where $\mathcal{P}(\Omega)$ is the set of all Borel probability measures on Ω.

C Proof of Characterization of Transport Maps to the Barycenter

Thm. 1 follows from the results of Agueh & Carlier (2011), but its statement differs from the original. We therefore derive it from these results. Throughout this section, we assume $\Omega = \mathbb{R}^p$ and $c(x, y) = \|x - y\|^2/2$.

We first state the results of Agueh & Carlier (2011) that we use.

Definition 3. *A probability measure $\nu \in \mathcal{P}_2(\Omega)$ vanishes on small sets if $\nu(A) = 0$ for every Borel set $A \subseteq \Omega$ with Hausdorff dimension at most $p - 1$.*

Theorem 6 (Duality of the barycenter problem by Agueh & Carlier (2011))**.** *Given a K-tuple of probability measures $\nu^{(\cdot)} = (\nu^{(1)}, ..., \nu^{(K)}) \in \mathcal{P}_2(\Omega)^K$, the following duality holds:*

$$\inf\left\{\sum_{i=1}^K \lambda^{(i)} \mathbb{K}_c(\nu^{(i)}, \nu) : \nu \in \mathcal{P}_2(\Omega)\right\}$$
$$= \sup\left\{\sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} : f^{(\cdot)} \in C_b^2(\Omega)^K, \sum_{i=1}^K \lambda^{(i)} f^{(i)} = 0\right\}, \tag{12}$$

where, for a function $f : \Omega \to \mathbb{R}$, the function $f^\dagger$ is defined by

$$f^\dagger(y) = \inf\left\{\frac{1}{2}\|y - x\|^2 - f(x) : x \in \Omega\right\} \text{ for } y \in \Omega.$$

Theorem 7 (Characterization of the barycenter by Agueh & Carlier (2011))**.** *Given a K-tuple of probability measures $\nu^{(\cdot)} = (\nu^{(1)}, ..., \nu^{(K)}) \in \mathcal{P}_2(\Omega)^K$, assume that there is $i \in [K]$ such that $\nu^{(i)}$ vanishes on small sets. Then, Eq. (11) admits the solution $\nu = \nabla\psi^{(i)} \sharp \nu^{(i)}$, where $\psi^{(i)} : \Omega \to \mathbb{R}$ is the convex function defined by*

$$\psi^{(i)}(x) = \frac{1}{2}\|x\|^2 - \left(f^{(i)}\right)^\dagger(x) \text{ for } x \in \Omega,$$

where $f^{(\cdot)}$ is a solution of Eq. (12).

Under the quadratic cost, the transform $f^\dagger$ coincides with f^c in Thm. 5.

Proof of Thm. 1. Let $\phi^{(\cdot)}$ be a solution of Eq. (2). We first show that the tuple $f^{(\cdot)}$ defined by

$$f^{(i)}(x) = \frac{1}{2}\|x\|^2 + C - \phi^{(i)}(x)$$

solves Eq. (12). We begin by verifying that $f^{(\cdot)}$ satisfies the constraint in Eq. (12). By the constraint on $\phi^{(\cdot)}$ in Eq. (2), we have

$$\sum_{i=1}^K \lambda^{(i)} f^{(i)}(x) = \frac{1}{2}\|x\|^2 + C - \sum_{i=1}^K \lambda^{(i)} \phi^{(i)}(x) = 0.$$

Since $\phi^{(i)} \in C_b^2(\Omega)$, it is continuous and there is $\kappa_i < \infty$ such that $|\phi^{(i)}(x)| \leq \kappa_i(1 + \|x\|^2)$ for all $x \in \Omega$. As a sum of continuous functions, $f^{(i)}$ is continuous, and

$$\left|f^{(i)}(x)\right| \leq \frac{1}{2}\|x\|^2 + |C| + \left|\phi^{(i)}(x)\right| \leq \left(\frac{1}{2} + |C| + \kappa_i\right)\left(1 + \|x\|^2\right) \text{ for all } x \in \Omega,$$

so $f^{(i)} \in C_b^2(\Omega)$.

Next, assume that there exists a solution $\bar{f}^{(\cdot)}$ of Eq. (12) such that the function $\bar{\phi}^{(i)}(x) = \frac{1}{2}\|x\|^2 + C - \bar{f}^{(i)}(x)$ satisfies the constraint in Eq. (2). Under this assumption, which we

verify at the end of the proof, we show that $f^{(\cdot)}$ attains the supremum in Eq. (12). For $f(x) = \frac{1}{2}\|x\|^2 + C - \phi(x)$, where ϕ is a convex function, we have

$$\begin{aligned} f^\dagger(y) =& \inf\left\{\frac{1}{2}\|y-x\|^2 - \frac{1}{2}\|x\|^2 - C + \phi(x) : x \in \Omega\right\} \\ =& \inf\left\{\frac{1}{2}\|y\|^2 - C - \langle y, x\rangle + \phi(x) : x \in \Omega\right\} \\ =& \frac{1}{2}\|y\|^2 - C - \sup\{\langle y, x\rangle - \phi(x) : x \in \Omega\} \\ =& \frac{1}{2}\|y\|^2 - C - \phi^*(y). \end{aligned}$$

Hence,

$$\begin{aligned} \sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} =& -\sum_{i=1}^K \lambda^{(i)} \int \left(\phi^{(i)}\right)^* d\nu^{(i)} + S, \\ \sum_{i=1}^K \lambda^{(i)} \int \left(\bar{f}^{(i)}\right)^\dagger d\nu^{(i)} =& -\sum_{i=1}^K \lambda^{(i)} \int \left(\bar{\phi}^{(i)}\right)^* d\nu^{(i)} + S, \end{aligned}$$

where $S = \sum_{i=1}^K \lambda^{(i)} \int \frac{1}{2}\|x\|^2 d\nu^{(i)} - C$. By the optimality of $\phi^{(\cdot)}$ in Eq. (2), we have

$$\begin{aligned} \sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} =& -\sum_{i=1}^K \lambda^{(i)} \int \left(\phi^{(i)}\right)^* d\nu^{(i)} + S \\ \geq& -\sum_{i=1}^K \lambda^{(i)} \int \left(\bar{\phi}^{(i)}\right)^* d\nu^{(i)} + S \\ =& \sum_{i=1}^K \lambda^{(i)} \int \left(\bar{f}^{(i)}\right)^\dagger d\nu^{(i)}. \end{aligned}$$

Conversely, by the optimality of $\bar{f}^{(\cdot)}$, we have

$$\sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} \leq \sum_{i=1}^K \lambda^{(i)} \int \left(\bar{f}^{(i)}\right)^\dagger d\nu^{(i)}.$$

Therefore,

$$\sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} = \sum_{i=1}^K \lambda^{(i)} \int \left(\bar{f}^{(i)}\right)^\dagger d\nu^{(i)}. \tag{13}$$

Hence, by Eq. (13), $f^{(\cdot)}$ attains the supremum in Eq. (12) and is thus a solution.

Since $f^{(\cdot)}$ is a solution of Eq. (12), the function $\psi^{(\cdot)}$ defined in Thm. 7 satisfies, by construction, $\psi^{(i)} = (\phi^{(i)})^* + C$. By Thm. 7, $\nabla\psi^{(i)}$ is the optimal transport map from $\nu^{(i)}$ to the barycenter $\bar{\nu}^{(\cdot)}$. Since $\nabla\psi^{(i)} = \nabla(\phi^{(i)})^*$, we conclude that $\nabla(\phi^{(i)})^*$ is the optimal transport map from $\nu^{(i)}$ to the barycenter $\bar{\nu}^{(\cdot)}$.

It remains to verify the assumption that there exists a solution $\bar{f}^{(\cdot)}$ of Eq. (12) such that $\bar{\phi}^{(i)}(x) = \frac{1}{2}\|x\|^2 + C - \bar{f}^{(i)}(x)$ satisfies the constraint in Eq. (2). We accomplish this by confirming that the equality in Eq. (12) of Thm. 6 continues to hold if $f^{(\cdot)} \in C_b^2(\Omega)^K$ is replaced by $f^{(\cdot)} \in C_b^{2\dagger}(\Omega)^K$, where $C_b^{2\dagger}(\Omega)$ denotes the subset of $C_b^2(\Omega)$ consisting of functions f for which there exists $g : \Omega \to \mathbb{R}$ such that $f(y) = \inf\{\frac{1}{2}\|y-x\|^2 - g(x) : x \in \Omega\}$ for all $y \in \Omega$. Granting this, let g satisfy $\bar{f}^{(i)} = g^\dagger$. We can then rewrite $\frac{1}{2}\|x\|^2 + C - \bar{f}^{(i)}(x)$ as

$$\frac{1}{2}\|x\|^2 + C - \bar{f}^{(i)}(x) = \frac{1}{2}\|x\|^2 + C - \inf\left\{\frac{1}{2}\|x-y\|^2 - g(y) : y \in \Omega\right\}$$

$$= \sup\left\{ \langle x, y\rangle - \frac{1}{2}\|y\|^2 - C + g(y) : y \in \Omega \right\},$$

which is the convex conjugate of $y \mapsto \frac{1}{2}\|y\|^2 + C - g(y)$. Since the convex conjugate of a function is convex and lower semi-continuous, the assumption is verified.

For $\gamma \in C_b^2(\Omega)$, define

$$b(\gamma) := \inf\left\{ -\sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} : f^{(\cdot)} \in C_b^{2\dagger}(\Omega)^K,\ \sum_{i=1}^K \lambda^{(i)} f^{(i)} = \gamma \right\},$$

and $J(\nu) := \sum_{i=1}^K \lambda^{(i)} \Phi_i(\nu)$ for $\nu \in \mathcal{P}_2(\Omega)$, where $\Phi_i(\nu) := \mathbb{K}_c(\nu^{(i)}, \nu)$. From Thm. 5, we have

$$\Phi_i(\nu) = \sup\left\{ \int f^\dagger d\nu^{(i)} + \int f d\nu : f \in C_b^{2\dagger}(\Omega) \right\}. \tag{14}$$

By construction and Eq. (14),

$$J(\nu) = \sup\left\{ -b(\gamma) + \int \gamma d\nu : \gamma \in C_b^2(\Omega) \right\},$$

that is, J is the convex conjugate of b with respect to the pairing $\langle \nu, f\rangle = \int f d\nu$. Each Φ_i is a supremum of affine functions and is thus convex and lower semi-continuous. We also have $\Phi_i(\nu) \geq 0$, since the quadratic cost takes only nonnegative values, and $\Phi_i(\nu) < \infty$, since $\nu^{(i)}, \nu \in \mathcal{P}_2(\Omega)$. Hence, J is a proper convex function. By the standard conjugate-subdifferential relation, $\gamma \in \partial J(\nu)$ if and only if $\nu \in \partial b(\gamma)$. Since J is convex and is minimized at $\bar{\nu}^{(\cdot)}$ over $\mathcal{P}_2(\Omega)$, we have $0 \in \partial J(\bar{\nu}^{(\cdot)})$. By the Fenchel–Young equality, we have

$$J(\bar{\nu}^{(\cdot)}) = -b(0) = \sup\left\{ \sum_{i=1}^K \lambda^{(i)} \int \left(f^{(i)}\right)^\dagger d\nu^{(i)} : f^{(\cdot)} \in C_b^{2\dagger}(\Omega)^K,\ \sum_{i=1}^K \lambda^{(i)} f^{(i)} = 0 \right\}.$$

This verifies that the equality of Eq. (12) in Thm. 6 holds even if we replace $f^{(\cdot)} \in C_b^2(\Omega)^K$ with $f^{(\cdot)} \in C_b^{2\dagger}(\Omega)^K$. □

D Properties of the Potential Functional

In this section, we establish several properties of the potential functional ϕ in Eq. (3). The first is a closed-form expression for the convex conjugate $\phi^*[T]$.

Lemma 3. *Let $T : \Omega \to \Omega$ be a continuous bijective gradient map that is monotone as in Def. 2 (c), i.e., $\langle T(x) - T(y), x - y\rangle \geq 0$ for any $x, y \in \Omega$. Then, $\nabla\phi[T] = T^{-1}$ and $\nabla\phi^*[T] = T$, and for any $x \in \Omega$,*

$$\phi^*[T](x) = \int_0^1 \langle T(T^{-1}(0) + t(x - T^{-1}(0))), x - T^{-1}(0)\rangle dt.$$

The next lemma gives an envelope of $\phi^*[T]$ that follows from the monotonicity and Lipschitzness of T.

Lemma 4. *Let $T : \Omega \to \Omega$ be a continuous bijective gradient map that is monotone and L-Lipschitz as in Def. 2 (c): for any $x, y \in \Omega$,*

$$0 \leq \langle T(x) - T(y), x - y\rangle \leq L\|x - y\|^2.$$

Then, for any $x \in \Omega$,

$$0 \leq \phi^*[T](x) \leq \frac{L}{2}\left\|x - T^{-1}(0)\right\|^2.$$

We now prove Lem. 3 and Lem. 4.

Proof of Lem. 3. Since T is a gradient map, there exists a differentiable function $\psi : \Omega \to \mathbb{R}$ such that $T = \nabla\psi$. The monotonicity of T implies that ψ is convex. Since T is bijective, the subdifferential of ψ^* at each $y \in \Omega$ is $\{T^{-1}(y)\}$, and thus $\nabla\psi^* = T^{-1}$. The continuity and bijectivity of T imply that T^{-1} is continuous. Hence, $\phi[T] = \psi^* - \psi^*(0)$ by the gradient theorem, and therefore $\nabla\phi[T] = T^{-1}$. Since ψ is convex and continuous, $\phi^*[T] = \psi^{**} + \psi^*(0) = \psi + \psi^*(0)$, which implies $\nabla\phi^*[T] = \nabla\psi = T$. Since $\nabla\phi[T](0) = T^{-1}(0)$, the Fenchel–Young equality gives $\phi^*[T](T^{-1}(0)) = -\phi[T](0) = 0$. Applying the gradient theorem along the path from $T^{-1}(0)$ to x yields the integral expression for $\phi^*[T](x)$. □

Proof of Lem. 4. Let $x_0 = T^{-1}(0)$. By Lem. 3, we have

$$\phi^*[T](x) = \int_0^1 \langle T(x_0 + t(x - x_0)) - T(x_0), x - x_0\rangle dt.$$

Fix $t \in (0, 1)$. Applying the monotonicity and Lipschitzness of T, we have

$$0 \le \langle T(x_0 + t(x - x_0)) - T(x_0), t(x - x_0)\rangle \le Lt^2\|x - x_0\|^2.$$

Dividing both sides by t and integrating over $t \in (0, 1)$ gives, for any $x \in \Omega$,

$$0 \le \phi^*[T](x) \le \int_0^1 Lt\|x - x_0\|^2 dt = \frac{L}{2}\left\|x - T^{-1}(0)\right\|^2,$$

which is the desired bound. □

E Stability Analysis

Proof of Thm. 3. The L-Lipschitzness of $T^{(i)}$ implies that $\phi^*[T^{(i)}]$ is L-smooth, and thus $\phi[T^{(i)}]$ is $1/L$-strongly convex. Therefore, for any $y, z \in \Omega$,

$$\phi\left[T^{(i)}\right](y) - \phi\left[T^{(i)}\right](z) - \left\langle \left(T^{(i)}\right)^{-1}(z), y - z\right\rangle \ge \frac{1}{2L}\|y - z\|^2.$$

Given $x \in \Omega$, taking $y = T_\nu^{(i)}(x)$ and $z = T^{(i)}(x)$ gives

$$\phi\left[T^{(i)}\right]\left(T_\nu^{(i)}(x)\right) - \phi\left[T^{(i)}\right]\left(T^{(i)}(x)\right) - \left\langle x, T_\nu^{(i)}(x) - T^{(i)}(x)\right\rangle \ge \frac{1}{2L}\left\|T_\nu^{(i)}(x) - T^{(i)}(x)\right\|^2.$$

By the Fenchel–Young equality, we have

$$\phi\left[T^{(i)}\right]\left(T_\nu^{(i)}(x)\right) + \phi^*\left[T^{(i)}\right](x) - \left\langle x, T_\nu^{(i)}(x)\right\rangle \ge \frac{1}{2L}\left\|T_\nu^{(i)}(x) - T^{(i)}(x)\right\|^2. \tag{15}$$

Since ϕ_ν is real-valued and convex, and hence lower semi-continuous, the Fenchel–Young equality gives, for any $x \in \Omega$,

$$\phi_\nu^*(x) + \phi_\nu\left(T_\nu^{(i)}(x)\right) - \left\langle x, T_\nu^{(i)}(x)\right\rangle = 0, \tag{16}$$

where we use the fact that $\nabla\phi_\nu^* = T_\nu^{(i)}$. Subtracting Eq. (16) from Eq. (15) gives

$$\phi^*\left[T^{(i)}\right](x) - \phi_\nu^*(x) + \phi\left[T^{(i)}\right]\left(T_\nu^{(i)}(x)\right) - \phi_\nu\left(T_\nu^{(i)}(x)\right) \ge \frac{1}{2L}\left\|T_\nu^{(i)}(x) - T^{(i)}(x)\right\|^2.$$

Integrating both sides with respect to $\nu^{(i)}$ gives

$$\int\left(\phi^*\left[T^{(i)}\right] - \phi_\nu^*\right)d\nu^{(i)} + \int\left(\phi\left[T^{(i)}\right] \circ T_\nu^{(i)} - \phi_\nu \circ T_\nu^{(i)}\right)d\nu^{(i)} \ge \frac{1}{2L}\left\|\left\|T_\nu^{(i)} - T^{(i)}\right\|\right\|^2_{L^2(\nu^{(i)})}.$$

Changing variables via $\bar\nu = T_\nu^{(i)}\sharp\nu^{(i)}$ and summing over $i \in [K]$ with weights $\lambda^{(i)}$, we obtain

$$\Phi\left(T^{(\cdot)}\right) - \Phi_\nu + \int\sum_{i=1}^K \lambda^{(i)}\left(\phi\left[T^{(i)}\right] - \phi_\nu\right)d\bar\nu \ge \frac{1}{2L}\left\|\left\|T_\nu^{(\cdot)} - T^{(\cdot)}\right\|\right\|^2_{L^2(\nu^{(\cdot)})}.$$

The claim is verified if

$$\int \sum_{i=1}^{K} \lambda^{(i)} \Big(\phi\Big[T^{(i)}\Big] - \phi_\nu \Big) d\bar{\nu} = 0. \tag{17}$$

From Lem. 3 and the congruency of $T^{(\cdot)}$ in the sense of Def. 2 (a), we have, for any $x \in \Omega$,

$$\sum_{i=1}^{K} \lambda^{(i)} \nabla\phi\Big[T^{(i)}\Big](x) = x.$$

Taking the inner product with x and integrating both sides over the line segment from 0 to x gives

$$\sum_{i=1}^{K} \lambda^{(i)} \phi\Big[T^{(i)}\Big](x) = \frac{1}{2}\|x\|^2.$$

Also, since ϕ_ν satisfies the congruency condition of Thm. 1 with $C = 0$, we have for $\bar{\nu}$-almost every $x \in \Omega$,

$$\sum_{i=1}^{K} \lambda^{(i)} \Big(\phi\Big[T^{(i)}\Big] - \phi_\nu \Big)(x) = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|x\|^2 = 0.$$

Eq. (17) follows, and the proof is complete. □

F Analysis of the Clipped Process

In this section, we prove Lem. 1. We first derive a tail bound for a centered, bounded multi-sample empirical average, which corresponds to $\Delta_{M,n}$ with $T_n^{(\cdot)}$ replaced by fixed maps. We then extend it to $\Delta_{M,n}$ by an ϵ-net argument.

F.1 Bound on Multi-Sample Empirical Averages

For each $i \in [K]$, let $g^{(i)} : \Omega \to \mathbb{R}$ be a measurable function, and let $\bar{g}^{(i)} = g^{(i)} - \mathbb{E}_{\nu^{(i)}}[g^{(i)}]$ be its centered version. Consider the multi-sample empirical average

$$G_n = \sum_{i=1}^{K} \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \bar{g}^{(i)}\Big(X_j^{(i)}\Big).$$

The following lemma gives a tail bound on G_n.

Lemma 5. *Define the $L^2(\nu^{(\cdot)})$ and $L^\infty(\nu^{(\cdot)})$ norms of $\bar{g}^{(\cdot)}$ by*

$$\Big\|\bar{g}^{(\cdot)}\Big\|^2_{L^2(\nu^{(\cdot)})} = \sum_{i=1}^{K} \lambda^{(i)} \Big\|\bar{g}^{(i)}\Big\|^2_{L^2(\nu^{(i)})}, \text{ and } \Big\|\bar{g}^{(\cdot)}\Big\|_{L^\infty(\nu^{(\cdot)})} = \max_{i\in[K]} \Big\|\bar{g}^{(i)}\Big\|_{L^\infty(\nu^{(i)})}.$$

Assume that $\|\bar{g}^{(\cdot)}\|_{L^2(\nu^{(\cdot)})}$ and $\|\bar{g}^{(\cdot)}\|_{L^\infty(\nu^{(\cdot)})}$ are finite. Then, for any $t > 0$, with probability at least $1 - e^{-t}$,

$$G_n \le \sqrt{\frac{2t}{\tilde{n}}} \Big\|\bar{g}^{(\cdot)}\Big\|_{L^2(\nu^{(\cdot)})} + \frac{t}{3\tilde{n}} \Big\|\bar{g}^{(\cdot)}\Big\|_{L^\infty(\nu^{(\cdot)})}.$$

The proof of Lem. 5 relies on Bernstein's inequality for the cumulant generating function. The cumulant generating function of a random variable Z is defined as $\psi_Z(\tau) = \ln \mathbb{E}_\nu[\exp(\tau Z)]$.

Lemma 6 (Bernstein's inequality for the cumulant generating function)**.** *Let X be a centered random variable over a probability space $(\Omega, \mathcal{F}, \nu)$. Then, for any $|\tau| \le 3/\|X\|_{L^\infty(\nu)}$, we have*

$$\psi_X(\tau) \le \frac{\mathbb{V}_\nu[X]\tau^2}{2(1 - \|X\|_{L^\infty(\nu)}|\tau|/3)}.$$

We now prove Lem. 5.

Proof of Lem. 5. We first bound the cumulant generating function of G_n. Since the samples are independent and identically distributed within each group, we have

$$\begin{aligned}\psi_{G_n}(\tau) &= \ln \mathbb{E}_{n\nu}[\exp(\tau G_n)] \\ &= \sum_{i=1}^{K} n_i \ln \mathbb{E}_{\nu^{(i)}}\left[\exp\left(\frac{\tau\lambda^{(i)}}{n_i}\bar{g}^{(i)}\left(X_1^{(i)}\right)\right)\right] \\ &= \sum_{i=1}^{K} n_i \psi_{\bar{g}^{(i)}(X_1^{(i)})}\left(\frac{\tau\lambda^{(i)}}{n_i}\right). \end{aligned} \tag{18}$$

By Lem. 6 and $\mathbb{V}_{\nu^{(i)}}[\bar{g}^{(i)}] = \|\bar{g}^{(i)}\|^2_{L^2(\nu^{(i)})}$, we have

$$\psi_{\bar{g}^{(i)}}\left(\frac{\tau\lambda^{(i)}}{n_i}\right) \le \frac{\|\bar{g}^{(i)}\|^2_{L^2(\nu^{(i)})}\left(\tau\lambda^{(i)}/n_i\right)^2}{2\left(1 - \|\bar{g}^{(i)}\|_{L^\infty(\nu^{(i)})}|\tau|\lambda^{(i)}/3n_i\right)}, \tag{19}$$

provided that $|\tau| < 3n_i/(\lambda^{(i)}\|\bar{g}^{(i)}\|_{L^\infty(\nu^{(i)})})$. Substituting Eq. (19) into Eq. (18) gives

$$\begin{aligned}\psi_{G_n}(\tau) &\le \sum_{i=1}^{K} \frac{\lambda^{(i)}\|\bar{g}^{(i)}\|^2_{L^2(\nu^{(i)})}\left(\lambda^{(i)}/n_i\right)\tau^2}{2\left(1 - \|\bar{g}^{(i)}\|_{L^\infty(\nu^{(i)})}|\tau|\lambda^{(i)}/3n_i\right)} \\ &\le \frac{\sum_{i=1}^{K}\lambda^{(i)}\|\bar{g}^{(i)}\|^2_{L^2(\nu^{(i)})}\tau^2}{2\tilde{n}\left(1 - \|\bar{g}^{(\cdot)}\|_{L^\infty(\nu^{(\cdot)})}|\tau|/3\tilde{n}\right)} \\ &= \frac{\left\|\bar{g}^{(\cdot)}\right\|^2_{L^2(\nu^{(\cdot)})}\tau^2}{2\tilde{n}\left(1 - \left\|\bar{g}^{(\cdot)}\right\|_{L^\infty(\nu^{(\cdot)})}|\tau|/3\tilde{n}\right)}, \end{aligned} \tag{20}$$

where the second inequality uses the fact that, for $a > 0$, the map $(x, y) \mapsto x/(1 - axy)$ is non-decreasing in both arguments on $\{x, y > 0 : xy < 1/a\}$.

Eq. (20) shows that G_n is a sub-gamma random variable with variance factor $\|\bar{g}^{(\cdot)}\|^2_{L^2(\nu^{(\cdot)})}/\tilde{n}$ and scale parameter $\|\bar{g}^{(\cdot)}\|_{L^\infty(\nu^{(\cdot)})}/3\tilde{n}$. By the Chernoff bound, for any $t > 0$, we have

$$\mathbb{P}_{n\nu}\left\{G_n > \sqrt{\frac{2t}{\tilde{n}}}\left\|\bar{g}^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})} + \frac{t}{3\tilde{n}}\left\|\bar{g}^{(\cdot)}\right\|_{L^\infty(\nu^{(\cdot)})}\right\} \le e^{-t}.$$

This completes the proof. □

F.2 Concentration of the Clipped Process

We now prove Lem. 1.

Proof of Lem. 1. Fix T, T' satisfying Def. 2 (b) and Def. 2 (c). We first bound $g(x) = \phi^*[T](x) - \phi^*[T'](x)$. By Lem. 4, we have $\phi^*[T], \phi^*[T'] \ge 0$ and $\phi^*[T](T^{-1}(0)) = \phi^*[T'](T'^{-1}(0)) = 0$. Evaluating g at $x = T^{-1}(0)$ and $x = T'^{-1}(0)$ thus gives $g(T^{-1}(0)) \le 0$ and $g(T'^{-1}(0)) \ge 0$. Since g is continuous, the intermediate value theorem yields $x_0 \in \Omega$ on the segment between $T^{-1}(0)$ and $T'^{-1}(0)$ such that $g(x_0) = 0$. For such an x_0, the gradient theorem gives

$$g(x) = \int_0^1 \langle (T - T')(x_0 + t(x - x_0)), x - x_0\rangle dt.$$

By the Cauchy–Schwarz inequality,

$$|g(x)| \le \sup\{\|(T - T')(x')\| : x' = x_0 + t(x - x_0), t \in [0, 1]\}\|x - x_0\|.$$

Since $x_0 \in B_R$ by Def. 2 (b) and B_R is convex, for $x \in B_R$ we have

$$|g(x)| \le 2R\|T - T'\|_{L^\infty(B_R)}. \tag{21}$$

Let $\tilde{T}_1^{(\cdot)}, \dots, \tilde{T}_{N_\epsilon}^{(\cdot)}$ be an ϵ-net of $\mathcal{T}_{R,L,D}$ with respect to the $\bar{L}^\infty(B_R)$-norm, where $N_\epsilon = N(\epsilon, \mathcal{T}_{R,L,D}, \|\cdot\|_{\bar{L}^\infty(B_R)})$. Let $\tilde{T}_n^{(\cdot)}$ be the element of the net closest to $T_n^{(\cdot)}$ in the $\bar{L}^\infty(B_R)$-norm. Define

$$\tilde{g}_{M,n}^{(i)}(x) = \left(\phi^*\left[T_{\mathcal{T}}^{(i)}\right](x) - \phi^*\left[\tilde{T}_n^{(i)}\right](x)\right) \wedge M \vee (-M),\ \bar{\tilde{g}}_{M,n}^{(i)}(x) = \tilde{g}_{M,n}^{(i)}(x) - \int \tilde{g}_{M,n}^{(i)} d\nu^{(i)},$$

and

$$\tilde{\Delta}_{M,n} = \sum_{i=1}^{K} \frac{\lambda^{(i)}}{n_i} \sum_{j=1}^{n_i} \bar{\tilde{g}}_{M,n}^{(i)}\left(X_j^{(i)}\right).$$

Since the clipping function is 1-Lipschitz, Eq. (21) gives, for $x \in B_R$,

$$\left|g_{M,n}^{(i)}(x) - \tilde{g}_{M,n}^{(i)}(x)\right| \le \left|\phi^*[T_n^{(i)}](x) - \phi^*[\tilde{T}_n^{(i)}](x)\right| \le 2R\left\|T_n^{(i)} - \tilde{T}_n^{(i)}\right\|_{L^\infty(B_R)}.$$

For $x \notin B_R$, since both $g_{M,n}^{(i)}$ and $\tilde{g}_{M,n}^{(i)}$ take values in $[-M, M]$, we have $|g_{M,n}^{(i)}(x) - \tilde{g}_{M,n}^{(i)}(x)| \le 2M = 8LR^2$. Write $\epsilon_i = \|\|T_n^{(i)} - \tilde{T}_n^{(i)}\|\|_{L^\infty(B_R)}$. Combining the two cases gives, for all $i \in [K]$ and $x \in \Omega$,

$$\left|g_{M,n}^{(i)}(x) - \tilde{g}_{M,n}^{(i)}(x)\right| \le 2R\epsilon_i \mathbb{1}\{x \in B_R\} + 8LR^2 \mathbb{1}\{x \notin B_R\}.$$

Using $\mathbb{1}\{x \in B_R\} \le 1$, $R^2\mathbb{1}\{x \notin B_R\} \le \|x\|^2\mathbb{1}\{x \notin B_R\}$, and the triangle inequality, together with the definitions of $\Xi_{R,n}$ and Ξ_R, we obtain

$$\left|\Delta_{M,n} - \tilde{\Delta}_{M,n}\right| \le 4R\sum_{i=1}^{K} \lambda^{(i)}\epsilon_i + 8L(\Xi_{R,n} + \Xi_R) \le 4R\epsilon + 8L(\Xi_{R,n} + \Xi_R), \tag{22}$$

where the last inequality follows from the Cauchy–Schwarz inequality $\sum_{i=1}^K \lambda^{(i)}\epsilon_i \le (\sum_{i=1}^K \lambda^{(i)}\epsilon_i^2)^{1/2} = \|T_n^{(\cdot)} - \tilde{T}_n^{(\cdot)}\|_{\bar{L}^\infty(B_R)} \le \epsilon$.

We next bound $\tilde{\Delta}_{M,n}$. By construction, $|\tilde{g}_{M,n}^{(i)}(x)| \le M$, and thus $\|\bar{\tilde{g}}_{M,n}^{(i)}\|_{L^\infty(\nu^{(i)})} \le 2M$. Moreover, $\tilde{g}_{M,n}^{(\cdot)}$ ranges over at most N_ϵ tuples of functions. Hence, by Lem. 5 and the union bound, for any $t > 0$, with probability at least $1 - N_\epsilon e^{-t}$,

$$\tilde{\Delta}_{M,n} \le \sqrt{\frac{2t}{\tilde{n}}}\left\|\bar{\tilde{g}}_{M,n}^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})} + \frac{2Mt}{3\tilde{n}}. \tag{23}$$

By Eq. (21) and the bound outside B_R derived above, we have

$$\left\|\tilde{g}_{M,n}^{(i)} - g_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})}^2 \le 4R^2\epsilon_i^2 + 64L^2R^2\Xi_R^{(i)}.$$

By the subadditivity of the standard deviation, we have

$$\begin{aligned}
\left\|\bar{\tilde{g}}_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})}^2 &\le \left(\left\|\bar{g}_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})} + \mathbb{V}_{\nu^{(i)}}\left[\tilde{g}_{M,n}^{(i)} - g_{M,n}^{(i)}\right]^{1/2}\right)^2 \\
&\le \left(\left\|\bar{g}_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})} + \left\|\tilde{g}_{M,n}^{(i)} - g_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})}\right)^2 \\
&\le \left(\left\|\bar{g}_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})} + \left(4R^2\epsilon_i^2 + 64L^2R^2\Xi_R^{(i)}\right)^{1/2}\right)^2.
\end{aligned}$$

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
&\left\|\tilde{\bar{g}}^{(i)}_{M,n}\right\|^2_{L^2(\nu^{(\cdot)})} \\
\le&\sum_{i=1}^{K}\lambda^{(i)}\left(\left\|\bar{g}^{(i)}_{M,n}\right\|_{L^2(\nu^{(i)})}+\left(4R^2\epsilon_i^2+64L^2R^2\Xi_R^{(i)}\right)^{1/2}\right)^2 \\
\le&\left\|\bar{g}^{(\cdot)}_{M,n}\right\|^2_{L^2(\nu^{(\cdot)})}+4R^2\epsilon^2+64L^2R^2\Xi_R \\
&\qquad\qquad+2\sum_{i=1}^{K}\lambda^{(i)}\left\|\bar{g}^{(i)}_{M,n}\right\|_{L^2(\nu^{(i)})}\left(4R^2\epsilon_i^2+64L^2R^2\Xi_R^{(i)}\right)^{1/2} \\
\le&\left\|\bar{g}^{(\cdot)}_{M,n}\right\|^2_{L^2(\nu^{(\cdot)})}+4R^2\epsilon^2+64L^2R^2\Xi_R+2\left\|\bar{g}^{(\cdot)}_{M,n}\right\|_{L^2(\nu^{(\cdot)})}\left(4R^2\epsilon^2+64L^2R^2\Xi_R\right)^{1/2} \\
\le&\left(\left\|\bar{g}^{(\cdot)}_{M,n}\right\|_{L^2(\nu^{(\cdot)})}+\left(4R^2\epsilon^2+64L^2R^2\Xi_R\right)^{1/2}\right)^2. \qquad (24)
\end{aligned}$$

Combining Eq. (22), Eq. (23), and Eq. (24) gives, with probability at least $1-N_\epsilon e^{-t}$,

$$\Delta_{M,n}\le 4R\epsilon+8L(\Xi_{R,n}+\Xi_R)+\sqrt{\frac{2t}{\tilde{n}}}\left(\left\|\bar{g}^{(\cdot)}_{M,n}\right\|_{L^2(\nu^{(\cdot)})}+\left(4R^2\epsilon^2+64L^2R^2\Xi_R\right)^{1/2}\right)+\frac{2Mt}{3\tilde{n}}.$$

Replacing t with $t+H_\epsilon$, where $H_\epsilon=\ln N_\epsilon$, completes the proof. □

G Analysis of the Beyond-Threshold Process

In this section, we prove Lem. 2.

Proof of Lem. 2. We bound $\Delta_{M^c,n}$ through an envelope of $g^{(i)}_{M^c,n}$ that does not depend on the data-dependent $T_n^{(\cdot)}$. By Lem. 4 and Def. 2 (b), we have $|g_n^{(i)}(x)|\le L(\|x\|+R)^2$. Since $|g^{(i)}_{M^c,n}(x)|=0\vee(|g_n^{(i)}(x)|-M)$ by Eq. (8), $|g^{(i)}_{M^c,n}(x)|>0$ only if $L(\|x\|+R)^2>M$, equivalently $\|x\|>\sqrt{M/L}-R$. With the choice $M=4LR^2$, $|g^{(i)}_{M^c,n}(x)|>0$ only if $x\notin B_R$. For $x\notin B_R$, we have

$$\begin{aligned}
\left|g^{(i)}_{M^c,n}(x)\right|\le&0\vee\left(L(\|x\|+R)^2-M\right)\\
=&0\vee L\|x\|^2\left(\left(1+\frac{R}{\|x\|}\right)^2-\left(\frac{2R}{\|x\|}\right)^2\right)\\
\le&\frac{4L\|x\|^2}{3}\mathbb{1}\{x\notin B_R\},
\end{aligned}$$

where we use $(1+u)^2-(2u)^2\le 4/3$ for $u\in[0,1]$ with $u=R/\|x\|$. Hence, $|g^{(i)}_{M^c,n}(x)|\le\frac{4L}{3}\|x\|^2\mathbb{1}\{x\notin B_R\}$ for all $x\in\Omega$. Since $\bar{g}^{(i)}_{M^c,n}=g^{(i)}_{M^c,n}-\int g^{(i)}_{M^c,n}d\nu^{(i)}$, this envelope gives

$$\begin{aligned}
\Delta_{M^c,n}\le&\sum_{i=1}^{K}\frac{\lambda^{(i)}}{n_i}\sum_{j=1}^{n_i}\left|g^{(i)}_{M^c,n}\left(X_j^{(i)}\right)\right|+\sum_{i=1}^{K}\lambda^{(i)}\int\left|g^{(i)}_{M^c,n}\right|d\nu^{(i)}\\
\le&\frac{4L}{3}\sum_{i=1}^{K}\frac{\lambda^{(i)}}{n_i}\sum_{j=1}^{n_i}\left\|X_j^{(i)}\right\|^2\mathbb{1}\left\{X_j^{(i)}\notin B_R\right\}+\frac{4L}{3}\Xi_R,
\end{aligned}$$

where the second inequality uses the definition of Ξ_R in Eq. (4). Since the envelope is deterministic, taking the expectation over the samples and using $\mathbb{E}_{n\nu}[\|X_j^{(i)}\|^2\mathbb{1}\{X_j^{(i)}\notin$

$B_R\}] = \Xi_R^{(i)}$ gives

$$\mathbb{E}_{n\nu}[\Delta_{M^c,n}] \le \frac{8L}{3}\Xi_R.$$

This is the claimed bound. □

H Details of the Potential Error Analysis

This section gives the full proof of Thm. 4.

Proof of Thm. 4. Throughout the proof, we set $M = 4LR^2$ and $A = A(T_\nu^{(\cdot)}; \mathcal{T}_{R,L,D})$. By Lem. 1 and $\Delta_n = \Delta_{M,n} + \Delta_{M^c,n}$, for any $t > 0$, with probability at least $1 - e^{-t}$,

$$\Delta_n \le 4R\epsilon + \Delta_{M^c,n} + 8L(\Xi_{R,n} + \Xi_R) + \sqrt{\frac{2(H_\epsilon + t)}{\tilde{n}}}\left(\left\|\bar{g}_{M,n}^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})} + \left(4R^2\epsilon^2 + 64L^2R^2\Xi_R\right)^{1/2}\right) + \frac{2M(H_\epsilon + t)}{3\tilde{n}}. \quad (25)$$

By Def. 1, we have

$$\begin{aligned}\left\|\bar{g}_{M,n}^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})}^2 &= \sum_{i=1}^K \lambda^{(i)} \mathbb{V}_{\nu^{(i)}}\left[g_{M,n}^{(i)}\right] \\ &\le C_\alpha M^{2(1-\alpha)} \sum_{i=1}^K \lambda^{(i)} \left\|\nabla g_{M,n}^{(i)}\right\|_{L^2(\nu^{(i)})}^{2\alpha} \\ &\le C_\alpha M^{2(1-\alpha)} \sum_{i=1}^K \lambda^{(i)} \left\|\nabla g_n^{(i)}\right\|_{L^2(\nu^{(i)})}^{2\alpha} \\ &\le C_\alpha M^{2(1-\alpha)} \left\|\nabla g_n^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})}^{2\alpha} = C_\alpha M^{2(1-\alpha)} \left\|T_n^{(\cdot)} - T_{\mathcal{T}}^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})}^{2\alpha}, \end{aligned} \quad (26)$$

where the first inequality uses $|g_{M,n}^{(i)}| \le M$, the second inequality uses the fact that the gradient of $g_{M,n}^{(i)}$ vanishes wherever $|g_n^{(i)}| > M$, and the third inequality follows from Jensen's inequality for the concave map $x \mapsto x^\alpha$. By Thm. 3 and the Cauchy–Schwarz inequality, we have

$$\left\|T_n^{(\cdot)} - T_{\mathcal{T}}^{(\cdot)}\right\|_{L^2(\nu^{(\cdot)})} \le \left(2L\left(\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu\right)\right)^{1/2} + (2LA)^{1/2}. \quad (27)$$

Combining Eq. (25), Eq. (26), Eq. (27), the decomposition Eq. (7), and the AM–GM inequality, we obtain, for any $\delta > 0$ and $t > 0$, with probability at least $1 - e^{-t}$,

$$\begin{aligned}\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu \le A + 4R\epsilon + \Delta_{M^c,n} + 8L(\Xi_{R,n} + \Xi_R) \\ + \delta\left(C_{L,R}\left(\left(\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu\right)^{1/2} + A^{1/2}\right)^\alpha \right. \\ \left. + \left(\frac{\epsilon^2}{L} + 16L\Xi_R\right)^{1/2}\right)^2 + \frac{2M(H_\epsilon + t)}{\tilde{n}}\left(\frac{1}{2\delta} + \frac{1}{3}\right), \end{aligned} \quad (28)$$

where $C_{L,R}$ is defined as in the statement. Let $E_\Phi = \Phi(T_n^{(\cdot)}) - \Phi_\nu$. We split the right-hand side of Eq. (28) into the part E'_Φ that does not depend on t and the part $E_\Phi(t)$ that does:

$$\begin{aligned}E'_\Phi = \Phi\left(T_{\mathcal{T}}^{(\cdot)}\right) - A + 4R\epsilon + \Delta_{M^c,n} + 8L(\Xi_{R,n} + \Xi_R) \\ + \delta\left(C_{L,R}\left(\left(\Phi\left(T_n^{(\cdot)}\right) - \Phi_\nu\right)^{1/2} + A^{1/2}\right)^\alpha \right.\end{aligned}$$

$$+\left(\frac{\epsilon^2}{L}+16L\Xi_R\right)^{1/2}\Bigg)^2+\frac{2MH_\epsilon}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right),$$

and

$$E_\Phi(t)=\frac{2Mt}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right).$$

Then, we have

$$\mathbb{E}_{n\nu}[E_\Phi]=\mathbb{E}_{n\nu}[E_\Phi-E'_\Phi]+\mathbb{E}_{n\nu}[E'_\Phi].$$

By the Cauchy–Schwarz inequality and Jensen's inequality, for a positive random variable X and non-random variables $a,b,c>0$, we have

$$\begin{aligned}
&\mathbb{E}_{n\nu}\left[\left(a(X^{1/2}+b^{1/2})^\alpha+c^{1/2}\right)^2\right]\\
=&a^2\mathbb{E}_{n\nu}\left[(X^{1/2}+b^{1/2})^{2\alpha}\right]+c+2ac^{1/2}\mathbb{E}_{n\nu}\left[\left(X^{1/2}+b^{1/2}\right)^\alpha\right]\\
\le&a^2\left(\mathbb{E}_{n\nu}\left[X+b+2X^{1/2}b^{1/2}\right]\right)^\alpha+c+2ac^{1/2}\left(\mathbb{E}_{n\nu}\left[X+b+2X^{1/2}b^{1/2}\right]\right)^{\alpha/2}\\
\le&a^2\left(\mathbb{E}_{n\nu}[X]^{1/2}+b^{1/2}\right)^{2\alpha}+c+2ac^{1/2}\left(\mathbb{E}_{n\nu}[X]^{1/2}+b^{1/2}\right)^\alpha\\
=&\left(a\left(\mathbb{E}_{n\nu}[X]^{1/2}+b^{1/2}\right)^\alpha+c^{1/2}\right)^2.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{E}_{n\nu}[E'_\Phi]\le A+4R\epsilon+\mathbb{E}_{n\nu}[\Delta_{M^c,n}]+8L(\mathbb{E}_{n\nu}[\Xi_{R,n}]+\Xi_R)\\
+\delta\Bigg(C_{L,R}\left(\mathbb{E}_{n\nu}[E_\Phi]^{1/2}+A^{1/2}\right)^\alpha\\
+\left(\frac{\epsilon^2}{L}+16L\Xi_R\right)^{1/2}\Bigg)^2+\frac{2MH_\epsilon}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right).
\end{aligned}$$

By Lem. 2, we have $\mathbb{E}_{n\nu}[\Delta_{M^c,n}]\le 8L\Xi_R/3$. By the linearity of the expectation and the definition of Ξ_R, we have $\mathbb{E}_{n\nu}[\Xi_{R,n}]=\Xi_R$. From Eq. (28), for any $t>0$, with probability at least $1-e^{-t}$, we have

$$E_\Phi-E'_\Phi\le E_\Phi(t).$$

Hence,

$$\begin{aligned}
\mathbb{E}_{n\nu}[E_\Phi-E'_\Phi]\le&\int_0^\infty\mathbb{P}\{E_\Phi-E'_\Phi>s\}ds\\
=&\frac{2M}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right)\int_0^\infty\mathbb{P}\{E_\Phi-E'_\Phi>E_\Phi(t)\}dt\\
\le&\frac{2M}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right)\int_0^\infty e^{-t}dt=\frac{2M}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right).
\end{aligned}$$

Consequently, letting $\bar{E}_\Phi=\mathbb{E}_{n\nu}[E_\Phi]$, we have

$$\begin{aligned}
\bar{E}_\Phi\le A+4R\epsilon+\frac{56L\Xi_R}{3}+\delta\Bigg(C_{L,R}\left(\bar{E}_\Phi^{1/2}+A^{1/2}\right)^\alpha+\left(\frac{\epsilon^2}{L}+16L\Xi_R\right)^{1/2}\Bigg)^2\\
+\frac{2M(1+H_\epsilon)}{\tilde{n}}\left(\frac{1}{2\delta}+\frac{1}{3}\right).
\end{aligned}$$

Since $\delta>0$ is arbitrary, we take

$$\delta^2=\frac{M(1+H_\epsilon)}{\tilde{n}}\left(C_{L,R}\left(\bar{E}_\Phi^{1/2}+A^{1/2}\right)^\alpha+\left(\frac{\epsilon^2}{L}+16L\Xi_R\right)^{1/2}\right)^{-2},$$

which gives

$$\bar{E}_\Phi \le A + 4R\epsilon + \frac{56L\Xi_R}{3} + \frac{2M(1+H_\epsilon)}{3\tilde{n}} + \sqrt{\frac{M(1+H_\epsilon)}{\tilde{n}}}\left(C_{L,R}\left(\bar{E}_\Phi^{1/2} + A^{1/2}\right)^\alpha + \left(\frac{\epsilon^2}{L} + 16L\Xi_R\right)^{1/2}\right).$$

By assumption, we have $\epsilon \le \gamma_{\tilde{n}} \frac{LR(1+H_\epsilon)}{\tilde{n}}$ and $\Xi_R \le \gamma_{\tilde{n}} \frac{R^2(1+H_\epsilon)}{\tilde{n}}$. Then, we have

$$\bar{E}_\Phi \le A + \left(\frac{2}{3} + \frac{17\gamma_{\tilde{n}}}{3} + \left(\frac{1+H_\epsilon}{4\tilde{n}}\gamma_{\tilde{n}}^2 + 4\gamma_{\tilde{n}}\right)^{1/2}\right)\frac{M(1+H_\epsilon)}{\tilde{n}} + C_{L,R}\sqrt{\frac{M(1+H_\epsilon)}{\tilde{n}}}\left(\bar{E}_\Phi^{1/2} + A^{1/2}\right)^\alpha.$$

Let $C_{\tilde{n}} = 1 + 12\sqrt{\gamma_{\tilde{n}}}$. Under the assumption that $(1 \vee C_{L,R})^2 \frac{M(1+H_\epsilon)}{\tilde{n}} \le 1$, noting that $M \ge 1$ and $\gamma_{\tilde{n}} \le 1$, we have

$$\begin{aligned}
&\frac{17\gamma_{\tilde{n}}}{2} + \frac{3}{2}\left(\frac{1+H_\epsilon}{4\tilde{n}}\gamma_{\tilde{n}}^2 + 4\gamma_{\tilde{n}}\right)^{1/2} \\
\le &\frac{17\gamma_{\tilde{n}}}{2} + \frac{3}{2}\left(\frac{1}{4} + 4\right)^{1/2}\sqrt{\gamma_{\tilde{n}}} \\
\le &\left(9 + \frac{9}{4}\right)\sqrt{\gamma_{\tilde{n}}} \le C_{\tilde{n}}.
\end{aligned}$$

Consequently, we have

$$\bar{E}_\Phi \le A + \frac{2C_{\tilde{n}}}{3}\frac{M(1+H_\epsilon)}{\tilde{n}} + C_{L,R}\sqrt{\frac{M(1+H_\epsilon)}{\tilde{n}}}\left(\bar{E}_\Phi^{1/2} + A^{1/2}\right)^\alpha.$$

Consider first the case $A = 0$. Let $B_{\tilde{n}} = \frac{1}{2}(1 + \sqrt{1 + \frac{8C_{\tilde{n}}}{3}})$. Assume that $\bar{E}_\Phi > (B_{\tilde{n}}^2(1 \vee C_{L,R})^2 \frac{M(1+H_\epsilon)}{\tilde{n}})^{\frac{1}{2-\alpha}}$. Then, we have

$$\left(B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right)^{\frac{1}{2-\alpha}} < \frac{2C_{\tilde{n}}}{3}B_{\tilde{n}}^{-2}\left(B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right) + B_{\tilde{n}}^{-1}\sqrt{B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}}\bar{E}_\Phi^{\alpha/2}.$$

Under the assumption that $(1 \vee C_{L,R})^2 \frac{M(1+H_\epsilon)}{\tilde{n}} \le 1/5$, using $B_{\tilde{n}} \le 5$ by definition, we have

$$\left(B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right)^{\frac{1}{2-\alpha}} < \frac{2C_{\tilde{n}}}{3}B_{\tilde{n}}^{-2}\left(B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right)^{\frac{1}{2-\alpha}} + B_{\tilde{n}}^{-1}\sqrt{B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}}\bar{E}_\Phi^{\alpha/2}.$$

Rearranging the terms gives

$$\left(B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right)^{\frac{\alpha}{2(2-\alpha)}} B_{\tilde{n}}\left(1 - \frac{2C_{\tilde{n}}}{3}B_{\tilde{n}}^{-2}\right) < \bar{E}_\Phi^{\alpha/2}.$$

Raising both sides to the power $2/\alpha$ gives

$$\left(B_{\tilde{n}}^2(1 \vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right)^{\frac{1}{2-\alpha}}\left(B_{\tilde{n}}\left(1 - \frac{2C_{\tilde{n}}}{3}B_{\tilde{n}}^{-2}\right)\right)^{\frac{2}{\alpha}} \le \bar{E}_\Phi.$$

This contradicts the previous assumption if $B_{\tilde{n}}\left(1-\frac{2C_{\tilde{n}}}{3}B_{\tilde{n}}^{-2}\right)\geq 1$. This condition is equivalent to

$$\begin{aligned}
&B_{\tilde{n}}^2 - B_{\tilde{n}} \geq \frac{2C_{\tilde{n}}}{3} \\
\iff &\left(B_{\tilde{n}} - \frac{1}{2}\right)^2 \geq \frac{1}{4}\left(1+\frac{8C_{\tilde{n}}}{3}\right) \\
\iff &B_{\tilde{n}} \geq \frac{1}{2}\left(1+\sqrt{1+\frac{8C_{\tilde{n}}}{3}}\right).
\end{aligned}$$

This condition always holds by the definition of $B_{\tilde{n}}$. Hence, we have

$$\bar{E}_\Phi \leq \left(B_{\tilde{n}}^2(1\vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}}\right)^{\frac{1}{2-\alpha}}.$$

Next, we consider the case where $(1\vee C_{L,R})^2\frac{M(1+H_\epsilon)}{\tilde{n}} \leq \gamma_{\tilde{n}}A^{2-\alpha}$. Then, we have

$$\frac{\bar{E}_\Phi}{A} \leq 1+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}(1\vee C_{L,R})^{-2}A^{1-\alpha} + C_{L,R}(1\vee C_{L,R})^{-1}\left(\left(\frac{\bar{E}_\Phi}{A}\right)^{1/2}+1\right)^\alpha\sqrt{\gamma_{\tilde{n}}}.$$

Since $1\vee C_{L,R}\geq 1$ and $C_{L,R}(1\vee C_{L,R})^{-1}\leq 1$, we have

$$\frac{\bar{E}_\Phi}{A} \leq 1+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha} + \left(\left(\frac{\bar{E}_\Phi}{A}\right)^{1/2}+1\right)^\alpha\sqrt{\gamma_{\tilde{n}}}.$$

If $\bar{E}_\Phi \leq A$, then the claim holds. Otherwise, we have

$$\begin{aligned}
\frac{\bar{E}_\Phi}{A} \leq& 1+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha} + \left(2\left(\frac{\bar{E}_\Phi}{A}\right)^{1/2}\right)^\alpha\sqrt{\gamma_{\tilde{n}}} \\
\leq& 1+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha} + 2\left(\frac{\bar{E}_\Phi}{A}\right)^{1/2}\sqrt{\gamma_{\tilde{n}}}.
\end{aligned}$$

Rearranging the terms gives

$$\left(\left(\frac{\bar{E}_\Phi}{A}\right)^{1/2} - \sqrt{\gamma_{\tilde{n}}}\right)^2 \leq 1+\gamma_{\tilde{n}}+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha}.$$

Hence,

$$\begin{aligned}
\frac{\bar{E}_\Phi}{A} \leq& \left(\sqrt{\gamma_{\tilde{n}}}+\sqrt{1+\gamma_{\tilde{n}}+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha}}\right)^2 \\
=& 1+2\gamma_{\tilde{n}}+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha}+\sqrt{\gamma_{\tilde{n}}}\sqrt{1+\gamma_{\tilde{n}}+\frac{2C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha}}.
\end{aligned}$$

The subadditivity of the square root and the AM–GM inequality give

$$\frac{\bar{E}_\Phi}{A} \leq 1+\frac{7}{4}\gamma_{\tilde{n}}+\frac{4C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha}+\sqrt{\gamma_{\tilde{n}}} \leq 1+\frac{11}{4}\sqrt{\gamma_{\tilde{n}}}+\frac{4C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha}.$$

By the definitions of $B_{\tilde{n}}$ and $C_{\tilde{n}}$, we have

$$\frac{2B_{\tilde{n}}}{1+\sqrt{11/3}} = \frac{1+\sqrt{1+8C_{\tilde{n}}/3}}{1+\sqrt{11/3}} = \frac{1+\sqrt{11/3+12\sqrt{\gamma_{\tilde{n}}}}}{1+\sqrt{11/3}} \leq 1+\frac{\sqrt{12\sqrt{\gamma_{\tilde{n}}}}}{1+\sqrt{11/3}} \leq 1+2\gamma_{\tilde{n}}^{1/4},$$

where we use the subadditivity of the square root. Also, we have

$$\left(\frac{1+\sqrt{11/3}}{2}\right)^2 = \frac{14+\sqrt{33}}{12}.$$

Since $C_{\tilde{n}} \leq 13$, we have

$$\begin{aligned} &1 + \frac{11}{4}\sqrt{\gamma_{\tilde{n}}} + \frac{4C_{\tilde{n}}\gamma_{\tilde{n}}}{3}A^{1-\alpha} \\ \leq &1 + 3\sqrt{\gamma_{\tilde{n}}} + 20\gamma_{\tilde{n}}A^{1-\alpha}. \end{aligned}$$

This concludes the proof. □

I Proof of the Main Theorem

In this section, we prove Thm. 2.

Proof of Thm. 2. The proof combines Thm. 4 and Thm. 3. Take $\epsilon = o(LR/\tilde{n})$. For any $L, R, D \geq 1$, by Def. 2 (d), there exists $\beta \geq 1$ such that $1 + H_\epsilon \leq (1 + o(1))D\ln^\beta(\tilde{n}/LR)$ owing to the assumption $LR = o(\tilde{n})$. Then, all the conditions of Thm. 4 are satisfied, and applying Thm. 4 and Thm. 3 gives the claimed bound. □